One Time of Interaction May Not Be Enough: Go Deep with an Interaction-over-Interaction Network for Response Selection in Dialogues

- https://www.aclweb.org/anthology/P19-1001/

Currently, researchers have paid great attention to retrieval-based dialogues in open-domain. In particular, people study the problem by investigating context-response matching for multi-turn response selection based on publicly recognized benchmark data sets. State-of-the-art methods require a response to interact with each utterance in a context from the beginning, but the interaction is performed in a shallow way. In this work, we let utterance-response interaction go deep by proposing an interaction-over-interaction network (IoI). The model performs matching by stacking multiple interaction blocks in which residual information from one time of interaction initiates the interaction process again. Thus, matching information within an utterance-response pair is extracted from the interaction of the pair in an iterative fashion, and the information flows along the chain of the blocks via representations. Evaluation results on three benchmark data sets indicate that IoI can significantly outperform state-of-the-art methods in terms of various matching metrics. Through further analysis, we also unveil how the depth of interaction affects the performance of IoI.

Incremental Transformer with Deliberation Decoder for Document Grounded Conversations

- https://www.aclweb.org/anthology/P19-1002/

Document Grounded Conversations is a task to generate dialogue responses when chatting about the content of a given document. Obviously, document knowledge plays a critical role in Document Grounded Conversations, while existing dialogue models do not exploit this kind of knowledge effectively enough. In this paper, we propose a novel Transformer-based architecture for multi-turn document grounded conversations. In particular, we devise an Incremental Transformer to encode multi-turn utterances along with knowledge in related documents. Motivated by the human cognitive process, we design a two-pass decoder (Deliberation Decoder) to improve context coherence and knowledge correctness. Our empirical study on a real-world Document Grounded Dataset proves that responses generated by our model significantly outperform competitive baselines on both context coherence and knowledge relevance.

Improving Multi-turn Dialogue Modelling with Utterance ReWriter

- https://www.aclweb.org/anthology/P19-1003/

Recent research has achieved impressive results in single-turn dialogue modelling. In the multi-turn setting, however, current models are still far from satisfactory. One major challenge is the frequently occurred coreference and information omission in our daily conversation, making it hard for machines to understand the real intention. In this paper, we propose rewriting the human utterance as a pre-process to help multi-turn dialgoue modelling. Each utterance is first rewritten to recover all coreferred and omitted information. The next processing steps are then performed based on the rewritten utterance. To properly train the utterance rewriter, we collect a new dataset with human annotations and introduce a Transformer-based utterance rewriting architecture using the pointer network. We show the proposed architecture achieves remarkably good performance on the utterance rewriting task. The trained utterance rewriter can be easily integrated into online chatbots and brings general improvement over different domains.

Do Neural Dialog Systems Use the Conversation History Effectively? An Empirical Study

- https://www.aclweb.org/anthology/P19-1004/

Neural generative models have been become increasingly popular when building conversational agents. They offer flexibility, can be easily adapted to new domains, and require minimal domain engineering. A common criticism of these systems is that they seldom understand or use the available dialog history effectively. In this paper, we take an empirical approach to understanding how these models use the available dialog history by studying the sensitivity of the models to artificially introduced unnatural changes or perturbations to their context at test time. We experiment with 10 different types of perturbations on 4 multi-turn dialog datasets and find that commonly used neural dialog architectures like recurrent and transformer-based seq2seq models are rarely sensitive to most perturbations such as missing or reordering utterances, shuffling words, etc. Also, by open-sourcing our code, we believe that it will serve as a useful diagnostic tool for evaluating dialog systems in the future.

Boosting Dialog Response Generation

- https://www.aclweb.org/anthology/P19-1005/

Neural models have become one of the most important approaches to dialog response generation. However, they still tend to generate the most common and generic responses in the corpus all the time. To address this problem, we designed an iterative training process and ensemble method based on boosting. We combined our method with different training and decoding paradigms as the base model, including mutual-information-based decoding and reward-augmented maximum likelihood learning. Empirical results show that our ap-

proach can significantly improve the diversity and relevance of the responses generated by all base models, backed by objective measurements and human evaluation.

Constructing Interpretive Spatio-Temporal Features for Multi-Turn Responses Selection

- https://www.aclweb.org/anthology/P19-1006/

Response selection plays an important role in fully automated dialogue systems. Given the dialogue context, the goal of response selection is to identify the best-matched next utterance (i.e., response) from multiple candidates. Despite the efforts of many previous useful models, this task remains challenging due to the huge semantic gap and also the large size of candidate set. To address these issues, we propose a Spatio-Temporal Matching network (STM) for response selection. In detail, soft alignment is first used to obtain the local relevance between the context and the response. And then, we construct spatio-temporal features by aggregating attention images in time dimension and make use of 3D convolution and pooling operations to extract matching information. Evaluation on two large-scale multi-turn response selection tasks has demonstrated that our proposed model significantly outperforms the state-of-the-art model. Particularly, visualization analysis shows that the spatio-temporal features enables matching information in segment pairs and time sequences, and have good interpretability for multi-turn text matching.

Semantic Parsing with Dual Learning

- https://www.aclweb.org/anthology/P19-1007/

Semantic parsing converts natural language queries into structured logical forms. The lack of training data is still one of the most serious problems in this area. In this work, we develop a semantic parsing framework with the dual learning algorithm, which enables a semantic parser to make full use of data (labeled and even unlabeled) through a dual-learning game. This game between a primal model (semantic parsing) and a dual model (logical form to query) forces them to regularize each other, and can achieve feedback signals from some prior-knowledge. By utilizing the prior-knowledge of logical form structures, we propose a novel reward signal at the surface and semantic levels which tends to generate complete and reasonable logical forms. Experimental results show that our approach achieves new state-of-the-art performance on ATIS dataset and gets competitive performance on OVERNIGHT dataset.

Semantic Expressive Capacity with Bounded Memory

- https://www.aclweb.org/anthology/P19-1008/

We investigate the capacity of mechanisms for compositional semantic parsing to describe relations between sentences and semantic representations. We prove that in order to represent certain relations, mechanisms which are syntactically projective must be able to remember an unbounded number of locations in the semantic representations, where nonprojective mechanisms need not. This is the first result of this kind, and has consequences both for grammar-based and for neural systems.

AMR Parsing as Sequence-to-Graph Transduction

- https://www.aclweb.org/anthology/P19-1009/

We propose an attention-based model that treats AMR parsing as sequence-to-graph transduction. Unlike most AMR parsers that rely on pre-trained aligners, external semantic resources, or data augmentation, our proposed parser is aligner-free, and it can be effectively trained with limited amounts of labeled AMR data. Our experimental results outperform all previously reported SMATCH scores, on both AMR 2.0 (76.3% on LDC2017T10) and AMR 1.0 (70.2% on LDC2014T12).

Generating Logical Forms from Graph Representations of Text and Entities

- https://www.aclweb.org/anthology/P19-1010/

Structured information about entities is critical for many semantic parsing tasks. We present an approach that uses a Graph Neural Network (GNN) architecture to incorporate information about relevant entities and their relations during parsing. Combined with a decoder copy mechanism, this approach provides a conceptually simple mechanism to generate logical forms with entities. We demonstrate that this approach is competitive with the state-of-the-art across several tasks without pre-training, and outperforms existing approaches when combined with BERT pre-training.

Learning Compressed Sentence Representations for On-Device Text Processing

- https://www.aclweb.org/anthology/P19-1011/

Vector representations of sentences, trained on massive text corpora, are widely used as generic sentence embeddings across a variety of NLP problems. The

learned representations are generally assumed to be continuous and real-valued, giving rise to a large memory footprint and slow retrieval speed, which hinders their applicability to low-resource (memory and computation) platforms, such as mobile devices. In this paper, we propose four different strategies to transform continuous and generic sentence embeddings into a binarized form, while preserving their rich semantic information. The introduced methods are evaluated across a wide range of downstream tasks, where the binarized sentence embeddings are demonstrated to degrade performance by only about 2% relative to their continuous counterparts, while reducing the storage requirement by over 98%. Moreover, with the learned binary representations, the semantic relatedness of two sentences can be evaluated by simply calculating their Hamming distance, which is more computational efficient compared with the inner product operation between continuous embeddings. Detailed analysis and case study further validate the effectiveness of proposed methods.

The (Non-)Utility of Structural Features in BiLSTM-based Dependency Parsers

- https://www.aclweb.org/anthology/P19-1012/

Classical non-neural dependency parsers put considerable effort on the design of feature functions. Especially, they benefit from information coming from structural features, such as features drawn from neighboring tokens in the dependency tree. In contrast, their BiLSTM-based successors achieve state-of-the-art performance without explicit information about the structural context. In this paper we aim to answer the question: How much structural context are the BiLSTM representations able to capture implicitly? We show that features drawn from partial subtrees become redundant when the BiLSTMs are used. We provide a deep insight into information flow in transition- and graph-based neural architectures to demonstrate where the implicit information comes from when the parsers make their decisions. Finally, with model ablations we demonstrate that the structural context is not only present in the models, but it significantly influences their performance.

Automatic Generation of High Quality CCGbanks for Parser Domain Adaptation

- https://www.aclweb.org/anthology/P19-1013/

We propose a new domain adaptation method for Combinatory Categorial Grammar (CCG) parsing, based on the idea of automatic generation of CCG corpora exploiting cheaper resources of dependency trees. Our solution is conceptually simple, and not relying on a specific parser architecture, making it applicable to the current best-performing parsers. We conduct extensive parsing

experiments with detailed discussion; on top of existing benchmark datasets on (1) biomedical texts and (2) question sentences, we create experimental datasets of (3) speech conversation and (4) math problems. When applied to the proposed method, an off-the-shelf CCG parser shows significant performance gains, improving from 90.7% to 96.6% on speech conversation, and from 88.5% to 96.8% on math problems.

A Joint Named-Entity Recognizer for Heterogeneous Tag-sets Using a Tag Hierarchy

- https://www.aclweb.org/anthology/P19-1014/

We study a variant of domain adaptation for named-entity recognition where multiple, heterogeneously tagged training sets are available. Furthermore, the test tag-set is not identical to any individual training tag-set. Yet, the relations between all tags are provided in a tag hierarchy, covering the test tags as a combination of training tags. This setting occurs when various datasets are created using different annotation schemes. This is also the case of extending a tag-set with a new tag by annotating only the new tag in a new dataset. We propose to use the given tag hierarchy to jointly learn a neural network that shares its tagging layer among all tag-sets. We compare this model to combining independent models and to a model based on the multitasking approach. Our experiments show the benefit of the tag-hierarchy model, especially when facing non-trivial consolidation of tag-sets.

Massively Multilingual Transfer for NER

- https://www.aclweb.org/anthology/P19-1015/

In cross-lingual transfer, NLP models over one or more source languages are applied to a low-resource target language. While most prior work has used a single source model or a few carefully selected models, here we consider a "massive" setting with many such models. This setting raises the problem of poor transfer, particularly from distant languages. We propose two techniques for modulating the transfer, suitable for zero-shot or few-shot learning, respectively. Evaluating on named entity recognition, we show that our techniques are much more effective than strong baselines, including standard ensembling, and our unsupervised method rivals oracle selection of the single best individual model.

Reliability-aware Dynamic Feature Composition for Name Tagging

- https://www.aclweb.org/anthology/P19-1016/

Word embeddings are widely used on a variety of tasks and can substantially improve the performance. However, their quality is not consistent throughout the vocabulary due to the long-tail distribution of word frequency. Without sufficient contexts, rare word embeddings are usually less reliable than those of common words. However, current models typically trust all word embeddings equally regardless of their reliability and thus may introduce noise and hurt the performance. Since names often contain rare and uncommon words, this problem is particularly critical for name tagging. In this paper, we propose a novel reliability-aware name tagging model to tackle this issue. We design a set of word frequency-based reliability signals to indicate the quality of each word embedding. Guided by the reliability signals, the model is able to dynamically select and compose features such as word embedding and character-level representation using gating mechanisms. For example, if an input word is rare, the model relies less on its word embedding and assigns higher weights to its character and contextual features. Experiments on OntoNotes 5.0 show that our model outperforms the baseline model by 2.7% absolute gain in F-score. In cross-genre experiments on five genres in OntoNotes, our model improves the performance for most genre pairs and obtains up to 5% absolute F-score gain.

Unsupervised Pivot Translation for Distant Languages

- https://www.aclweb.org/anthology/P19-1017/

Unsupervised neural machine translation (NMT) has attracted a lot of attention recently. While state-of-the-art methods for unsupervised translation usually perform well between similar languages (e.g., English-German translation), they perform poorly between distant languages, because unsupervised alignment does not work well for distant languages. In this work, we introduce unsupervised pivot translation for distant languages, which translates a language to a distant language through multiple hops, and the unsupervised translation on each hop is relatively easier than the original direct translation. We propose a learning to route (LTR) method to choose the translation path between the source and target languages. LTR is trained on language pairs whose best translation path is available and is applied on the unseen language pairs for path selection. Experiments on 20 languages and 294 distant language pairs demonstrate the advantages of the unsupervised pivot translation for distant languages, as well as the effectiveness of the proposed LTR for path selection. Specifically, in the best case, LTR achieves an improvement of 5.58 BLEU points over the conventional direct unsupervised method.

Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces

- https://www.aclweb.org/anthology/P19-1018/

Recent work on bilingual lexicon induction (BLI) has frequently depended either on aligned bilingual lexicons or on distribution matching, often with an assumption about the isometry of the two spaces. We propose a technique to quantitatively estimate this assumption of the isometry between two embedding spaces and empirically show that this assumption weakens as the languages in question become increasingly etymologically distant. We then propose Bilingual Lexicon Induction with Semi-Supervision (BLISS) — a semi-supervised approach that relaxes the isometric assumption while leveraging both limited aligned bilingual lexicons and a larger set of unaligned word embeddings, as well as a novel hubness filtering technique. Our proposed method obtains state of the art results on 15 of 18 language pairs on the MUSE dataset, and does particularly well when the embedding spaces don't appear to be isometric. In addition, we also show that adding supervision stabilizes the learning procedure, and is effective even with minimal supervision.

An Effective Approach to Unsupervised Machine Translation

- https://www.aclweb.org/anthology/P19-1019/

While machine translation has traditionally relied on large amounts of parallel corpora, a recent research line has managed to train both Neural Machine Translation (NMT) and Statistical Machine Translation (SMT) systems using monolingual corpora only. In this paper, we identify and address several deficiencies of existing unsupervised SMT approaches by exploiting subword information, developing a theoretically well founded unsupervised tuning method, and incorporating a joint refinement procedure. Moreover, we use our improved SMT system to initialize a dual NMT model, which is further fine-tuned through on-the-fly back-translation. Together, we obtain large improvements over the previous state-of-the-art in unsupervised machine translation. For instance, we get 22.5 BLEU points in English-to-German WMT 2014, 5.5 points more than the previous best unsupervised system, and 0.5 points more than the (supervised) shared task winner back in 2014.

Effective Adversarial Regularization for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1020/

A regularization technique based on adversarial perturbation, which was initially developed in the field of image processing, has been successfully applied to text classification tasks and has yielded attractive improvements. We aim to further leverage this promising methodology into more sophisticated and critical neural models in the natural language processing field, i.e., neural machine

translation (NMT) models. However, it is not trivial to apply this methodology to such models. Thus, this paper investigates the effectiveness of several possible configurations of applying the adversarial perturbation and reveals that the adversarial regularization technique can significantly and consistently improve the performance of widely used NMT models, such as LSTM-based and Transformer-based models.

Revisiting Low-Resource Neural Machine Translation: A Case Study

- https://www.aclweb.org/anthology/P19-1021/

It has been shown that the performance of neural machine translation (NMT) drops starkly in low-resource conditions, underperforming phrase-based statistical machine translation (PBSMT) and requiring large amounts of auxiliary data to achieve competitive results. In this paper, we re-assess the validity of these results, arguing that they are the result of lack of system adaptation to low-resource settings. We discuss some pitfalls to be aware of when training low-resource NMT systems, and recent techniques that have shown to be especially helpful in low-resource settings, resulting in a set of best practices for low-resource NMT. In our experiments on German–English with different amounts of IWSLT14 training data, we show that, without the use of any auxiliary monolingual or multilingual data, an optimized NMT system can outperform PBSMT with far less data than previously claimed. We also apply these techniques to a low-resource Korean–English dataset, surpassing previously reported results by 4 BLEU.

Domain Adaptive Inference for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1022/

We investigate adaptive ensemble weighting for Neural Machine Translation, addressing the case of improving performance on a new and potentially unknown domain without sacrificing performance on the original domain. We adapt sequentially across two Spanish-English and three English-German tasks, comparing unregularized fine-tuning, L2 and Elastic Weight Consolidation. We then report a novel scheme for adaptive NMT ensemble decoding by extending Bayesian Interpolation with source information, and report strong improvements across test domains without access to the domain label.

Neural Relation Extraction for Knowledge Base Enrichment

- https://www.aclweb.org/anthology/P19-1023/

We study relation extraction for knowledge base (KB) enrichment. Specifically, we aim to extract entities and their relationships from sentences in the form of triples and map the elements of the extracted triples to an existing KB in an end-to-end manner. Previous studies focus on the extraction itself and rely on Named Entity Disambiguation (NED) to map triples into the KB space. This way, NED errors may cause extraction errors that affect the overall precision and recall.To address this problem, we propose an end-to-end relation extraction model for KB enrichment based on a neural encoder-decoder model. We collect high-quality training data by distant supervision with co-reference resolution and paraphrase detection. We propose an n-gram based attention model that captures multi-word entity names in a sentence. Our model employs jointly learned word and entity embeddings to support named entity disambiguation. Finally, our model uses a modified beam search and a triple classifier to help generate high-quality triples. Our model outperforms state-of-the-art baselines by 15.51% and 8.38% in terms of F1 score on two real-world datasets.

Attention Guided Graph Convolutional Networks for Relation Extraction

- https://www.aclweb.org/anthology/P19-1024/

Dependency trees convey rich structural information that is proven useful for extracting relations among entities in text. However, how to effectively make use of relevant information while ignoring irrelevant information from the dependency trees remains a challenging research question. Existing approaches employing rule based hard-pruning strategies for selecting relevant partial dependency structures may not always yield optimal results. In this work, we propose Attention Guided Graph Convolutional Networks (AGGCNs), a novel model which directly takes full dependency trees as inputs. Our model can be understood as a soft-pruning approach that automatically learns how to selectively attend to the relevant sub-structures useful for the relation extraction task. Extensive results on various tasks including cross-sentence n-ary relation extraction and large-scale sentence-level relation extraction show that our model is able to better leverage the structural information of the full dependency trees, giving significantly better results than previous approaches.

Spatial Aggregation Facilitates Discovery of Spatial Topics

- https://www.aclweb.org/anthology/P19-1025/

Spatial aggregation refers to merging of documents created at the same spatial location. We show that by spatial aggregation of a large collection of documents and applying a traditional topic discovery algorithm on the aggregated data we can efficiently discover spatially distinct topics. By looking at topic discovery through matrix factorization lenses we show that spatial aggregation allows low

rank approximation of the original document-word matrix, in which spatially distinct topics are preserved and non-spatial topics are aggregated into a single topic. Our experiments on synthetic data confirm this observation. Our experiments on 4.7 million tweets collected during the Sandy Hurricane in 2012 show that spatial and temporal aggregation allows rapid discovery of relevant spatial and temporal topics during that period. Our work indicates that different forms of document aggregation might be effective in rapid discovery of various types of distinct topics from large collections of documents.

Relation Embedding with Dihedral Group in Knowledge Graph

- https://www.aclweb.org/anthology/P19-1026/

Link prediction is critical for the application of incomplete knowledge graph (KG) in the downstream tasks. As a family of effective approaches for link predictions, embedding methods try to learn low-rank representations for both entities and relations such that the bilinear form defined therein is a well-behaved scoring function. Despite of their successful performances, existing bilinear forms overlook the modeling of relation compositions, resulting in lacks of interpretability for reasoning on KG. To fulfill this gap, we propose a new model called DihEdral, named after dihedral symmetry group. This new model learns knowledge graph embeddings that can capture relation compositions by nature. Furthermore, our approach models the relation embeddings parametrized by discrete values, thereby decrease the solution space drastically. Our experiments show that DihEdral is able to capture all desired properties such as (skew-) symmetry, inversion and (non-) Abelian composition, and outperforms existing bilinear form based approach and is comparable to or better than deep learning models such as ConvE.

Sequence Tagging with Contextual and Non-Contextual Subword Representations: A Multilingual Evaluation

- https://www.aclweb.org/anthology/P19-1027/

Pretrained contextual and non-contextual subword embeddings have become available in over 250 languages, allowing massively multilingual NLP. However, while there is no dearth of pretrained embeddings, the distinct lack of systematic evaluations makes it difficult for practitioners to choose between them. In this work, we conduct an extensive evaluation comparing non-contextual subword embeddings, namely FastText and BPEmb, and a contextual representation method, namely BERT, on multilingual named entity recognition and part-of-speech tagging. We find that overall, a combination of BERT, BPEmb, and character representations works best across languages and tasks. A more detailed analysis reveals different strengths and weaknesses: Multilingual BERT

performs well in medium- to high-resource languages, but is outperformed by non-contextual subword embeddings in a low-resource setting.

Augmenting Neural Networks with First-order Logic

- https://www.aclweb.org/anthology/P19-1028/

Today, the dominant paradigm for training neural networks involves minimizing task loss on a large dataset. Using world knowledge to inform a model, and yet retain the ability to perform end-to-end training remains an open question. In this paper, we present a novel framework for introducing declarative knowledge to neural network architectures in order to guide training and prediction. Our framework systematically compiles logical statements into computation graphs that augment a neural network without extra learnable parameters or manual redesign. We evaluate our modeling strategy on three tasks: machine comprehension, natural language inference, and text chunking. Our experiments show that knowledge-augmented networks can strongly improve over baselines, especially in low-data regimes.

Self-Regulated Interactive Sequence-to-Sequence Learning

- https://www.aclweb.org/anthology/P19-1029/

Not all types of supervision signals are created equal: Different types of feedback have different costs and effects on learning. We show how self-regulation strategies that decide when to ask for which kind of feedback from a teacher (or from oneself) can be cast as a learning-to-learn problem leading to improved cost-aware sequence-to-sequence learning. In experiments on interactive neural machine translation, we find that the self-regulator discovers an -greedy strategy for the optimal cost-quality trade-off by mixing different feedback types including corrections, error markups, and self-supervision. Furthermore, we demonstrate its robustness under domain shift and identify it as a promising alternative to active learning.

You Only Need Attention to Traverse Trees

- https://www.aclweb.org/anthology/P19-1030/

In recent NLP research, a topic of interest is universal sentence encoding, sentence representations that can be used in any supervised task. At the word sequence level, fully attention-based models suffer from two problems: a quadratic increase in memory consumption with respect to the sentence length and an inability to capture and use syntactic information. Recursive neural nets can extract very good syntactic information by traversing a tree structure. To this

end, we propose Tree Transformer, a model that captures phrase level syntax for constituency trees as well as word-level dependencies for dependency trees by doing recursive traversal only with attention. Evaluation of this model on four tasks gets noteworthy results compared to the standard transformer and LSTM-based models as well as tree-structured LSTMs. Ablation studies to find whether positional information is inherently encoded in the trees and which type of attention is suitable for doing the recursive traversal are provided.

Cross-Domain Generalization of Neural Constituency Parsers

- https://www.aclweb.org/anthology/P19-1031/

Neural parsers obtain state-of-the-art results on benchmark treebanks for constituency parsing—but to what degree do they generalize to other domains? We present three results about the generalization of neural parsers in a zero-shot setting: training on trees from one corpus and evaluating on out-of-domain corpora. First, neural and non-neural parsers generalize comparably to new domains. Second, incorporating pre-trained encoder representations into neural parsers substantially improves their performance across all domains, but does not give a larger relative improvement for out-of-domain treebanks. Finally, despite the rich input representations they learn, neural parsers still benefit from structured output prediction of output trees, yielding higher exact match accuracy and stronger generalization both to larger text spans and to out-of-domain corpora. We analyze generalization on English and Chinese corpora, and in the process obtain state-of-the-art parsing results for the Brown, Genia, and English Web treebanks.

Adaptive Attention Span in Transformers

- https://www.aclweb.org/anthology/P19-1032/

We propose a novel self-attention mechanism that can learn its optimal attention span. This allows us to extend significantly the maximum context size used in Transformer, while maintaining control over their memory footprint and computational time. We show the effectiveness of our approach on the task of character level language modeling, where we achieve state-of-the-art performances on text8 and enwiki8 by using a maximum context of 8k characters.

Neural News Recommendation with Long- and Short-term User Representations

- https://www.aclweb.org/anthology/P19-1033/

Personalized news recommendation is important to help users find their interested news and improve reading experience. A key problem in news recommendation is learning accurate user representations to capture their interests. Users usually have both long-term preferences and short-term interests. However, existing news recommendation methods usually learn single representations of users, which may be insufficient. In this paper, we propose a neural news recommendation approach which can learn both long- and short-term user representations. The core of our approach is a news encoder and a user encoder. In the news encoder, we learn representations of news from their titles and topic categories, and use attention network to select important words. In the user encoder, we propose to learn long-term user representations from the embeddings of their IDs.In addition, we propose to learn short-term user representations from their recently browsed news via GRU network. Besides, we propose two methods to combine long-term and short-term user representations. The first one is using the long-term user representation to initialize the hidden state of the GRU network in short-term user representation. The second one is concatenating both long- and short-term user representations as a unified user vector. Extensive experiments on a real-world dataset show our approach can effectively improve the performance of neural news recommendation.

Automatic Domain Adaptation Outperforms Manual Domain Adaptation for Predicting Financial Outcomes

- https://www.aclweb.org/anthology/P19-1034/

In this paper, we automatically create sentiment dictionaries for predicting financial outcomes. We compare three approaches: (i) manual adaptation of the domain-general dictionary H4N, (ii) automatic adaptation of H4N and (iii) a combination consisting of first manual, then automatic adaptation. In our experiments, we demonstrate that the automatically adapted sentiment dictionary outperforms the previous state of the art in predicting the financial outcomes excess return and volatility. In particular, automatic adaptation performs better than manual adaptation. In our analysis, we find that annotation based on an expert's a priori belief about a word's meaning can be incorrect – annotation should be performed based on the word's contexts in the target domain instead.

Manipulating the Difficulty of C-Tests

- https://www.aclweb.org/anthology/P19-1035/

We propose two novel manipulation strategies for increasing and decreasing the difficulty of C-tests automatically. This is a crucial step towards generating learner-adaptive exercises for self-directed language learning and preparing language assessment tests. To reach the desired difficulty level, we manipulate

the size and the distribution of gaps based on absolute and relative gap difficulty predictions. We evaluate our approach in corpus-based experiments and in a user study with 60 participants. We find that both strategies are able to generate C-tests with the desired difficulty level.

Towards Unsupervised Text Classification Leveraging Experts and Word Embeddings

- https://www.aclweb.org/anthology/P19-1036/

Text classification aims at mapping documents into a set of predefined categories. Supervised machine learning models have shown great success in this area but they require a large number of labeled documents to reach adequate accuracy. This is particularly true when the number of target categories is in the tens or the hundreds. In this work, we explore an unsupervised approach to classify documents into categories simply described by a label. The proposed method is inspired by the way a human proceeds in this situation: It draws on textual similarity between the most relevant words in each document and a dictionary of keywords for each category reflecting its semantics and lexical field. The novelty of our method hinges on the enrichment of the category labels through a combination of human expertise and language models, both generic and domain specific. Our experiments on 5 standard corpora show that the proposed method increases F1-score over relying solely on human expertise and can also be on par with simple supervised approaches. It thus provides a practical alternative to situations where low cost text categorization is needed, as we illustrate with our application to operational risk incidents classification.

Neural Text Simplification of Clinical Letters with a Domain Specific Phrase Table

- https://www.aclweb.org/anthology/P19-1037/

Clinical letters are infamously impenetrable for the lay patient. This work uses neural text simplification methods to automatically improve the understandability of clinical letters for patients. We take existing neural text simplification software and augment it with a new phrase table that links complex medical terminology to simpler vocabulary by mining SNOMED-CT. In an evaluation task using crowdsourcing, we show that the results of our new system are ranked easier to understand (average rank 1.93) than using the original system (2.34) without our phrase table. We also show improvement against baselines including the original text (2.79) and using the phrase table without the neural text simplification software (2.94). Our methods can easily be transferred outside of the clinical domain by using domain-appropriate resources to provide effective neural text simplification for any domain without the need for costly annotation.

What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues

- https://www.aclweb.org/anthology/P19-1038/

Predicting financial risk is an essential task in financial market. Prior research has shown that textual information in a firm's financial statement can be used to predict its stock's risk level. Nowadays, firm CEOs communicate information not only verbally through press releases and financial reports, but also nonverbally through investor meetings and earnings conference calls. There are anecdotal evidences that CEO's vocal features, such as emotions and voice tones, can reveal the firm's performance. However, how vocal features can be used to predict risk levels, and to what extent, is still unknown. To fill the gap, we obtain earnings call audio recordings and textual transcripts for S&P 500 companies in recent years. We propose a multimodal deep regression model (MDRM) that jointly model CEO's verbal (from text) and vocal (from audio) information in a conference call. Empirical results show that our model that jointly considers verbal and vocal features achieves significant and substantial prediction error reduction. We also discuss several interesting findings and the implications to financial markets. The processed earnings conference calls data (text and audio) are released for readers who are interested in reproducing the results or designing trading strategy.

Detecting Concealed Information in Text and Speech

- https://www.aclweb.org/anthology/P19-1039/

Motivated by infamous cheating scandals in the media industry, the wine industry, and political campaigns, we address the problem of detecting concealed information in technical settings. In this work, we explore acoustic-prosodic and linguistic indicators of information concealment by collecting a unique corpus of professionals practicing for oral exams while concealing information. We reveal subtle signs of concealing information in speech and text, compare and contrast them with those in deception detection literature, uncovering the link between concealing information and deception. We then present a series of experiments that automatically detect concealed information from text and speech. We compare the use of acoustic-prosodic, linguistic, and individual feature sets, using different machine learning models. Finally, we present a multi-task learning framework with acoustic, linguistic, and individual features, that outperforms human performance by over 15%.

Evidence-based Trustworthiness

- https://www.aclweb.org/anthology/P19-1040/

The information revolution brought with it information pollution. Information retrieval and extraction help us cope with abundant information from diverse sources. But some sources are of anonymous authorship, and some are of uncertain accuracy, so how can we determine what we should actually believe? Not all information sources are equally trustworthy, and simply accepting the majority view is often wrong. This paper develops a general framework for estimating the trustworthiness of information sources in an environment where multiple sources provide claims and supporting evidence, and each claim can potentially be produced by multiple sources. We consider two settings: one in which information sources directly assert claims, and a more realistic and challenging one, in which claims are inferred from evidence provided by sources, via (possibly noisy) NLP techniques. Our key contribution is to develop a family of probabilistic models that jointly estimate the trustworthiness of sources, and the credibility of claims they assert. This is done while accounting for the (possibly noisy) NLP needed to infer claims from evidence supplied by sources. We evaluate our framework on several datasets, showing strong results and significant improvement over baselines.

Disentangled Representation Learning for Non-Parallel Text Style Transfer

- https://www.aclweb.org/anthology/P19-1041/

This paper tackles the problem of disentangling the latent representations of style and content in language models. We propose a simple yet effective approach, which incorporates auxiliary multi-task and adversarial objectives, for style prediction and bag-of-words prediction, respectively. We show, both qualitatively and quantitatively, that the style and content are indeed disentangled in the latent space. This disentangled latent representation learning can be applied to style transfer on non-parallel corpora. We achieve high performance in terms of transfer accuracy, content preservation, and language fluency, in comparison to various previous approaches.

Cross-Sentence Grammatical Error Correction

- https://www.aclweb.org/anthology/P19-1042/

Automatic grammatical error correction (GEC) research has made remarkable progress in the past decade. However, all existing approaches to GEC correct errors by considering a single sentence alone and ignoring crucial cross-sentence context. Some errors can only be corrected reliably using cross-sentence context and models can also benefit from the additional contextual information in correcting other errors. In this paper, we address this serious limitation of existing approaches and improve strong neural encoder-decoder models by appropriately

modeling wider contexts. We employ an auxiliary encoder that encodes previous sentences and incorporate the encoding in the decoder via attention and gating mechanisms. Our approach results in statistically significant improvements in overall GEC performance over strong baselines across multiple test sets. Analysis of our cross-sentence GEC model on a synthetic dataset shows high performance in verb tense corrections that require cross-sentence context.

This Email Could Save Your Life: Introducing the Task of Email Subject Line Generation

- https://www.aclweb.org/anthology/P19-1043/

Given the overwhelming number of emails, an effective subject line becomes essential to better inform the recipient of the email's content. In this paper, we propose and study the task of email subject line generation: automatically generating an email subject line from the email body. We create the first dataset for this task and find that email subject line generation favor extremely ive summary which differentiates it from news headline generation or news single document summarization. We then develop a novel deep learning method and compare it to several baselines as well as recent state-of-the-art text summarization systems. We also investigate the efficacy of several automatic metrics based on correlations with human judgments and propose a new automatic evaluation metric. Our system outperforms competitive baselines given both automatic and human evaluations. To our knowledge, this is the first work to tackle the problem of effective email subject line generation.

Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change

- https://www.aclweb.org/anthology/P19-1044/

State-of-the-art models of lexical semantic change detection suffer from noise stemming from vector space alignment. We have empirically tested the Temporal Referencing method for lexical semantic change and show that, by avoiding alignment, it is less affected by this noise. We show that, trained on a diachronic corpus, the skip-gram with negative sampling architecture with temporal referencing outperforms alignment models on a synthetic task as well as a manual testset. We introduce a principled way to simulate lexical semantic change and systematically control for possible biases.

Adversarial Attention Modeling for Multi-dimensional Emotion Regression

- https://www.aclweb.org/anthology/P19-1045/

In this paper, we propose a neural network-based approach, namely Adversarial Attention Network, to the task of multi-dimensional emotion regression, which automatically rates multiple emotion dimension scores for an input text. Especially, to determine which words are valuable for a particular emotion dimension, an attention layer is trained to weight the words in an input sequence. Furthermore, adversarial training is employed between two attention layers to learn better word weights via a discriminator. In particular, a shared attention layer is incorporated to learn public word weights between two emotion dimensions. Empirical evaluation on the EMOBANK corpus shows that our approach achieves notable improvements in r-values on both EMOBANK Reader's and Writer's multi-dimensional emotion regression tasks in all domains over the state-of-the-art baselines.

Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing

- https://www.aclweb.org/anthology/P19-1046/

We propose a general strategy named 'divide, conquer and combine' for multimodal fusion. Instead of directly fusing features at holistic level, we conduct fusion hierarchically so that both local and global interactions are considered for a comprehensive interpretation of multimodal embeddings. In the 'divide' and 'conquer' stages, we conduct local fusion by exploring the interaction of a portion of the aligned feature vectors across various modalities lying within a sliding window, which ensures that each part of multimodal embeddings are explored sufficiently. On its basis, global fusion is conducted in the 'combine' stage to explore the interconnection across local interactions, via an Attentive Bi-directional Skip-connected LSTM that directly connects distant local interactions and integrates two levels of attention mechanism. In this way, local interactions can exchange information sufficiently and thus obtain an overall view of multimodal information. Our method achieves state-of-the-art performance on multimodal affective computing with higher efficiency.

Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls

- https://www.aclweb.org/anthology/P19-1047/

Every fiscal quarter, companies hold earnings calls in which company executives respond to questions from analysts. After these calls, analysts often change their price target recommendations, which are used in equity re- search reports to help investors make deci- sions. In this paper, we examine analysts' decision making behavior as it pertains to the language content of earnings calls. We

identify a set of 20 pragmatic features of analysts' questions which we correlate with analysts' pre-call investor recommendations. We also analyze the degree to which semantic and pragmatic features from an earnings call complement market data in predicting analysts' post-call changes in price targets. Our results show that earnings calls are moderately predictive of analysts' decisions even though these decisions are influenced by a number of other factors including private communication with company executives and market conditions. A breakdown of model errors indicates disparate performance on calls from different market sectors.

An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis

- https://www.aclweb.org/anthology/P19-1048/

Aspect-based sentiment analysis produces a list of aspect terms and their corresponding sentiments for a natural language sentence. This task is usually done in a pipeline manner, with aspect term extraction performed first, followed by sentiment predictions toward the extracted aspect terms. While easier to develop, such an approach does not fully exploit joint information from the two subtasks and does not use all available sources of training information that might be helpful, such as document-level labeled sentiment corpus. In this paper, we propose an interactive multi-task learning network (IMN) which is able to jointly learn multiple related tasks simultaneously at both the token level as well as the document level. Unlike conventional multi-task learning methods that rely on learning common features for the different tasks, IMN introduces a message passing architecture where information is iteratively passed to different tasks through a shared set of latent variables. Experimental results demonstrate superior performance of the proposed method against multiple baselines on three benchmark datasets.

Decompositional Argument Mining: A General Purpose Approach for Argument Graph Construction

- https://www.aclweb.org/anthology/P19-1049/

This work presents an approach decomposing propositions into four functional components and identify the patterns linking those components to determine argument structure. The entities addressed by a proposition are target concepts and the features selected to make a point about the target concepts are aspects. A line of reasoning is followed by providing evidence for the points made about the target concepts via aspects. Opinions on target concepts and opinions on aspects are used to support or attack the ideas expressed by target concepts and aspects. The relations between aspects, target concepts, opinions on target

concepts and aspects are used to infer the argument relations. Propositions are connected iteratively to form a graph structure. The approach is generic in that it is not tuned for a specific corpus and evaluated on three different corpora from the literature: AAEC, AMT, US2016G1tv and achieved an F score of 0.79, 0.77 and 0.64, respectively.

MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations

- https://www.aclweb.org/anthology/P19-1050/

Emotion recognition in conversations is a challenging task that has recently gained popularity due to its potential applications. Until now, however, a large-scale multimodal multi-party emotional conversational database containing more than two speakers per dialogue was missing. Thus, we propose the Multimodal EmotionLines Dataset (MELD), an extension and enhancement of EmotionLines. MELD contains about 13,000 utterances from 1,433 dialogues from the TV-series Friends. Each utterance is annotated with emotion and sentiment labels, and encompasses audio, visual and textual modalities. We propose several strong multimodal baselines and show the importance of contextual and multimodal information for emotion recognition in conversations. The full dataset is available for use at http://affective-meld.github.io.

Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification

- https://www.aclweb.org/anthology/P19-1051/

Open-domain targeted sentiment analysis aims to detect opinion targets along with their sentiment polarities from a sentence. Prior work typically formulates this task as a sequence tagging problem. However, such formulation suffers from problems such as huge search space and sentiment inconsistency. To address these problems, we propose a span-based extract-then-classify framework, where multiple opinion targets are directly extracted from the sentence under the supervision of target span boundaries, and corresponding polarities are then classified using their span representations. We further investigate three approaches under this framework, namely the pipeline, joint, and collapsed models. Experiments on three benchmark datasets show that our approach consistently outperforms the sequence tagging baseline. Moreover, we find that the pipeline model achieves the best performance compared with the other two models.

Transfer Capsule Network for Aspect Level Sentiment Classification

- https://www.aclweb.org/anthology/P19-1052/

Aspect-level sentiment classification aims to determine the sentiment polarity of a sentence towards an aspect. Due to the high cost in annotation, the lack of aspect-level labeled data becomes a major obstacle in this area. On the other hand, document-level labeled data like reviews are easily accessible from online websites. These reviews encode sentiment knowledge in abundant contexts. In this paper, we propose a Transfer Capsule Network (TransCap) model for transferring document-level knowledge to aspect-level sentiment classification. To this end, we first develop an aspect routing approach to encapsulate the sentence-level semantic representations into semantic capsules from both the aspect-level and document-level data. We then extend the dynamic routing approach to adaptively couple the semantic capsules with the class capsules under the transfer learning framework. Experiments on SemEval datasets demonstrate the effectiveness of TransCap.

Progressive Self-Supervised Attention Learning for Aspect-Level Sentiment Analysis

- https://www.aclweb.org/anthology/P19-1053/

In aspect-level sentiment classification (ASC), it is prevalent to equip dominant neural models with attention mechanisms, for the sake of acquiring the importance of each context word on the given aspect. However, such a mechanism tends to excessively focus on a few frequent words with sentiment polarities, while ignoring infrequent ones. In this paper, we propose a progressive self-supervised attention learning approach for neural ASC models, which automatically mines useful attention supervision information from a training corpus to refine attention mechanisms. Specifically, we iteratively conduct sentiment predictions on all training instances. Particularly, at each iteration, the context word with the maximum attention weight is extracted as the one with active/misleading influence on the correct/incorrect prediction of every instance, and then the word itself is masked for subsequent iterations. Finally, we augment the conventional training objective with a regularization term, which enables ASC models to continue equally focusing on the extracted active context words while decreasing weights of those misleading ones. Experimental results on multiple datasets show that our proposed approach yields better attention mechanisms, leading to substantial improvements over the two state-of-the-art neural ASC models. Source code and trained models are available at https://github.com/DeepLearnXMU/PSSAttention.

Classification and Clustering of Arguments with Contextualized Word Embeddings

- https://www.aclweb.org/anthology/P19-1054/

We experiment with two recent contextualized word embedding methods (ELMo and BERT) in the context of open-domain argument search. For the first time, we show how to leverage the power of contextualized word embeddings to classify and cluster topic-dependent arguments, achieving impressive results on both tasks and across multiple datasets. For argument classification, we improve the state-of-the-art for the UKP Sentential Argument Mining Corpus by 20.8 percentage points and for the IBM Debater - Evidence Sentences dataset by 7.4 percentage points. For the understudied task of argument clustering, we propose a pre-training step which improves by 7.8 percentage points over strong baselines on a novel dataset, and by 12.3 percentage points for the Argument Facet Similarity (AFS) Corpus.

Sentiment Tagging with Partial Labels using Modular Architectures

- https://www.aclweb.org/anthology/P19-1055/

Many NLP learning tasks can be decomposed into several distinct sub-tasks, each associated with a partial label. In this paper we focus on a popular class of learning problems, sequence prediction applied to several sentiment analysis tasks, and suggest a modular learning approach in which different sub-tasks are learned using separate functional modules, combined to perform the final task while sharing information. Our experiments show this approach helps constrain the learning process and can alleviate some of the supervision efforts.

DOER: Dual Cross-Shared RNN for Aspect Term-Polarity Co-Extraction

- https://www.aclweb.org/anthology/P19-1056/

This paper focuses on two related subtasks of aspect-based sentiment analysis, namely aspect term extraction and aspect sentiment classification, which we call aspect term-polarity co-extraction. The former task is to extract aspects of a product or service from an opinion document, and the latter is to identify the polarity expressed in the document about these extracted aspects. Most existing algorithms address them as two separate tasks and solve them one by one, or only perform one task, which can be complicated for real applications. In this paper, we treat these two tasks as two sequence labeling problems and propose a novel Dual crOss-sharEd RNN framework (DOER) to generate all aspect term-polarity pairs of the input sentence simultaneously. Specifically, DOER

involves a dual recurrent neural network to extract the respective representation of each task, and a cross-shared unit to consider the relationship between them. Experimental results demonstrate that the proposed framework outperforms state-of-the-art baselines on three benchmark datasets.

A Corpus for Modeling User and Language Effects in Argumentation on Online Debating

- https://www.aclweb.org/anthology/P19-1057/

Existing argumentation datasets have succeeded in allowing researchers to develop computational methods for analyzing the content, structure and linguistic features of argumentative text. They have been much less successful in fostering studies of the effect of "user" traits — characteristics and beliefs of the participants — on the debate/argument outcome as this type of user information is generally not available. This paper presents a dataset of 78,376 debates generated over a 10-year period along with surprisingly comprehensive participant profiles. We also complete an example study using the dataset to analyze the effect of selected user traits on the debate outcome in comparison to the linguistic features typically employed in studies of this kind.

Topic Tensor Network for Implicit Discourse Relation Recognition in Chinese

- https://www.aclweb.org/anthology/P19-1058/

In the literature, most of the previous studies on English implicit discourse relation recognition only use sentence-level representations, which cannot provide enough semantic information in Chinese due to its unique paratactic characteristics. In this paper, we propose a topic tensor network to recognize Chinese implicit discourse relations with both sentence-level and topic-level representations. In particular, besides encoding arguments (discourse units) using a gated convolutional network to obtain sentence-level representations, we train a simplified topic model to infer the latent topic-level representations. Moreover, we feed the two pairs of representations to two factored tensor networks, respectively, to capture both the sentence-level interactions and topic-level relevance using multi-slice tensors. Experimentation on CDTB, a Chinese discourse corpus, shows that our proposed model significantly outperforms several state-of-the-art baselines in both micro and macro F1-scores.

Learning from Omission

- https://www.aclweb.org/anthology/P19-1059/

Pragmatic reasoning allows humans to go beyond the literal meaning when interpret- ing language in context. Previous work has shown that such reasoning can improve the performance of already-trained language understanding systems. Here, we explore whether pragmatic reasoning during training can improve the quality of learned meanings. Our experiments on reference game data show that end-to-end pragmatic training produces more accurate utterance interpretation models, especially when data is sparse and language is complex.

Multi-Task Learning for Coherence Modeling

- https://www.aclweb.org/anthology/P19-1060/

We address the task of assessing discourse coherence, an aspect of text quality that is essential for many NLP tasks, such as summarization and language assessment. We propose a hierarchical neural network trained in a multi-task fashion that learns to predict a document-level coherence score (at the network's top layers) along with word-level grammatical roles (at the bottom layers), taking advantage of inductive transfer between the two tasks. We assess the extent to which our framework generalizes to different domains and prediction tasks, and demonstrate its effectiveness not only on standard binary evaluation coherence tasks, but also on real-world tasks involving the prediction of varying degrees of coherence, achieving a new state of the art.

Data Programming for Learning Discourse Structure

- https://www.aclweb.org/anthology/P19-1061/

This paper investigates the advantages and limits of data programming for the task of learning discourse structure. The data programming paradigm implemented in the Snorkel framework allows a user to label training data using expert-composed heuristics, which are then transformed via the "generative step" into probability distributions of the class labels given the training candidates. These results are later generalized using a discriminative model. Snorkel's attractive promise to create a large amount of annotated data from a smaller set of training data by unifying the output of a set of heuristics has yet to be used for computationally difficult tasks, such as that of discourse attachment, in which one must decide where a given discourse unit attaches to other units in a text in order to form a coherent discourse structure. Although approaching this problem using Snorkel requires significant modifications to the structure of the heuristics, we show that weak supervision methods can be more than competitive with classical supervised learning approaches to the attachment problem.

Evaluating Discourse in Structured Text Representations

- https://www.aclweb.org/anthology/P19-1062/

Discourse structure is integral to understanding a text and is helpful in many NLP tasks. Learning latent representations of discourse is an attractive alternative to acquiring expensive labeled discourse data. Liu and Lapata (2018) propose a structured attention mechanism for text classification that derives a tree over a text, akin to an RST discourse tree. We examine this model in detail, and evaluate on additional discourse-relevant tasks and datasets, in order to assess whether the structured attention improves performance on the end task and whether it captures a text's discourse structure. We find the learned latent trees have little to no structure and instead focus on lexical cues; even after obtaining more structured trees with proposed model modifications, the trees are still far from capturing discourse structure when compared to discourse dependency trees from an existing discourse parser. Finally, ablation studies show the structured attention provides little benefit, sometimes even hurting performance.

Know What You Don't Know: Modeling a Pragmatic Speaker that Refers to Objects of Unknown Categories

- https://www.aclweb.org/anthology/P19-1063/

Zero-shot learning in Language & Vision is the task of correctly labelling (or naming) objects of novel categories. Another strand of work in L&V aims at pragmatically informative rather than "correct" object descriptions, e.g. in reference games. We combine these lines of research and model zero-shot reference games, where a speaker needs to successfully refer to a novel object in an image. Inspired by models of "rational speech acts", we extend a neural generator to become a pragmatic speaker reasoning about uncertain object categories. As a result of this reasoning, the generator produces fewer nouns and names of distractor categories as compared to a literal speaker. We show that this conversational strategy for dealing with novel objects often improves communicative success, in terms of resolution accuracy of an automatic listener.

End-to-end Deep Reinforcement Learning Based Coreference Resolution

- https://www.aclweb.org/anthology/P19-1064/

Recent neural network models have significantly advanced the task of coreference resolution. However, current neural coreference models are usually trained with heuristic loss functions that are computed over a sequence of local decisions. In this paper, we introduce an end-to-end reinforcement learning based

coreference resolution model to directly optimize coreference evaluation metrics. Specifically, we modify the state-of-the-art higher-order mention ranking approach in Lee et al. (2018) to a reinforced policy gradient model by incorporating the reward associated with a sequence of coreference linking actions. Furthermore, we introduce maximum entropy regularization for adequate exploration to prevent the model from prematurely converging to a bad local optimum. Our proposed model achieves new state-of-the-art performance on the English OntoNotes v5.0 benchmark.

Implicit Discourse Relation Identification for Open-domain Dialogues

- https://www.aclweb.org/anthology/P19-1065/

Discourse relation identification has been an active area of research for many years, and the challenge of identifying implicit relations remains largely an unsolved task, especially in the context of an open-domain dialogue system. Previous work primarily relies on a corpora of formal text which is inherently non-dialogic, i.e., news and journals. This data however is not suitable to handle the nuances of informal dialogue nor is it capable of navigating the plethora of valid topics present in open-domain dialogue. In this paper, we designed a novel discourse relation identification pipeline specifically tuned for open-domain dialogue systems. We firstly propose a method to automatically extract the implicit discourse relation argument pairs and labels from a dataset of dialogic turns, resulting in a novel corpus of discourse relation pairs; the first of its kind to attempt to identify the discourse relations connecting the dialogic turns in open-domain discourse. Moreover, we have taken the first steps to leverage the dialogue features unique to our task to further improve the identification of such relations by performing feature ablation and incorporating dialogue features to enhance the state-of-the-art model.

Coreference Resolution with Entity Equalization

- https://www.aclweb.org/anthology/P19-1066/

A key challenge in coreference resolution is to capture properties of entity clusters, and use those in the resolution process. Here we provide a simple and effective approach for achieving this, via an "Entity Equalization" mechanism. The Equalization approach represents each mention in a cluster via an approximation of the sum of all mentions in the cluster. We show how this can be done in a fully differentiable end-to-end manner, thus enabling high-order inferences in the resolution process. Our approach, which also employs BERT embeddings, results in new state-of-the-art results on the CoNLL-2012 coreference resolution task, improving average F1 by 3.6%.

A Cross-Domain Transferable Neural Coherence Model

- https://www.aclweb.org/anthology/P19-1067/

Coherence is an important aspect of text quality and is crucial for ensuring its readability. One important limitation of existing coherence models is that training on one domain does not easily generalize to unseen categories of text. Previous work advocates for generative models for cross-domain generalization, because for discriminative models, the space of incoherent sentence orderings to discriminate against during training is prohibitively large. In this work, we propose a local discriminative neural model with a much smaller negative sampling space that can efficiently learn against incorrect orderings. The proposed coherence model is simple in structure, yet it significantly outperforms previous state-of-art methods on a standard benchmark dataset on the Wall Street Journal corpus, as well as in multiple new challenging settings of transfer to unseen categories of discourse on Wikipedia articles.

MOROCO: The Moldavian and Romanian Dialectal Corpus

- https://www.aclweb.org/anthology/P19-1068/

In this work, we introduce the MOldavian and ROmanian Dialectal COrpus (MOROCO), which is freely available for download at https://github.com/butnaruandrei/MOROCO. The corpus contains 33564 samples of text (with over 10 million tokens) collected from the news domain. The samples belong to one of the following six topics: culture, finance, politics, science, sports and tech. The data set is divided into 21719 samples for training, 5921 samples for validation and another 5924 samples for testing. For each sample, we provide corresponding dialectal and category labels. This allows us to perform empirical studies on several classification tasks such as (i) binary discrimination of Moldavian versus Romanian text samples, (ii) intra-dialect multi-class categorization by topic and (iii) cross-dialect multi-class categorization by topic. We perform experiments using a shallow approach based on string kernels, as well as a novel deep approach based on character-level convolutional neural networks containing Squeeze-and-Excitation blocks. We also present and analyze the most discriminative features of our best performing model, before and after named entity removal.

Just "OneSeC" for Producing Multilingual Sense-Annotated Data

- https://www.aclweb.org/anthology/P19-1069/

The well-known problem of knowledge acquisition is one of the biggest issues in Word Sense Disambiguation (WSD), where annotated data are still scarce in English and almost absent in other languages. In this paper we formulate the assumption of One Sense per Wikipedia Category and present OneSeC, a language-independent method for the automatic extraction of hundreds of thousands of sentences in which a target word is tagged with its meaning. Our automatically-generated data consistently lead a supervised WSD model to state-of-the-art performance when compared with other automatic and semi-automatic methods. Moreover, our approach outperforms its competitors on multilingual and domain-specific settings, where it beats the existing state of the art on all languages and most domains. All the training data are available for research purposes at http://trainomatic.org/onesec.

How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions

- https://www.aclweb.org/anthology/P19-1070/

Cross-lingual word embeddings (CLEs) facilitate cross-lingual transfer of NLP models. Despite their ubiquitous downstream usage, increasingly popular projection-based CLE models are almost exclusively evaluated on bilingual lexicon induction (BLI). Even the BLI evaluations vary greatly, hindering our ability to correctly interpret performance and properties of different CLE models. In this work, we take the first step towards a comprehensive evaluation of CLE models: we thoroughly evaluate both supervised and unsupervised CLE models, for a large number of language pairs, on BLI and three downstream tasks, providing new insights concerning the ability of cutting-edge CLE models to support cross-lingual NLP. We empirically demonstrate that the performance of CLE models largely depends on the task at hand and that optimizing CLE models for BLI may hurt downstream performance. We indicate the most robust supervised and unsupervised CLE models and emphasize the need to reassess simple baselines, which still display competitive performance across the board. We hope our work catalyzes further research on CLE evaluation and model analysis.

SP-10K: A Large-scale Evaluation Set for Selectional Preference Acquisition

- https://www.aclweb.org/anthology/P19-1071/

Selectional Preference (SP) is a commonly observed language phenomenon and proved to be useful in many natural language processing tasks. To provide a better evaluation method for SP models, we introduce SP-10K, a large-scale

evaluation set that provides human ratings for the plausibility of 10,000 SP pairs over five SP relations, covering 2,500 most frequent verbs, nouns, and adjectives in American English. Three representative SP acquisition methods based on pseudo-disambiguation are evaluated with SP-10K. To demonstrate the importance of our dataset, we investigate the relationship between SP-10K and the commonsense knowledge in ConceptNet5 and show the potential of using SP to represent the commonsense knowledge. We also use the Winograd Schema Challenge to prove that the proposed new SP relations are essential for the hard pronoun coreference resolution problem.

A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains

- https://www.aclweb.org/anthology/P19-1072/

We perform an interdisciplinary large-scale evaluation for detecting lexical semantic divergences in a diachronic and in a synchronic task: semantic sense changes across time, and semantic sense changes across domains. Our work addresses the superficialness and lack of comparison in assessing models of diachronic lexical change, by bringing together and extending benchmark models on a common state-of-the-art evaluation task. In addition, we demonstrate that the same evaluation task and modelling approaches can successfully be utilised for the synchronic detection of domain-specific sense divergences in the field of term extraction.

Errudite: Scalable, Reproducible, and Testable Error Analysis

- https://www.aclweb.org/anthology/P19-1073/

Though error analysis is crucial to understanding and improving NLP models, the common practice of manual, subjective categorization of a small sample of errors can yield biased and incomplete conclusions. This paper codifies model and task agnostic principles for informative error analysis, and presents Errudite, an interactive tool for better supporting this process. First, error groups should be precisely defined for reproducibility; Errudite supports this with an expressive domain-specific language. Second, to avoid spurious conclusions, a large set of instances should be analyzed, including both positive and negative examples; Errudite enables systematic grouping of relevant instances with filtering queries. Third, hypotheses about the cause of errors should be explicitly tested; Errudite supports this via automated counterfactual rewriting. We validate our approach with a user study, finding that Errudite (1) enables users to perform high quality and reproducible error analyses with less effort, (2) reveals

substantial ambiguities in prior published error analyses practices, and (3) enhances the error analysis experience by allowing users to test and revise prior beliefs.

DocRED: A Large-Scale Document-Level Relation Extraction Dataset

- https://www.aclweb.org/anthology/P19-1074/

Multiple entities in a document generally exhibit complex inter-sentence relations, and cannot be well handled by existing relation extraction (RE) methods that typically focus on extracting intra-sentence relations for single entity pairs. In order to accelerate the research on document-level RE, we introduce DocRED, a new dataset constructed from Wikipedia and Wikidata with three features: (1) DocRED annotates both named entities and relations, and is the largest human-annotated dataset for document-level RE from plain text; (2) DocRED requires reading multiple sentences in a document to extract entities and infer their relations by synthesizing all information of the document; (3) along with the human-annotated data, we also offer large-scale distantly supervised data, which enables DocRED to be adopted for both supervised and weakly supervised scenarios. In order to verify the challenges of document-level RE, we implement recent state-of-the-art methods for RE and conduct a thorough evaluation of these methods on DocRED. Empirical results show that DocRED is challenging for existing RE methods, which indicates that document-level RE remains an open problem and requires further efforts. Based on the detailed analysis on the experiments, we discuss multiple promising directions for future research. We make DocRED and the code for our baselines publicly available at https://github.com/thunlp/DocRED.

ChID: A Large-scale Chinese IDiom Dataset for Cloze Test

- https://www.aclweb.org/anthology/P19-1075/

Cloze-style reading comprehension in Chinese is still limited due to the lack of various corpora. In this paper we propose a large-scale Chinese cloze test dataset ChID, which studies the comprehension of idiom, a unique language phenomenon in Chinese. In this corpus, the idioms in a passage are replaced by blank symbols and the correct answer needs to be chosen from well-designed candidate idioms. We carefully study how the design of candidate idioms and the representation of idioms affect the performance of state-of-the-art models. Results show that the machine accuracy is substantially worse than that of human, indicating a large space for further research.

Automatic Evaluation of Local Topic Quality

- https://www.aclweb.org/anthology/P19-1076/

Topic models are typically evaluated with respect to the global topic distributions that they generate, using metrics such as coherence, but without regard to local (token-level) topic assignments. Token-level assignments are important for downstream tasks such as classification. Even recent models, which aim to improve the quality of these token-level topic assignments, have been evaluated only with respect to global metrics. We propose a task designed to elicit human judgments of token-level topic assignments. We use a variety of topic model types and parameters and discover that global metrics agree poorly with human assignments. Since human evaluation is expensive we propose a variety of automated metrics to evaluate topic models at a local level. Finally, we correlate our proposed metrics with human judgments from the task on several datasets. We show that an evaluation based on the percent of topic switches correlates most strongly with human judgment of local topic quality. We suggest that this new metric, which we call consistency, be adopted alongside global metrics such as topic coherence when evaluating new topic models.

Crowdsourcing and Aggregating Nested Markable Annotations

- https://www.aclweb.org/anthology/P19-1077/

One of the key steps in language resource creation is the identification of the text segments to be annotated, or markables, which depending on the task may vary from nominal chunks for named entity resolution to (potentially nested) noun phrases in coreference resolution (or mentions) to larger text segments in text segmentation. Markable identification is typically carried out semi-automatically, by running a markable identifier and correcting its output by hand–which is increasingly done via annotators recruited through crowdsourcing and aggregating their responses. In this paper, we present a method for identifying markables for coreference annotation that combines high-performance automatic markable detectors with checking with a Game-With-A-Purpose (GWAP) and aggregation using a Bayesian annotation model. The method was evaluated both on news data and data from a variety of other genres and results in an improvement on F1 of mention boundaries of over seven percentage points when compared with a state-of-the-art, domain-independent automatic mention detector, and almost three points over an in-domain mention detector. One of the key contributions of our proposal is its applicability to the case in which markables are nested, as is the case with coreference markables; but the GWAP and several of the proposed markable detectors are task and language-independent and are thus applicable to a variety of other annotation scenarios.

Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems

- https://www.aclweb.org/anthology/P19-1078/

Over-dependence on domain ontology and lack of sharing knowledge across domains are two practical and yet less studied problems of dialogue state tracking. Existing approaches generally fall short when tracking unknown slot values during inference and often have difficulties in adapting to new domains. In this paper, we propose a Transferable Dialogue State Generator (TRADE) that generates dialogue states from utterances using copy mechanism, facilitating transfer when predicting (domain, slot, value) triplets not encountered during training. Our model is composed of an utterance encoder, a slot gate, and a state generator, which are shared across domains. Empirical results demonstrate that TRADE achieves state-of-the-art 48.62% joint goal accuracy for the five domains of MultiWOZ, a human-human dialogue dataset. In addition, we show the transferring ability by simulating zero-shot and few-shot dialogue state tracking for unseen domains. TRADE achieves 60.58% joint goal accuracy in one of the zero-shot domains, and is able to adapt to few-shot cases without forgetting already trained domains.

Multi-Task Networks with Universe, Group, and Task Feature Learning

- https://www.aclweb.org/anthology/P19-1079/

We present methods for multi-task learning that take advantage of natural groupings of related tasks. Task groups may be defined along known properties of the tasks, such as task domain or language. Such task groups represent supervised information at the inter-task level and can be encoded into the model. We investigate two variants of neural network architectures that accomplish this, learning different feature spaces at the levels of individual tasks, task groups, as well as the universe of all tasks: (1) parallel architectures encode each input simultaneously into feature spaces at different levels; (2) serial architectures encode each input successively into feature spaces at different levels in the task hierarchy. We demonstrate the methods on natural language understanding (NLU) tasks, where a grouping of tasks into different task domains leads to improved performance on ATIS, Snips, and a large in-house dataset.

Constrained Decoding for Neural NLG from Compositional Representations in Task-Oriented Dialogue

- https://www.aclweb.org/anthology/P19-1080/

Generating fluent natural language responses from structured semantic representations is a critical step in task-oriented conversational systems. Avenues like the E2E NLG Challenge have encouraged the development of neural approaches, particularly sequence-to-sequence (Seq2Seq) models for this problem. The semantic representations used, however, are often underspecified, which places a higher burden on the generation model for sentence planning, and also limits the extent to which generated responses can be controlled in a live system. In this paper, we (1) propose using tree-structured semantic representations, like those used in traditional rule-based NLG systems, for better discourse-level structuring and sentence-level planning; (2) introduce a challenging dataset using this representation for the weather domain; (3) introduce a constrained decoding approach for Seq2Seq models that leverages this representation to improve semantic correctness; and (4) demonstrate promising results on our dataset and the E2E dataset.

OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs

- https://www.aclweb.org/anthology/P19-1081/

We study a conversational reasoning model that strategically traverses through a large-scale common fact knowledge graph (KG) to introduce engaging and contextually diverse entities and attributes. For this study, we collect a new Open-ended Dialog <-> KG parallel corpus called OpenDialKG, where each utterance from 15K human-to-human role-playing dialogs is manually annotated with ground-truth reference to corresponding entities and paths from a large-scale KG with 1M+ facts. We then propose the DialKG Walker model that learns the symbolic transitions of dialog contexts as structured traversals over KG, and predicts natural entities to introduce given previous dialog contexts via a novel domain-agnostic, attention-based graph path decoder. Automatic and human evaluations show that our model can retrieve more natural and human-like responses than the state-of-the-art baselines or rule-based models, in both in-domain and cross-domain tasks. The proposed model also generates a KG walk path for each entity retrieved, providing a natural way to explain conversational reasoning.

Coupling Retrieval and Meta-Learning for Context-Dependent Semantic Parsing

- https://www.aclweb.org/anthology/P19-1082/

In this paper, we present an approach to incorporate retrieved datapoints as supporting evidence for context-dependent semantic parsing, such as generating source code conditioned on the class environment. Our approach naturally

combines a retrieval model and a meta-learner, where the former learns to find similar datapoints from the training data, and the latter considers retrieved datapoints as a pseudo task for fast adaptation. Specifically, our retriever is a context-aware encoder-decoder model with a latent variable which takes context environment into consideration, and our meta-learner learns to utilize retrieved datapoints in a model-agnostic meta-learning paradigm for fast adaptation. We conduct experiments on CONCODE and CSQA datasets, where the context refers to class environment in JAVA codes and conversational history, respectively. We use sequence-to-action model as the base semantic parser, which performs the state-of-the-art accuracy on both datasets. Results show that both the context-aware retriever and the meta-learning strategy improve accuracy, and our approach performs better than retrieve-and-edit baselines.

Knowledge-aware Pronoun Coreference Resolution

- https://www.aclweb.org/anthology/P19-1083/

Resolving pronoun coreference requires knowledge support, especially for particular domains (e.g., medicine). In this paper, we explore how to leverage different types of knowledge to better resolve pronoun coreference with a neural model. To ensure the generalization ability of our model, we directly incorporate knowledge in the format of triplets, which is the most common format of modern knowledge graphs, instead of encoding it with features or rules as that in conventional approaches. Moreover, since not all knowledge is helpful in certain contexts, to selectively use them, we propose a knowledge attention module, which learns to select and use informative knowledge based on contexts, to enhance our model. Experimental results on two datasets from different domains prove the validity and effectiveness of our model, where it outperforms state-of-the-art baselines by a large margin. Moreover, since our model learns to use external knowledge rather than only fitting the training data, it also demonstrates superior performance to baselines in the cross-domain setting.

Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference

- https://www.aclweb.org/anthology/P19-1084/

Natural Language Inference (NLI) datasets often contain hypothesis-only biases—artifacts that allow models to achieve non-trivial performance without learning whether a premise entails a hypothesis. We propose two probabilistic methods to build models that are more robust to such biases and better transfer across datasets. In contrast to standard approaches to NLI, our methods predict the probability of a premise given a hypothesis and NLI label, discouraging models from ignoring the premise. We evaluate our methods on synthetic and existing NLI datasets by training on datasets containing biases

and testing on datasets containing no (or different) hypothesis-only biases. Our results indicate that these methods can make NLI models more robust to dataset-specific artifacts, transferring better than a baseline architecture in 9 out of 12 NLI datasets. Additionally, we provide an extensive analysis of the interplay of our methods with known biases in NLI datasets, as well as the effects of encouraging models to ignore biases and fine-tuning on target datasets.

GEAR: Graph-based Evidence Aggregating and Reasoning for Fact Verification

- https://www.aclweb.org/anthology/P19-1085/

Fact verification (FV) is a challenging task which requires to retrieve relevant evidence from plain text and use the evidence to verify given claims. Many claims require to simultaneously integrate and reason over several pieces of evidence for verification. However, previous work employs simple models to extract information from evidence without letting evidence communicate with each other, e.g., merely concatenate the evidence for processing. Therefore, these methods are unable to grasp sufficient relational and logical information among the evidence. To alleviate this issue, we propose a graph-based evidence aggregating and reasoning (GEAR) framework which enables information to transfer on a fully-connected evidence graph and then utilizes different aggregators to collect multi-evidence information. We further employ BERT, an effective pre-trained language representation model, to improve the performance. Experimental results on a large-scale benchmark dataset FEVER have demonstrated that GEAR could leverage multi-evidence information for FV and thus achieves the promising result with a test FEVER score of 67.10%. Our code is available at https://github.com/thunlp/GEAR.

SherLIiC: A Typed Event-Focused Lexical Inference Benchmark for Evaluating Natural Language Inference

- https://www.aclweb.org/anthology/P19-1086/

We present SherLIiC, a testbed for lexical inference in context (LIiC), consisting of 3985 manually annotated inference rule candidates (InfCands), accompanied by (i) ~960k unlabeled InfCands, and (ii) ~190k typed textual relations between Freebase entities extracted from the large entity-linked corpus ClueWeb09. Each InfCand consists of one of these relations, expressed as a lemmatized dependency path, and two argument placeholders, each linked to one or more Freebase types. Due to our candidate selection process based on strong distributional evidence, SherLIiC is much harder than existing testbeds because distributional evidence is of little utility in the classification of InfCands. We also show that, due to

its construction, many of SherLIiC's correct InfCands are novel and missing from existing rule bases. We evaluate a large number of strong baselines on SherLIiC, ranging from semantic vector space models to state of the art neural models of natural language inference (NLI). We show that SherLIiC poses a tough challenge to existing NLI systems.

Extracting Symptoms and their Status from Clinical Conversations

- https://www.aclweb.org/anthology/P19-1087/

This paper describes novel models tailored for a new application, that of extracting the symptoms mentioned in clinical conversations along with their status. Lack of any publicly available corpus in this privacy-sensitive domain led us to develop our own corpus, consisting of about 3K conversations annotated by professional medical scribes. We propose two novel deep learning approaches to infer the symptom names and their status: (1) a new hierarchical span-attribute tagging (SA-T) model, trained using curriculum learning, and (2) a variant of sequence-to-sequence model which decodes the symptoms and their status from a few speaker turns within a sliding window over the conversation. This task stems from a realistic application of assisting medical providers in capturing symptoms mentioned by patients from their clinical conversations. To reflect this application, we define multiple metrics. From inter-rater agreement, we find that the task is inherently difficult. We conduct comprehensive evaluations on several contrasting conditions and observe that the performance of the models range from an F-score of 0.5 to 0.8 depending on the condition. Our analysis not only reveals the inherent challenges of the task, but also provides useful directions to improve the models.

What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations

- https://www.aclweb.org/anthology/P19-1088/

The quality of a counseling intervention relies highly on the active collaboration between clients and counselors. In this paper, we explore several linguistic aspects of the collaboration process occurring during counseling conversations. Specifically, we address the differences between high-quality and low-quality counseling. Our approach examines participants' turn-by-turn interaction, their linguistic alignment, the sentiment expressed by speakers during the conversation, as well as the different topics being discussed. Our results suggest important language differences in low- and high-quality counseling, which we further use to derive linguistic features able to capture the differences between the two

groups. These features are then used to build automatic classifiers that can predict counseling quality with accuracies of up to 88%.

Finding Your Voice: The Linguistic Development of Mental Health Counselors

- https://www.aclweb.org/anthology/P19-1089/

Mental health counseling is an enterprise with profound societal importance where conversations play a primary role. In order to acquire the conversational skills needed to face a challenging range of situations, mental health counselors must rely on training and on continued experience with actual clients. However, in the absence of large scale longitudinal studies, the nature and significance of this developmental process remain unclear. For example, prior literature suggests that experience might not translate into consequential changes in counselor behavior. This has led some to even argue that counseling is a profession without expertise. In this work, we develop a computational framework to quantify the extent to which individuals change their linguistic behavior with experience and to study the nature of this evolution. We use our framework to conduct a large longitudinal study of mental health counseling conversations, tracking over 3,400 counselors across their tenure. We reveal that overall, counselors do indeed change their conversational behavior to become more diverse across interactions, developing an individual voice that distinguishes them from other counselors. Furthermore, a finer-grained investigation shows that the rate and nature of this diversification vary across functionally different conversational components.

Towards Automating Healthcare Question Answering in a Noisy Multilingual Low-Resource Setting

- https://www.aclweb.org/anthology/P19-1090/

We discuss ongoing work into automating a multilingual digital helpdesk service available via text messaging to pregnant and breastfeeding mothers in South Africa. Our anonymized dataset consists of short informal questions, often in low-resource languages, with unreliable language labels, spelling errors and code-mixing, as well as template answers with some inconsistencies. We explore cross-lingual word embeddings, and train parametric and non-parametric models on 90K samples for answer selection from a set of 126 templates. Preliminary results indicate that LSTMs trained end-to-end perform best, with a test accuracy of 62.13% and a recall@5 of 89.56%, and demonstrate that we can accelerate response time by several orders of magnitude.

Joint Entity Extraction and Assertion Detection for Clinical Text

- https://www.aclweb.org/anthology/P19-1091/

Negative medical findings are prevalent in clinical reports, yet discriminating them from positive findings remains a challenging task for in-formation extraction. Most of the existing systems treat this task as a pipeline of two separate tasks, i.e., named entity recognition (NER)and rule-based negation detection. We consider this as a multi-task problem and present a novel end-to-end neural model to jointly extract entities and negations. We extend a standard hierarchical encoder-decoder NER model and first adopt a shared encoder followed by separate decoders for the two tasks. This architecture performs considerably better than the previous rule-based and machine learning-based systems. To overcome the problem of increased parameter size especially for low-resource settings, we propose the Conditional Softmax Shared Decoder architecture which achieves state-of-art results for NER and negation detection on the 2010 i2b2/VA challenge dataset and a proprietary de-identified clinical dataset.

HEAD-QA: A Healthcare Dataset for Complex Reasoning

- https://www.aclweb.org/anthology/P19-1092/

We present HEAD-QA, a multi-choice question answering testbed to encourage research on complex reasoning. The questions come from exams to access a specialized position in the Spanish healthcare system, and are challenging even for highly specialized humans. We then consider monolingual (Spanish) and cross-lingual (to English) experiments with information retrieval and neural techniques. We show that: (i) HEAD-QA challenges current methods, and (ii) the results lag well behind human performance, demonstrating its usefulness as a benchmark for future work.

Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network

- https://www.aclweb.org/anthology/P19-1093/

With the advancement in argument detection, we suggest to pay more attention to the challenging task of identifying the more convincing arguments. Machines capable of responding and interacting with humans in helpful ways have become ubiquitous. We now expect them to discuss with us the more delicate questions in our world, and they should do so armed with effective arguments. But what makes an argument more persuasive? What will convince you? In this paper, we present a new data set, IBM-EviConv, of pairs of evidence labeled for convincingness, designed to be more challenging than existing alternatives.

We also propose a Siamese neural network architecture shown to outperform several baselines on both a prior convincingness data set and our own. Finally, we provide insights into our experimental results and the various kinds of argumentative value our method is capable of detecting.

From Surrogacy to Adoption; From Bitcoin to Cryptocurrency: Debate Topic Expansion

- https://www.aclweb.org/anthology/P19-1094/

When debating a controversial topic, it is often desirable to expand the boundaries of discussion. For example, we may consider the pros and cons of possible alternatives to the debate topic, make generalizations, or give specific examples. We introduce the task of Debate Topic Expansion - finding such related topics for a given debate topic, along with a novel annotated dataset for the task. We focus on relations between Wikipedia concepts, and show that they differ from well-studied lexical-semantic relations such as hypernyms, hyponyms and antonyms. We present algorithms for finding both consistent and contrastive expansions and demonstrate their effectiveness empirically. We suggest that debate topic expansion may have various use cases in argumentation mining.

Multimodal and Multi-view Models for Emotion Recognition

- https://www.aclweb.org/anthology/P19-1095/

Studies on emotion recognition (ER) show that combining lexical and acoustic information results in more robust and accurate models. The majority of the studies focus on settings where both modalities are available in training and evaluation. However, in practice, this is not always the case; getting ASR output may represent a bottleneck in a deployment pipeline due to computational complexity or privacy-related constraints. To address this challenge, we study the problem of efficiently combining acoustic and lexical modalities during training while still providing a deployable acoustic model that does not require lexical inputs. We first experiment with multimodal models and two attention mechanisms to assess the extent of the benefits that lexical information can provide. Then, we frame the task as a multi-view learning problem to induce semantic information from a multimodal model into our acoustic-only network using a contrastive loss function. Our multimodal model outperforms the previous state of the art on the USC-IEMOCAP dataset reported on lexical and acoustic information. Additionally, our multi-view-trained acoustic network significantly surpasses models that have been exclusively trained with acoustic features.

Emotion-Cause Pair Extraction: A New Task to Emotion Analysis in Texts

- https://www.aclweb.org/anthology/P19-1096/

Emotion cause extraction (ECE), the task aimed at extracting the potential causes behind certain emotions in text, has gained much attention in recent years due to its wide applications. However, it suffers from two shortcomings: 1) the emotion must be annotated before cause extraction in ECE, which greatly limits its applications in real-world scenarios; 2) the way to first annotate emotion and then extract the cause ignores the fact that they are mutually indicative. In this work, we propose a new task: emotion-cause pair extraction (ECPE), which aims to extract the potential pairs of emotions and corresponding causes in a document. We propose a 2-step approach to address this new ECPE task, which first performs individual emotion extraction and cause extraction via multi-task learning, and then conduct emotion-cause pairing and filtering. The experimental results on a benchmark emotion cause corpus prove the feasibility of the ECPE task as well as the effectiveness of our approach.

Argument Invention from First Principles

- https://www.aclweb.org/anthology/P19-1097/

Competitive debaters often find themselves facing a challenging task – how to debate a topic they know very little about, with only minutes to prepare, and without access to books or the Internet? What they often do is rely on ”first principles”, commonplace arguments which are relevant to many topics, and which they have refined in past debates. In this work we aim to explicitly define a taxonomy of such principled recurring arguments, and, given a controversial topic, to automatically identify which of these arguments are relevant to the topic. As far as we know, this is the first time that this approach to argument invention is formalized and made explicit in the context of NLP. The main goal of this work is to show that it is possible to define such a taxonomy. While the taxonomy suggested here should be thought of as a ”first attempt” it is nonetheless coherent, covers well the relevant topics and coincides with what professional debaters actually argue in their speeches, and facilitates automatic argument invention for new topics.

Improving the Similarity Measure of Determinantal Point Processes for Extractive Multi-Document Summarization

- https://www.aclweb.org/anthology/P19-1098/

The most important obstacles facing multi-document summarization include excessive redundancy in source descriptions and the looming shortage of training

data. These obstacles prevent encoder-decoder models from being used directly, but optimization-based methods such as determinantal point processes (DPPs) are known to handle them well. In this paper we seek to strengthen a DPP-based method for extractive multi-document summarization by presenting a novel similarity measure inspired by capsule networks. The approach measures redundancy between a pair of sentences based on surface form and semantic information. We show that our DPP system with improved similarity measure performs competitively, outperforming strong summarization baselines on benchmark datasets. Our findings are particularly meaningful for summarizing documents created by multiple authors containing redundant yet lexically diverse expressions.

Global Optimization under Length Constraint for Neural Text Summarization

- https://www.aclweb.org/anthology/P19-1099/

We propose a global optimization method under length constraint (GOLC) for neural text summarization models. GOLC increases the probabilities of generating summaries that have high evaluation scores, ROUGE in this paper, within a desired length. We compared GOLC with two optimization methods, a maximum log-likelihood and a minimum risk training, on CNN/Daily Mail and a Japanese single document summarization data set of The Mainichi Shimbun Newspapers. The experimental results show that a state-of-the-art neural summarization model optimized with GOLC generates fewer overlength summaries while maintaining the fastest processing speed; only 6.70% overlength summaries on CNN/Daily and 7.8% on long summary of Mainichi, compared to the approximately 20% to 50% on CNN/Daily Mail and 10% to 30% on Mainichi with the other optimization methods. We also demonstrate the importance of the generation of in-length summaries for post-editing with the dataset Mainich that is created with strict length constraints. The ex- perimental results show approximately 30% to 40% improved post-editing time by use of in-length summaries.

Searching for Effective Neural Extractive Summarization: What Works and What's Next

- https://www.aclweb.org/anthology/P19-1100/

The recent years have seen remarkable success in the use of deep neural networks on text summarization. However, there is no clear understanding of why they perform so well, or how they might be improved. In this paper, we seek to better understand how neural extractive summarization systems could benefit from different types of model architectures, transferable knowledge and learning

schemas. Besides, we find an effective way to improve the current framework and achieve the state-of-the-art result on CNN/DailyMail by a large margin based on our observations and analysis. Hopefully, our work could provide more hints for future research on extractive summarization.

A Simple Theoretical Model of Importance for Summarization

- https://www.aclweb.org/anthology/P19-1101/

Research on summarization has mainly been driven by empirical approaches, crafting systems to perform well on standard datasets with the notion of information Importance remaining latent. We argue that establishing theoretical models of Importance will advance our understanding of the task and help to further improve summarization systems. To this end, we propose simple but rigorous definitions of several concepts that were previously used only intuitively in summarization: Redundancy, Relevance, and Informativeness. Importance arises as a single quantity naturally unifying these concepts. Additionally, we provide intuitions to interpret the proposed quantities and experiments to demonstrate the potential of the framework to inform and guide subsequent works.

Multi-News: A Large-Scale Multi-Document Summarization Dataset and ive Hierarchical Model

- https://www.aclweb.org/anthology/P19-1102/

Automatic generation of summaries from multiple news articles is a valuable tool as the number of online publications grows rapidly. Single document summarization (SDS) systems have benefited from advances in neural encoder-decoder model thanks to the availability of large datasets. However, multi-document summarization (MDS) of news articles has been limited to datasets of a couple of hundred examples. In this paper, we introduce Multi-News, the first large-scale MDS news dataset. Additionally, we propose an end-to-end model which incorporates a traditional extractive summarization model with a standard SDS model and achieves competitive results on MDS datasets. We benchmark several methods on Multi-News and hope that this work will promote advances in summarization in the multi-document setting.

Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency

- https://www.aclweb.org/anthology/P19-1103/

We address the problem of adversarial attacks on text classification, which is rarely studied comparing to attacks on image classification. The challenge of this task is to generate adversarial examples that maintain lexical correctness, grammatical correctness and semantic similarity. Based on the synonyms substitution strategy, we introduce a new word replacement order determined by both the word saliency and the classification probability, and propose a greedy algorithm called probability weighted word saliency (PWWS) for text adversarial attack. Experiments on three popular datasets using convolutional as well as LSTM models show that PWWS reduces the classification accuracy to the most extent, and keeps a very low word substitution rate. A human evaluation study shows that our generated adversarial examples maintain the semantic similarity well and are hard for humans to perceive. Performing adversarial training using our perturbed datasets improves the robustness of the models. At last, our method also exhibits a good transferability on the generated adversarial examples.

Heuristic Authorship Obfuscation

- https://www.aclweb.org/anthology/P19-1104/

Authorship verification is the task of determining whether two texts were written by the same author. We deal with the adversary task, called authorship obfuscation: preventing verification by altering a to-be-obfuscated text. Our new obfuscation approach (1) models writing style difference as the Jensen-Shannon distance between the character n-gram distributions of texts, and (2) manipulates an author's subconsciously encoded writing style in a sophisticated manner using heuristic search. To obfuscate, we analyze the huge space of textual variants for a paraphrased version of the to-be-obfuscated text that has a sufficient Jensen-Shannon distance at minimal costs in terms of text quality. We analyze, quantify, and illustrate the rationale of this approach, define paraphrasing operators, derive obfuscation thresholds, and develop an effective obfuscation framework. Our authorship obfuscation approach defeats state-of-the-art verification approaches, including unmasking and compression models, while keeping text changes at a minimum.

Text Categorization by Learning Predominant Sense of Words as Auxiliary Task

- https://www.aclweb.org/anthology/P19-1105/

Distributions of the senses of words are often highly skewed and give a strong influence of the domain of a document. This paper follows the assumption and presents a method for text categorization by leveraging the predominant sense of words depending on the domain, i.e., domain-specific senses. The key idea is that the features learned from predominant senses are possible to discriminate

the domain of the document and thus improve the overall performance of text categorization. We propose multi-task learning framework based on the neural network model, transformer, which trains a model to simultaneously categorize documents and predicts a predominant sense for each word. The experimental results using four benchmark datasets show that our method is comparable to the state-of-the-art categorization approach, especially our model works well for categorization of multi-label documents.

DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions

- https://www.aclweb.org/anthology/P19-1106/

Automatically validating a research artefact is one of the frontiers in Artificial Intelligence (AI) that directly brings it close to competing with human intellect and intuition. Although criticised sometimes, the existing peer review system still stands as the benchmark of research validation. The present-day peer review process is not straightforward and demands profound domain knowledge, expertise, and intelligence of human reviewer(s), which is somewhat elusive with the current state of AI. However, the peer review texts, which contains rich sentiment information of the reviewer, reflecting his/her overall attitude towards the research in the paper, could be a valuable entity to predict the acceptance or rejection of the manuscript under consideration. Here in this work, we investigate the role of reviewer sentiment embedded within peer review texts to predict the peer review outcome. Our proposed deep neural architecture takes into account three channels of information: the paper, the corresponding reviews, and review's polarity to predict the overall recommendation score as well as the final decision. We achieve significant performance improvement over the baselines (29% error reduction) proposed in a recently released dataset of peer reviews. An AI of this kind could assist the editors/program chairs as an additional layer of confidence, especially when non-responding/missing reviewers are frequent in present day peer review.

Gated Embeddings in End-to-End Speech Recognition for Conversational-Context Fusion

- https://www.aclweb.org/anthology/P19-1107/

We present a novel conversational-context aware end-to-end speech recognizer based on a gated neural network that incorporates conversational-context/word/speech embeddings. Unlike conventional speech recognition models, our model learns longer conversational-context information that spans across sentences and is consequently better at recognizing long conversations. Specifically, we propose to use text-based external word and/or sentence

embeddings (i.e., fastText, BERT) within an end-to-end framework, yielding significant improvement in word error rate with better conversational-context representation. We evaluated the models on the Switchboard conversational speech corpus and show that our model outperforms standard end-to-end speech recognition models.

Figurative Usage Detection of Symptom Words to Improve Personal Health Mention Detection

- https://www.aclweb.org/anthology/P19-1108/

Personal health mention detection deals with predicting whether or not a given sentence is a report of a health condition. Past work mentions errors in this prediction when symptom words, i.e., names of symptoms of interest, are used in a figurative sense. Therefore, we combine a state-of-the-art figurative usage detection with CNN-based personal health mention detection. To do so, we present two methods: a pipeline-based approach and a feature augmentation-based approach. The introduction of figurative usage detection results in an average improvement of 2.21% F-score of personal health mention detection, in the case of the feature augmentation-based approach. This paper demonstrates the promise of using figurative usage detection to improve personal health mention detection.

Complex Word Identification as a Sequence Labelling Task

- https://www.aclweb.org/anthology/P19-1109/

Complex Word Identification (CWI) is concerned with detection of words in need of simplification and is a crucial first step in a simplification pipeline. It has been shown that reliable CWI systems considerably improve text simplification. However, most CWI systems to date address the task on a word-by-word basis, not taking the context into account. In this paper, we present a novel approach to CWI based on sequence modelling. Our system is capable of performing CWI in context, does not require extensive feature engineering and outperforms state-of-the-art systems on this task.

Neural News Recommendation with Topic-Aware News Representation

- https://www.aclweb.org/anthology/P19-1110/

News recommendation can help users find interested news and alleviate information overload. The topic information of news is critical for learning accurate news and user representations for news recommendation. However, it is not

considered in many existing news recommendation methods. In this paper, we propose a neural news recommendation approach with topic-aware news representations. The core of our approach is a topic-aware news encoder and a user encoder. In the news encoder we learn representations of news from their titles via CNN networks and apply attention networks to select important words. In addition, we propose to learn topic-aware news representations by jointly training the news encoder with an auxiliary topic classification task. In the user encoder we learn the representations of users from their browsed news and use attention networks to select informative news for user representation learning. Extensive experiments on a real-world dataset validate the effectiveness of our approach.

Poetry to Prose Conversion in Sanskrit as a Linearisation Task: A Case for Low-Resource Languages

- https://www.aclweb.org/anthology/P19-1111/

The word ordering in a Sanskrit verse is often not aligned with its corresponding prose order. Conversion of the verse to its corresponding prose helps in better comprehension of the construction. Owing to the resource constraints, we formulate this task as a word ordering (linearisation) task. In doing so, we completely ignore the word arrangement at the verse side. kāvya guru, the approach we propose, essentially consists of a pipeline of two pretraining steps followed by a seq2seq model. The first pretraining step learns task-specific token embeddings from pretrained embeddings. In the next step, we generate multiple possible hypotheses for possible word arrangements of the input %using another pretraining step. We then use them as inputs to a neural seq2seq model for the final prediction. We empirically show that the hypotheses generated by our pretraining step result in predictions that consistently outperform predictions based on the original order in the verse. Overall, kāvya guru outperforms current state of the art models in linearisation for the poetry to prose conversion task in Sanskrit.

Learning Emphasis Selection for Written Text in Visual Media from Crowd-Sourced Label Distributions

- https://www.aclweb.org/anthology/P19-1112/

In visual communication, text emphasis is used to increase the comprehension of written text to convey the author's intent. We study the problem of emphasis selection, i.e. choosing candidates for emphasis in short written text, to enable automated design assistance in authoring. Without knowing the author's intent and only considering the input text, multiple emphasis selections are valid. We propose a model that employs end-to-end label distribution learning (LDL) on

crowd-sourced data and predicts a selection distribution, capturing the inter-subjectivity (common-sense) in the audience as well as the ambiguity of the input. We compare the model with several baselines in which the problem is transformed to single-label learning by mapping label distributions to absolute labels via majority voting.

Rumor Detection by Exploiting User Credibility Information, Attention and Multi-task Learning

- https://www.aclweb.org/anthology/P19-1113/

In this study, we propose a new multi-task learning approach for rumor detection and stance classification tasks. This neural network model has a shared layer and two task specific layers. We incorporate the user credibility information into the rumor detection layer, and we also apply attention mechanism in the rumor detection process. The attended information include not only the hidden states in the rumor detection layer, but also the hidden states from the stance detection layer. The experiments on two datasets show that our proposed model outperforms the state-of-the-art rumor detection approaches.

Context-specific Language Modeling for Human Trafficking Detection from Online Advertisements

- https://www.aclweb.org/anthology/P19-1114/

Human trafficking is a worldwide crisis. Traffickers exploit their victims by anonymously offering sexual services through online advertisements. These ads often contain clues that law enforcement can use to separate out potential trafficking cases from volunteer sex advertisements. The problem is that the sheer volume of ads is too overwhelming for manual processing. Ideally, a centralized semi-automated tool can be used to assist law enforcement agencies with this task. Here, we present an approach using natural language processing to identify trafficking ads on these websites. We propose a classifier by integrating multiple text feature sets, including the publicly available pre-trained textual language model Bi-directional Encoder Representation from transformers (BERT). In this paper, we demonstrate that a classifier using this composite feature set has significantly better performance compared to any single feature set alone.

Self-Attentional Models for Lattice Inputs

- https://www.aclweb.org/anthology/P19-1115/

Lattices are an efficient and effective method to encode ambiguity of upstream systems in natural language processing tasks, for example to compactly cap-

ture multiple speech recognition hypotheses, or to represent multiple linguistic analyses. Previous work has extended recurrent neural networks to model lattice inputs and achieved improvements in various tasks, but these models suffer from very slow computation speeds. This paper extends the recently proposed paradigm of self-attention to handle lattice inputs. Self-attention is a sequence modeling technique that relates inputs to one another by computing pairwise similarities and has gained popularity for both its strong results and its computational efficiency. To extend such models to handle lattices, we introduce probabilistic reachability masks that incorporate lattice structure into the model and support lattice scores if available. We also propose a method for adapting positional embeddings to lattice structures. We apply the proposed model to a speech translation task and find that it outperforms all examined baselines while being much faster to compute than previous neural lattice models during both training and inference.

When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion

- https://www.aclweb.org/anthology/P19-1116/

Though machine translation errors caused by the lack of context beyond one sentence have long been acknowledged, the development of context-aware NMT systems is hampered by several problems. Firstly, standard metrics are not sensitive to improvements in consistency in document-level translations. Secondly, previous work on context-aware NMT assumed that the sentence-aligned parallel data consisted of complete documents while in most practical scenarios such document-level data constitutes only a fraction of the available parallel data. To address the first issue, we perform a human study on an English-Russian subtitles dataset and identify deixis, ellipsis and lexical cohesion as three main sources of inconsistency. We then create test sets targeting these phenomena. To address the second shortcoming, we consider a set-up in which a much larger amount of sentence-level data is available compared to that aligned at the document level. We introduce a model that is suitable for this scenario and demonstrate major gains over a context-agnostic baseline on our new benchmarks without sacrificing performance as measured with BLEU.

A Compact and Language-Sensitive Multilingual Translation Method

- https://www.aclweb.org/anthology/P19-1117/

Multilingual neural machine translation (Multi-NMT) with one encoder-decoder model has made remarkable progress due to its simple deployment. However,

this multilingual translation paradigm does not make full use of language commonality and parameter sharing between encoder and decoder. Furthermore, this kind of paradigm cannot outperform the individual models trained on bilingual corpus in most cases. In this paper, we propose a compact and language-sensitive method for multilingual translation. To maximize parameter sharing, we first present a universal representor to replace both encoder and decoder models. To make the representor sensitive for specific languages, we further introduce language-sensitive embedding, attention, and discriminator with the ability to enhance model performance. We verify our methods on various translation scenarios, including one-to-many, many-to-many and zero-shot. Extensive experiments demonstrate that our proposed methods remarkably outperform strong standard multilingual translation systems on WMT and IWSLT datasets. Moreover, we find that our model is especially helpful in low-resource and zero-shot translation scenarios.

Unsupervised Parallel Sentence Extraction with Parallel Segment Detection Helps Machine Translation

- https://www.aclweb.org/anthology/P19-1118/

Mining parallel sentences from comparable corpora is important. Most previous work relies on supervised systems, which are trained on parallel data, thus their applicability is problematic in low-resource scenarios. Recent developments in building unsupervised bilingual word embeddings made it possible to mine parallel sentences based on cosine similarities of source and target language words. We show that relying only on this information is not enough, since sentences often have similar words but different meanings. We detect continuous parallel segments in sentence pair candidates and rely on them when mining parallel sentences. We show better mining accuracy on three language pairs in a standard shared task on artificial data. We also provide the first experiments showing that parallel sentences mined from real life sources improve unsupervised MT. Our code is available, we hope it will be used to support low-resource MT research.

Unsupervised Bilingual Word Embedding Agreement for Unsupervised Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1119/

Unsupervised bilingual word embedding (UBWE), together with other technologies such as back-translation and denoising, has helped unsupervised neural machine translation (UNMT) achieve remarkable results in several language pairs. In previous methods, UBWE is first trained using non-parallel monolingual corpora and then this pre-trained UBWE is used to initialize the word embedding in the encoder and decoder of UNMT. That is, the training of UBWE and UNMT

are separate. In this paper, we first empirically investigate the relationship between UBWE and UNMT. The empirical findings show that the performance of UNMT is significantly affected by the performance of UBWE. Thus, we propose two methods that train UNMT with UBWE agreement. Empirical results on several language pairs show that the proposed methods significantly outperform conventional UNMT.

Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies

- https://www.aclweb.org/anthology/P19-1120/

Transfer learning or multilingual model is essential for low-resource neural machine translation (NMT), but the applicability is limited to cognate languages by sharing their vocabularies. This paper shows effective techniques to transfer a pretrained NMT model to a new, unrelated language without shared vocabularies. We relieve the vocabulary mismatch by using cross-lingual word embedding, train a more language-agnostic encoder by injecting artificial noises, and generate synthetic data easily from the pretraining data without back-translation. Our methods do not require restructuring the vocabulary or retraining the model. We improve plain NMT transfer by up to +5.1% BLEU in five low-resource translation tasks, outperforming multilingual joint training by a large margin. We also provide extensive ablation studies on pretrained embedding, synthetic data, vocabulary size, and parameter freezing for a better understanding of NMT transfer.

Improved Zero-shot Neural Machine Translation via Ignoring Spurious Correlations

- https://www.aclweb.org/anthology/P19-1121/

Zero-shot translation, translating between language pairs on which a Neural Machine Translation (NMT) system has never been trained, is an emergent property when training the system in multilingual settings. However, naive training for zero-shot NMT easily fails, and is sensitive to hyper-parameter setting. The performance typically lags far behind the more conventional pivot-based approach which translates twice using a third language as a pivot. In this work, we address the degeneracy problem due to capturing spurious correlations by quantitatively analyzing the mutual information between language IDs of the source and decoded sentences. Inspired by this analysis, we propose to use two simple but effective approaches: (1) decoder pre-training; (2) back-translation. These methods show significant improvement (4 22 BLEU points) over the vanilla zero-shot translation on three challenging multilingual datasets, and achieve similar or better results than the pivot-based approach.

Syntactically Supervised Transformers for Faster Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1122/

Standard decoders for neural machine translation autoregressively generate a single target token per timestep, which slows inference especially for long outputs. While architectural advances such as the Transformer fully parallelize the decoder computations at training time, inference still proceeds sequentially. Recent developments in non- and semi-autoregressive decoding produce multiple tokens per timestep independently of the others, which improves inference speed but deteriorates translation quality. In this work, we propose the syntactically supervised Transformer (SynST), which first autoregressively predicts a chunked parse tree before generating all of the target tokens in one shot conditioned on the predicted parse. A series of controlled experiments demonstrates that SynST decodes sentences ~5x faster than the baseline autoregressive Transformer while achieving higher BLEU scores than most competing methods on En-De and En-Fr datasets.

Dynamically Composing Domain-Data Selection with Clean-Data Selection by "Co-Curricular Learning" for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1123/

Noise and domain are important aspects of data quality for neural machine translation. Existing research focus separately on domain-data selection, clean-data selection, or their static combination, leaving the dynamic interaction across them not explicitly examined. This paper introduces a "co-curricular learning" method to compose dynamic domain-data selection with dynamic clean-data selection, for transfer learning across both capabilities. We apply an EM-style optimization procedure to further refine the "co-curriculum". Experiment results and analysis with two domains demonstrate the effectiveness of the method and the properties of data scheduled by the co-curriculum.

On the Word Alignment from Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1124/

Prior researches suggest that neural machine translation (NMT) captures word alignment through its attention mechanism, however, this paper finds attention may almost fail to capture word alignment for some NMT models. This paper thereby proposes two methods to induce word alignment which are general and agnostic to specific NMT models. Experiments show that both methods induce much better word alignment than attention. This paper further visualizes the

translation through the word alignment induced by NMT. In particular, it analyzes the effect of alignment errors on translation errors at word level and its quantitative analysis over many testing examples consistently demonstrate that alignment errors are likely to lead to translation errors measured by different metrics.

Imitation Learning for Non-Autoregressive Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1125/

Non-autoregressive translation models (NAT) have achieved impressive inference speedup. A potential issue of the existing NAT algorithms, however, is that the decoding is conducted in parallel, without directly considering previous context. In this paper, we propose an imitation learning framework for non-autoregressive machine translation, which still enjoys the fast translation speed but gives comparable translation performance compared to its auto-regressive counterpart. We conduct experiments on the IWSLT16, WMT14 and WMT16 datasets. Our proposed model achieves a significant speedup over the autoregressive models, while keeping the translation quality comparable to the autoregressive models. By sampling sentence length in parallel at inference time, we achieve the performance of 31.85 BLEU on WMT16 Ro→En and 30.68 BLEU on IWSLT16 En→De.

Monotonic Infinite Lookback Attention for Simultaneous Machine Translation

- https://www.aclweb.org/anthology/P19-1126/

Simultaneous machine translation begins to translate each source sentence before the source speaker is finished speaking, with applications to live and streaming scenarios. Simultaneous systems must carefully schedule their reading of the source sentence to balance quality against latency. We present the first simultaneous translation system to learn an adaptive schedule jointly with a neural machine translation (NMT) model that attends over all source tokens read thus far. We do so by introducing Monotonic Infinite Lookback (MILk) attention, which maintains both a hard, monotonic attention head to schedule the reading of the source sentence, and a soft attention head that extends from the monotonic head back to the beginning of the source. We show that MILk's adaptive schedule allows it to arrive at latency-quality trade-offs that are favorable to those of a recently proposed wait-k strategy for many latency values.

Global Textual Relation Embedding for Relational Understanding

- https://www.aclweb.org/anthology/P19-1127/

Pre-trained embeddings such as word embeddings and sentence embeddings are fundamental tools facilitating a wide range of downstream NLP tasks. In this work, we investigate how to learn a general-purpose embedding of textual relations, defined as the shortest dependency path between entities. Textual relation embedding provides a level of knowledge between word/phrase level and sentence level, and we show that it can facilitate downstream tasks requiring relational understanding of the text. To learn such an embedding, we create the largest distant supervision dataset by linking the entire English ClueWeb09 corpus to Freebase. We use global co-occurrence statistics between textual and knowledge base relations as the supervision signal to train the embedding. Evaluation on two relational understanding tasks demonstrates the usefulness of the learned textual relation embedding. The data and code can be found at https://github.com/czyssrs/GloREPlus

Graph Neural Networks with Generated Parameters for Relation Extraction

- https://www.aclweb.org/anthology/P19-1128/

In this paper, we propose a novel graph neural network with generated parameters (GP-GNNs). The parameters in the propagation module, i.e. the transition matrices used in message passing procedure, are produced by a generator taking natural language sentences as inputs. We verify GP-GNNs in relation extraction from text, both on bag- and instance-settings. Experimental results on a human-annotated dataset and two distantly supervised datasets show that multi-hop reasoning mechanism yields significant improvements. We also perform a qualitative analysis to demonstrate that our model could discover more accurate relations by multi-hop relational reasoning.

Entity-Relation Extraction as Multi-Turn Question Answering

- https://www.aclweb.org/anthology/P19-1129/

In this paper, we propose a new paradigm for the task of entity-relation extraction. We cast the task as a multi-turn question answering problem, i.e., the extraction of entities and elations is transformed to the task of identifying answer spans from the context. This multi-turn QA formalization comes with several key advantages: firstly, the question query encodes important information for the entity/relation class we want to identify; secondly, QA provides a

natural way of jointly modeling entity and relation; and thirdly, it allows us to exploit the well developed machine reading comprehension (MRC) models. Experiments on the ACE and the CoNLL04 corpora demonstrate that the proposed paradigm significantly outperforms previous best models. We are able to obtain the state-of-the-art results on all of the ACE04, ACE05 and CoNLL04 datasets, increasing the SOTA results on the three datasets to 49.6 (+1.2), 60.3 (+0.7) and 69.2 (+1.4), respectively. Additionally, we construct and will release a newly developed dataset RESUME, which requires multi-step reasoning to construct entity dependencies, as opposed to the single-step dependency extraction in the triplet exaction in previous datasets. The proposed multi-turn QA model also achieves the best performance on the RESUME dataset.

Exploiting Entity BIO Tag Embeddings and Multi-task Learning for Relation Extraction with Imbalanced Data

- https://www.aclweb.org/anthology/P19-1130/

In practical scenario, relation extraction needs to first identify entity pairs that have relation and then assign a correct relation class. However, the number of non-relation entity pairs in context (negative instances) usually far exceeds the others (positive instances), which negatively affects a model's performance. To mitigate this problem, we propose a multi-task architecture which jointly trains a model to perform relation identification with cross-entropy loss and relation classification with ranking loss. Meanwhile, we observe that a sentence may have multiple entities and relation mentions, and the patterns in which the entities appear in a sentence may contain useful semantic information that can be utilized to distinguish between positive and negative instances. Thus we further incorporate the embeddings of character-wise/word-wise BIO tag from the named entity recognition task into character/word embeddings to enrich the input representation. Experiment results show that our proposed approach can significantly improve the performance of a baseline model with more than 10% absolute increase in F1-score, and outperform the state-of-the-art models on ACE 2005 Chinese and English corpus. Moreover, BIO tag embeddings are particularly effective and can be used to improve other models as well.

Joint Type Inference on Entities and Relations via Graph Convolutional Networks

- https://www.aclweb.org/anthology/P19-1131/

We develop a new paradigm for the task of joint entity relation extraction. It first identifies entity spans, then performs a joint inference on entity types and relation types. To tackle the joint type inference task, we propose a novel graph convolutional network (GCN) running on an entity-relation bipartite graph. By

introducing a binary relation classification task, we are able to utilize the structure of entity-relation bipartite graph in a more efficient and interpretable way. Experiments on ACE05 show that our model outperforms existing joint models in entity performance and is competitive with the state-of-the-art in relation performance.

Extracting Multiple-Relations in One-Pass with Pre-Trained Transformers

- https://www.aclweb.org/anthology/P19-1132/

Many approaches to extract multiple relations from a paragraph require multiple passes over the paragraph. In practice, multiple passes are computationally expensive and this makes difficult to scale to longer paragraphs and larger text corpora. In this work, we focus on the task of multiple relation extractions by encoding the paragraph only once. We build our solution upon the pre-trained self-attentive models (Transformer), where we first add a structured prediction layer to handle extraction between multiple entity pairs, then enhance the paragraph embedding to capture multiple relational information associated with each entity with entity-aware attention. We show that our approach is not only scalable but can also perform state-of-the-art on the standard benchmark ACE 2005.

Unsupervised Information Extraction: Regularizing Discriminative Approaches with Relation Distribution Losses

- https://www.aclweb.org/anthology/P19-1133/

Unsupervised relation extraction aims at extracting relations between entities in text. Previous unsupervised approaches are either generative or discriminative. In a supervised setting, discriminative approaches, such as deep neural network classifiers, have demonstrated substantial improvement. However, these models are hard to train without supervision, and the currently proposed solutions are unstable. To overcome this limitation, we introduce a skewness loss which encourages the classifier to predict a relation with confidence given a sentence, and a distribution distance loss enforcing that all relations are predicted in average. These losses improve the performance of discriminative based models, and enable us to train deep neural networks satisfactorily, surpassing current state of the art on three different datasets.

Fine-tuning Pre-Trained Transformer Language Models to Distantly Supervised Relation Extraction

- https://www.aclweb.org/anthology/P19-1134/

Distantly supervised relation extraction is widely used to extract relational facts from text, but suffers from noisy labels. Current relation extraction methods try to alleviate the noise by multi-instance learning and by providing supporting linguistic and contextual information to more efficiently guide the relation classification. While achieving state-of-the-art results, we observed these models to be biased towards recognizing a limited set of relations with high precision, while ignoring those in the long tail. To address this gap, we utilize a pre-trained language model, the OpenAI Generative Pre-trained Transformer (GPT) (Radford et al., 2018). The GPT and similar models have been shown to capture semantic and syntactic features, and also a notable amount of "common-sense" knowledge, which we hypothesize are important features for recognizing a more diverse set of relations. By extending the GPT to the distantly supervised setting, and fine-tuning it on the NYT10 dataset, we show that it predicts a larger set of distinct relation types with high confidence. Manual and automated evaluation of our model shows that it achieves a state-of-the-art AUC score of 0.422 on the NYT10 dataset, and performs especially well at higher recall levels.

ARNOR: Attention Regularization based Noise Reduction for Distant Supervision Relation Classification

- https://www.aclweb.org/anthology/P19-1135/

Distant supervision is widely used in relation classification in order to create large-scale training data by aligning a knowledge base with an unlabeled corpus. However, it also introduces amounts of noisy labels where a contextual sentence actually does not express the labeled relation. In this paper, we propose ARNOR, a novel Attention Regularization based NOise Reduction framework for distant supervision relation classification. ARNOR assumes that a trustable relation label should be explained by the neural attention model. Specifically, our ARNOR framework iteratively learns an interpretable model and utilizes it to select trustable instances. We first introduce attention regularization to force the model to pay attention to the patterns which explain the relation labels, so as to make the model more interpretable. Then, if the learned model can clearly locate the relation patterns of a candidate instance in the training set, we will select it as a trustable instance for further training step. According to the experiments on NYT data, our ARNOR framework achieves significant improvements over state-of-the-art methods in both relation classification performance and noise reduction effect.

GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction

- https://www.aclweb.org/anthology/P19-1136/

In this paper, we present GraphRel, an end-to-end relation extraction model which uses graph convolutional networks (GCNs) to jointly learn named entities and relations. In contrast to previous baselines, we consider the interaction between named entities and relations via a 2nd-phase relation-weighted GCN to better extract relations. Linear and dependency structures are both used to extract both sequential and regional features of the text, and a complete word graph is further utilized to extract implicit features among all word pairs of the text. With the graph-based approach, the prediction for overlapping relations is substantially improved over previous sequential approaches. We evaluate GraphRel on two public datasets: NYT and WebNLG. Results show that GraphRel maintains high precision while increasing recall substantially. Also, GraphRel outperforms previous work by 3.2% and 5.8% (F1 score), achieving a new state-of-the-art for relation extraction.

DIAG-NRE: A Neural Pattern Diagnosis Framework for Distantly Supervised Neural Relation Extraction

- https://www.aclweb.org/anthology/P19-1137/

Pattern-based labeling methods have achieved promising results in alleviating the inevitable labeling noises of distantly supervised neural relation extraction. However, these methods require significant expert labor to write relation-specific patterns, which makes them too sophisticated to generalize quickly. To ease the labor-intensive workload of pattern writing and enable the quick generalization to new relation types, we propose a neural pattern diagnosis framework, DIAG-NRE, that can automatically summarize and refine high-quality relational patterns from noise data with human experts in the loop. To demonstrate the effectiveness of DIAG-NRE, we apply it to two real-world datasets and present both significant and interpretable improvements over state-of-the-art methods.

Multi-grained Named Entity Recognition

- https://www.aclweb.org/anthology/P19-1138/

This paper presents a novel framework, MGNER, for Multi-Grained Named Entity Recognition where multiple entities or entity mentions in a sentence could be non-overlapping or totally nested. Different from traditional approaches regarding NER as a sequential labeling task and annotate entities consecutively, MGNER detects and recognizes entities on multiple granularities: it is able to recognize named entities without explicitly assuming non-overlapping or totally nested structures. MGNER consists of a Detector that examines all possible word segments and a Classifier that categorizes entities. In addition, contextual information and a self-attention mechanism are utilized throughout the framework to improve the NER performance. Experimental results show that

MGNER outperforms current state-of-the-art baselines up to 4.4% in terms of the F1 score among nested/non-overlapping NER tasks.

ERNIE: Enhanced Language Representation with Informative Entities

- https://www.aclweb.org/anthology/P19-1139/

Neural language representation models such as BERT pre-trained on large-scale corpora can well capture rich semantic patterns from plain text, and be fine-tuned to consistently improve the performance of various NLP tasks. However, the existing pre-trained language models rarely consider incorporating knowledge graphs (KGs), which can provide rich structured knowledge facts for better language understanding. We argue that informative entities in KGs can enhance language representation with external knowledge. In this paper, we utilize both large-scale textual corpora and KGs to train an enhanced language representation model (ERNIE), which can take full advantage of lexical, syntactic, and knowledge information simultaneously. The experimental results have demonstrated that ERNIE achieves significant improvements on various knowledge-driven tasks, and meanwhile is comparable with the state-of-the-art model BERT on other common NLP tasks. The code and datasets will be available in the future.

Multi-Channel Graph Neural Network for Entity Alignment

- https://www.aclweb.org/anthology/P19-1140/

Entity alignment typically suffers from the issues of structural heterogeneity and limited seed alignments. In this paper, we propose a novel Multi-channel Graph Neural Network model (MuGNN) to learn alignment-oriented knowledge graph (KG) embeddings by robustly encoding two KGs via multiple channels. Each channel encodes KGs via different relation weighting schemes with respect to self-attention towards KG completion and cross-KG attention for pruning exclusive entities respectively, which are further combined via pooling techniques. Moreover, we also infer and transfer rule knowledge for completing two KGs consistently. MuGNN is expected to reconcile the structural differences of two KGs, and thus make better use of seed alignments. Extensive experiments on five publicly available datasets demonstrate our superior performance (5% Hits@1 up on average). Source code and data used in the experiments can be accessed at https://github.com/thunlp/MuGNN .

A Neural Multi-digraph Model for Chinese NER with Gazetteers

- https://www.aclweb.org/anthology/P19-1141/

Gazetteers were shown to be useful resources for named entity recognition (NER). Many existing approaches to incorporating gazetteers into machine learning based NER systems rely on manually defined selection strategies or handcrafted templates, which may not always lead to optimal effectiveness, especially when multiple gazetteers are involved. This is especially the case for the task of Chinese NER, where the words are not naturally tokenized, leading to additional ambiguities. To automatically learn how to incorporate multiple gazetteers into an NER system, we propose a novel approach based on graph neural networks with a multi-digraph structure that captures the information that the gazetteers offer. Experiments on various datasets show that our model is effective in incorporating rich gazetteer information while resolving ambiguities, outperforming previous approaches.

Improved Language Modeling by Decoding the Past

- https://www.aclweb.org/anthology/P19-1142/

Highly regularized LSTMs achieve impressive results on several benchmark datasets in language modeling. We propose a new regularization method based on decoding the last token in the context using the predicted distribution of the next token. This biases the model towards retaining more contextual information, in turn improving its ability to predict the next token. With negligible overhead in the number of parameters and training time, our Past Decode Regularization (PDR) method improves perplexity on the Penn Treebank dataset by up to 1.8 points and by up to 2.3 points on the WikiText-2 dataset, over strong regularized baselines using a single softmax. With a mixture-of-softmax model, we show gains of up to 1.0 perplexity points on these datasets. In addition, our method achieves 1.169 bits-per-character on the Penn Treebank Character dataset for character level language modeling.

Training Hybrid Language Models by Marginalizing over Segmentations

- https://www.aclweb.org/anthology/P19-1143/

In this paper, we study the problem of hybrid language modeling, that is using models which can predict both characters and larger units such as character ngrams or words. Using such models, multiple potential segmentations usually exist for a given string, for example one using words and one using characters only. Thus, the probability of a string is the sum of the probabilities of all the

possible segmentations. Here, we show how it is possible to marginalize over the segmentations efficiently, in order to compute the true probability of a sequence. We apply our technique on three datasets, comprising seven languages, showing improvements over a strong character level language model.

Improving Neural Language Models by Segmenting, Attending, and Predicting the Future

- https://www.aclweb.org/anthology/P19-1144/

Common language models typically predict the next word given the context. In this work, we propose a method that improves language modeling by learning to align the given context and the following phrase. The model does not require any linguistic annotation of phrase segmentation. Instead, we define syntactic heights and phrase segmentation rules, enabling the model to automatically induce phrases, recognize their task-specific heads, and generate phrase embeddings in an unsupervised learning manner. Our method can easily be applied to language models with different network architectures since an independent module is used for phrase induction and context-phrase alignment, and no change is required in the underlying language modeling network. Experiments have shown that our model outperformed several strong baseline models on different data sets. We achieved a new state-of-the-art performance of 17.4 perplexity on the Wikitext-103 dataset. Additionally, visualizing the outputs of the phrase induction module showed that our model is able to learn approximate phrase-level structural knowledge without any annotation.

Lightweight and Efficient Neural Natural Language Processing with Quaternion Networks

- https://www.aclweb.org/anthology/P19-1145/

Many state-of-the-art neural models for NLP are heavily parameterized and thus memory inefficient. This paper proposes a series of lightweight and memory efficient neural architectures for a potpourri of natural language processing (NLP) tasks. To this end, our models exploit computation using Quaternion algebra and hypercomplex spaces, enabling not only expressive inter-component interactions but also significantly (75%) reduced parameter size due to lesser degrees of freedom in the Hamilton product. We propose Quaternion variants of models, giving rise to new architectures such as the Quaternion attention Model and Quaternion Transformer. Extensive experiments on a battery of NLP tasks demonstrates the utility of proposed Quaternion-inspired models, enabling up to 75% reduction in parameter size without significant loss in performance.

Sparse Sequence-to-Sequence Models

- https://www.aclweb.org/anthology/P19-1146/

Sequence-to-sequence models are a powerful workhorse of NLP. Most variants employ a softmax transformation in both their attention mechanism and output layer, leading to dense alignments and strictly positive output probabilities. This density is wasteful, making models less interpretable and assigning probability mass to many implausible outputs. In this paper, we propose sparse sequence-to-sequence models, rooted in a new family of -entmax transformations, which includes softmax and sparsemax as particular cases, and is sparse for any > 1. We provide fast algorithms to evaluate these transformations and their gradients, which scale well for large vocabulary sizes. Our models are able to produce sparse alignments and to assign nonzero probability to a short list of plausible outputs, sometimes rendering beam search exact. Experiments on morphological inflection and machine translation reveal consistent gains over dense models.

On the Robustness of Self-Attentive Models

- https://www.aclweb.org/anthology/P19-1147/

This work examines the robustness of self-attentive neural networks against adversarial input perturbations. Specifically, we investigate the attention and feature extraction mechanisms of state-of-the-art recurrent neural networks and self-attentive architectures for sentiment analysis, entailment and machine translation under adversarial attacks. We also propose a novel attack algorithm for generating more natural adversarial examples that could mislead neural models but not humans. Experimental results show that, compared to recurrent neural models, self-attentive models are more robust against adversarial perturbation. In addition, we provide theoretical explanations for their superior robustness to support our claims.

Exact Hard Monotonic Attention for Character-Level Transduction

- https://www.aclweb.org/anthology/P19-1148/

Many common character-level, string-to-string transduction tasks, e.g., grapheme-to-phoneme conversion and morphological inflection, consist almost exclusively of monotonic transduction. Neural sequence-to-sequence models with soft attention, non-monotonic models, outperform popular monotonic models. In this work, we ask the following question: Is monotonicity really a helpful inductive bias in these tasks? We develop a hard attention sequence-to-sequence model that enforces strict monotonicity and learns

alignment jointly. With the help of dynamic programming, we are able to compute the exact marginalization over all alignments. Our models achieve state-of-the-art performance on morphological inflection. Furthermore, we find strong performance on two other character-level transduction tasks. Code is available at https://github.com/shijie-wu/neural-transducer.

A Lightweight Recurrent Network for Sequence Modeling

- https://www.aclweb.org/anthology/P19-1149/

Recurrent networks have achieved great success on various sequential tasks with the assistance of complex recurrent units, but suffer from severe computational inefficiency due to weak parallelization. One direction to alleviate this issue is to shift heavy computations outside the recurrence. In this paper, we propose a lightweight recurrent network, or LRN. LRN uses input and forget gates to handle long-range dependencies as well as gradient vanishing and explosion, with all parameter related calculations factored outside the recurrence. The recurrence in LRN only manipulates the weight assigned to each token, tightly connecting LRN with self-attention networks. We apply LRN as a drop-in replacement of existing recurrent units in several neural sequential models. Extensive experiments on six NLP tasks show that LRN yields the best running efficiency with little or no loss in model performance.

Towards Scalable and Reliable Capsule Networks for Challenging NLP Applications

- https://www.aclweb.org/anthology/P19-1150/

Obstacles hindering the development of capsule networks for challenging NLP applications include poor scalability to large output spaces and less reliable routing processes. In this paper, we introduce: (i) an agreement score to evaluate the performance of routing processes at instance-level; (ii) an adaptive optimizer to enhance the reliability of routing; (iii) capsule compression and partial routing to improve the scalability of capsule networks. We validate our approach on two NLP tasks, namely: multi-label text classification and question answering. Experimental results show that our approach considerably improves over strong competitors on both tasks. In addition, we gain the best results in low-resource settings with few training instances.

Soft Representation Learning for Sparse Transfer

- https://www.aclweb.org/anthology/P19-1151/

Transfer learning is effective for improving the performance of tasks that are related, and Multi-task learning (MTL) and Cross-lingual learning (CLL) are important instances. This paper argues that hard-parameter sharing, of hard-coding layers shared across different tasks or languages, cannot generalize well, when sharing with a loosely related task. Such case, which we call sparse transfer, might actually hurt performance, a phenomenon known as negative transfer. Our contribution is using adversarial training across tasks, to "soft-code" shared and private spaces, to avoid the shared space gets too sparse. In CLL, our proposed architecture considers another challenge of dealing with low-quality input.

Learning Representations from Imperfect Time Series Data via Tensor Rank Regularization

- https://www.aclweb.org/anthology/P19-1152/

There has been an increased interest in multimodal language processing including multimodal dialog, question answering, sentiment analysis, and speech recognition. However, naturally occurring multimodal data is often imperfect as a result of imperfect modalities, missing entries or noise corruption. To address these concerns, we present a regularization method based on tensor rank minimization. Our method is based on the observation that high-dimensional multimodal time series data often exhibit correlations across time and modalities which leads to low-rank tensor representations. However, the presence of noise or incomplete values breaks these correlations and results in tensor representations of higher rank. We design a model to learn such tensor representations and effectively regularize their rank. Experiments on multimodal language data show that our model achieves good results across various levels of imperfection.

Towards Lossless Encoding of Sentences

- https://www.aclweb.org/anthology/P19-1153/

A lot of work has been done in the field of image compression via machine learning, but not much attention has been given to the compression of natural language. Compressing text into lossless representations while making features easily retrievable is not a trivial task, yet has huge benefits. Most methods designed to produce feature rich sentence embeddings focus solely on performing well on downstream tasks and are unable to properly reconstruct the original sequence from the learned embedding. In this work, we propose a near lossless method for encoding long sequences of texts as well as all of their sub-sequences into feature rich representations. We test our method on sentiment analysis and show good performance across all sub-sentence and sentence embeddings.

Open Vocabulary Learning for Neural Chinese Pinyin IME

- https://www.aclweb.org/anthology/P19-1154/

Pinyin-to-character (P2C) conversion is the core component of pinyin-based Chinese input method engine (IME). However, the conversion is seriously compromised by the ambiguities of Chinese characters corresponding to pinyin as well as the predefined fixed vocabularies. To alleviate such inconveniences, we propose a neural P2C conversion model augmented by an online updated vocabulary with a sampling mechanism to support open vocabulary learning during IME working. Our experiments show that the proposed method outperforms commercial IMEs and state-of-the-art traditional models on standard corpus and true inputting history dataset in terms of multiple metrics and thus the online updated vocabulary indeed helps our IME effectively follows user inputting behavior.

Using LSTMs to Assess the Obligatoriness of Phonological Distinctive Features for Phonotactic Learning

- https://www.aclweb.org/anthology/P19-1155/

To ascertain the importance of phonetic information in the form of phonological distinctive features for the purpose of segment-level phonotactic acquisition, we compare the performance of two recurrent neural network models of phonotactic learning: one that has access to distinctive features at the start of the learning process, and one that does not. Though the predictions of both models are significantly correlated with human judgments of non-words, the feature-naive model significantly outperforms the feature-aware one in terms of probability assigned to a held-out test set of English words, suggesting that distinctive features are not obligatory for learning phonotactic patterns at the segment level.

Better Character Language Modeling through Morphology

- https://www.aclweb.org/anthology/P19-1156/

We incorporate morphological supervision into character language models (CLMs) via multitasking and show that this addition improves bits-per-character (BPC) performance across 24 languages, even when the morphology data and language modeling data are disjoint. Analyzing the CLMs shows that inflected words benefit more from explicitly modeling morphology than uninflected words, and that morphological supervision improves performance even as the amount of language modeling data grows. We then transfer morphological supervision across languages to improve performance in the low-resource setting.

Historical Text Normalization with Delayed Rewards

- https://www.aclweb.org/anthology/P19-1157/

Training neural sequence-to-sequence models with simple token-level log-likelihood is now a standard approach to historical text normalization, albeit often outperformed by phrase-based models. Policy gradient training enables direct optimization for exact matches, and while the small datasets in historical text normalization are prohibitive of from-scratch reinforcement learning, we show that policy gradient fine-tuning leads to significant improvements across the board. Policy gradient training, in particular, leads to more accurate normalizations for long or unseen words.

Stochastic Tokenization with a Language Model for Neural Text Classification

- https://www.aclweb.org/anthology/P19-1158/

For unsegmented languages such as Japanese and Chinese, tokenization of a sentence has a significant impact on the performance of text classification. Sentences are usually segmented with words or subwords by a morphological analyzer or byte pair encoding and then encoded with word (or subword) representations for neural networks. However, segmentation is potentially ambiguous, and it is unclear whether the segmented tokens achieve the best performance for the target task. In this paper, we propose a method to simultaneously learn tokenization and text classification to address these problems. Our model incorporates a language model for unsupervised tokenization into a text classifier and then trains both models simultaneously. To make the model robust against infrequent tokens, we sampled segmentation for each sentence stochastically during training, which resulted in improved performance of text classification. We conducted experiments on sentiment analysis as a text classification task and show that our method achieves better performance than previous methods.

Mitigating Gender Bias in Natural Language Processing: Literature Review

- https://www.aclweb.org/anthology/P19-1159/

As Natural Language Processing (NLP) and Machine Learning (ML) tools rise in popularity, it becomes increasingly vital to recognize the role they play in shaping societal biases and stereotypes. Although NLP models have shown success in modeling various applications, they propagate and may even amplify gender bias found in text corpora. While the study of bias in artificial intelligence is not new, methods to mitigate gender bias in NLP are relatively nascent. In this paper, we review contemporary studies on recognizing and mitigating

gender bias in NLP. We discuss gender bias based on four forms of representation bias and analyze methods recognizing gender bias. Furthermore, we discuss the advantages and drawbacks of existing gender debiasing methods. Finally, we discuss future studies for recognizing and mitigating gender bias in NLP.

Gender-preserving Debiasing for Pre-trained Word Embeddings

- https://www.aclweb.org/anthology/P19-1160/

Word embeddings learnt from massive text collections have demonstrated significant levels of discriminative biases such as gender, racial or ethnic biases, which in turn bias the down-stream NLP applications that use those word embeddings. Taking gender-bias as a working example, we propose a debiasing method that preserves non-discriminative gender-related information, while removing stereotypical discriminative gender biases from pre-trained word embeddings. Specifically, we consider four types of information: feminine, masculine, gender-neutral and stereotypical, which represent the relationship between gender vs. bias, and propose a debiasing method that (a) preserves the gender-related information in feminine and masculine words, (b) preserves the neutrality in gender-neutral words, and (c) removes the biases from stereotypical words. Experimental results on several previously proposed benchmark datasets show that our proposed method can debias pre-trained word embeddings better than existing SoTA methods proposed for debiasing word embeddings while preserving gender-related but non-discriminative information.

Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology

- https://www.aclweb.org/anthology/P19-1161/

Gender stereotypes are manifest in most of the world's languages and are consequently propagated or amplified by NLP systems. Although research has focused on mitigating gender stereotypes in English, the approaches that are commonly employed produce ungrammatical sentences in morphologically rich languages. We present a novel approach for converting between masculine-inflected and feminine-inflected sentences in such languages. For Spanish and Hebrew, our approach achieves F1 scores of 82% and 73% at the level of tags and accuracies of 90% and 87% at the level of forms. By evaluating our approach using four different languages, we show that, on average, it reduces gender stereotyping by a factor of 2.5 without any sacrifice to grammaticality.

A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings

- https://www.aclweb.org/anthology/P19-1162/

Word embedding models have gained a lot of traction in the Natural Language Processing community, however, they suffer from unintended demographic biases. Most approaches to evaluate these biases rely on vector space based metrics like the Word Embedding Association Test (WEAT). While these approaches offer great geometric insights into unintended biases in the embedding vector space, they fail to offer an interpretable meaning for how the embeddings could cause discrimination in downstream NLP applications. In this work, we present a transparent framework and metric for evaluating discrimination across protected groups with respect to their word embedding bias. Our metric (Relative Negative Sentiment Bias, RNSB) measures fairness in word embeddings via the relative negative sentiment associated with demographic identity terms from various protected groups. We show that our framework and metric enable useful analysis into the bias in word embeddings.

The Risk of Racial Bias in Hate Speech Detection

- https://www.aclweb.org/anthology/P19-1163/

We investigate how annotators' insensitivity to differences in dialect can lead to racial bias in automatic hate speech detection models, potentially amplifying harm against minority populations. We first uncover unexpected correlations between surface markers of African American English (AAE) and ratings of toxicity in several widely-used hate speech datasets. Then, we show that models trained on these corpora acquire and propagate these biases, such that AAE tweets and tweets by self-identified African Americans are up to two times more likely to be labelled as offensive compared to others. Finally, we propose *dialect* and *race priming* as ways to reduce the racial bias in annotation, showing that when annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive.

Evaluating Gender Bias in Machine Translation

- https://www.aclweb.org/anthology/P19-1164/

We present the first challenge set and evaluation protocol for the analysis of gender bias in machine translation (MT). Our approach uses two recent coreference resolution datasets composed of English sentences which cast participants into non-stereotypical gender roles (e.g., "The doctor asked the nurse to help her in the operation"). We devise an automatic gender bias evaluation method for eight target languages with grammatical gender, based on morphological

analysis (e.g., the use of female inflection for the word "doctor"). Our analyses show that four popular industrial MT systems and two recent state-of-the-art academic MT models are significantly prone to gender-biased translation errors for all tested target languages. Our data and code are publicly available at https://github.com/gabrielStanovsky/mt_gender.

LSTMEmbed: Learning Word and Sense Representations from a Large Semantically Annotated Corpus with Long Short-Term Memories

- https://www.aclweb.org/anthology/P19-1165/

While word embeddings are now a de facto standard representation of words in most NLP tasks, recently the attention has been shifting towards vector representations which capture the different meanings, i.e., senses, of words. In this paper we explore the capabilities of a bidirectional LSTM model to learn representations of word senses from semantically annotated corpora. We show that the utilization of an architecture that is aware of word order, like an LSTM, enables us to create better representations. We assess our proposed model on various standard benchmarks for evaluating semantic representations, reaching state-of-the-art performance on the SemEval-2014 word-to-sense similarity task. We release the code and the resulting word and sense embeddings at http://lcl.uniroma1.it/LSTMEmbed.

Understanding Undesirable Word Embedding Associations

- https://www.aclweb.org/anthology/P19-1166/

Word embeddings are often criticized for capturing undesirable word associations such as gender stereotypes. However, methods for measuring and removing such biases remain poorly understood. We show that for any embedding model that implicitly does matrix factorization, debiasing vectors post hoc using subspace projection (Bolukbasi et al., 2016) is, under certain conditions, equivalent to training on an unbiased corpus. We also prove that WEAT, the most common association test for word embeddings, systematically overestimates bias. Given that the subspace projection method is provably effective, we use it to derive a new measure of association called the relational inner product association (RIPA). Experiments with RIPA reveal that, on average, skipgram with negative sampling (SGNS) does not make most words any more gendered than they are in the training corpus. However, for gender-stereotyped words, SGNS actually amplifies the gender association in the corpus.

Unsupervised Discovery of Gendered Language through Latent-Variable Modeling

- https://www.aclweb.org/anthology/P19-1167/

Studying the ways in which language is gendered has long been an area of interest in sociolinguistics. Studies have explored, for example, the speech of male and female characters in film and the language used to describe male and female politicians. In this paper, we aim not to merely study this phenomenon qualitatively, but instead to quantify the degree to which the language used to describe men and women is different and, moreover, different in a positive or negative way. To that end, we introduce a generative latent-variable model that jointly represents adjective (or verb) choice, with its sentiment, given the natural gender of a head (or dependent) noun. We find that there are significant differences between descriptions of male and female nouns and that these differences align with common gender stereotypes: Positive adjectives used to describe women are more often related to their bodies than adjectives used to describe men.

Topic Sensitive Attention on Generic Corpora Corrects Sense Bias in Pretrained Embeddings

- https://www.aclweb.org/anthology/P19-1168/

Given a small corpus D_T pertaining to a limited set of focused topics, our goal is to train embeddings that accurately capture the sense of words in the topic in spite of the limited size of D_T. These embeddings may be used in various tasks involving D_T. A popular strategy in limited data settings is to adapt pretrained embeddings E trained on a large corpus. To correct for sense drift, fine-tuning, regularization, projection, and pivoting have been proposed recently. Among these, regularization informed by a word's corpus frequency performed well, but we improve upon it using a new regularizer based on the stability of its cooccurrence with other words. However, a thorough comparison across ten topics, spanning three tasks, with standardized settings of hyper-parameters, reveals that even the best embedding adaptation strategies provide small gains beyond well-tuned baselines, which many earlier comparisons ignored. In a bold departure from adapting pretrained embeddings, we propose using D_T to probe, attend to, and borrow fragments from any large, topic-rich source corpus (such as Wikipedia), which need not be the corpus used to pretrain embeddings. This step is made scalable and practical by suitable indexing. We reach the surprising conclusion that even limited corpus augmentation is more useful than adapting embeddings, which suggests that non-dominant sense information may be irrevocably obliterated from pretrained embeddings and cannot be salvaged by adaptation.

SphereRE: Distinguishing Lexical Relations with Hyperspherical Relation Embeddings

- https://www.aclweb.org/anthology/P19-1169/

Lexical relations describe how meanings of terms relate to each other. Typical examples include hypernymy, synonymy, meronymy, etc. Automatic distinction of lexical relations is vital for NLP applications, and also challenging due to the lack of contextual signals to discriminate between such relations. In this work, we present a neural representation learning model to distinguish lexical relations among term pairs based on Hyperspherical Relation Embeddings (SphereRE). Rather than learning embeddings for individual terms, the model learns representations of relation triples by mapping them to the hyperspherical embedding space, where relation triples of different lexical relations are well separated. Experiments over several benchmarks confirm SphereRE outperforms state-of-the-arts.

Multilingual Factor Analysis

- https://www.aclweb.org/anthology/P19-1170/

In this work we approach the task of learning multilingual word representations in an offline manner by fitting a generative latent variable model to a multilingual dictionary. We model equivalent words in different languages as different views of the same word generated by a common latent variable representing their latent lexical meaning. We explore the task of alignment by querying the fitted model for multilingual embeddings achieving competitive results across a variety of tasks. The proposed model is robust to noise in the embedding space making it a suitable method for distributed representations learned from noisy corpora.

Meaning to Form: Measuring Systematicity as Information

- https://www.aclweb.org/anthology/P19-1171/

A longstanding debate in semiotics centers on the relationship between linguistic signs and their corresponding semantics: is there an arbitrary relationship between a word form and its meaning, or does some systematic phenomenon pervade? For instance, does the character bigram ‘gl’ have any systematic relationship to the meaning of words like ‘glisten’, ‘gleam’ and ‘glow’? In this work, we offer a holistic quantification of the systematicity of the sign using mutual information and recurrent neural networks. We employ these in a data-driven and massively multilingual approach to the question, examining 106 languages. We find a statistically significant reduction in entropy when modeling a word

form conditioned on its semantic representation. Encouragingly, we also recover well-attested English examples of systematic affixes. We conclude with the meta-point: Our approximate effect size (measured in bits) is quite small—despite some amount of systematicity between form and meaning, an arbitrary relationship and its resulting benefits dominate human language.

Learning Morphosyntactic Analyzers from the Bible via Iterative Annotation Projection across 26 Languages

- https://www.aclweb.org/anthology/P19-1172/

A large percentage of computational tools are concentrated in a very small subset of the planet's languages. Compounding the issue, many languages lack the high-quality linguistic annotation necessary for the construction of such tools with current machine learning methods. In this paper, we address both issues simultaneously: leveraging the high accuracy of English taggers and parsers, we project morphological information onto translations of the Bible in 26 varied test languages. Using an iterative discovery, constraint, and training process, we build inflectional lexica in the target languages. Through a combination of iteration, ensembling, and reranking, we see double-digit relative error reductions in lemmatization and morphological analysis over a strong initial system.

Adversarial Multitask Learning for Joint Multi-Feature and Multi-Dialect Morphological Modeling

- https://www.aclweb.org/anthology/P19-1173/

Morphological tagging is challenging for morphologically rich languages due to the large target space and the need for more training data to minimize model sparsity. Dialectal variants of morphologically rich languages suffer more as they tend to be more noisy and have less resources. In this paper we explore the use of multitask learning and adversarial training to address morphological richness and dialectal variations in the context of full morphological tagging. We use multitask learning for joint morphological modeling for the features within two dialects, and as a knowledge-transfer scheme for cross-dialectal modeling. We use adversarial training to learn dialect invariant features that can help the knowledge-transfer scheme from the high to low-resource variants. We work with two dialectal variants: Modern Standard Arabic (high-resource "dialect"') and Egyptian Arabic (low-resource dialect) as a case study. Our models achieve state-of-the-art results for both. Furthermore, adversarial training provides more significant improvement when using smaller training datasets in particular.

Neural Machine Translation with Reordering Embeddings

- https://www.aclweb.org/anthology/P19-1174/

The reordering model plays an important role in phrase-based statistical machine translation. However, there are few works that exploit the reordering information in neural machine translation. In this paper, we propose a reordering mechanism to learn the reordering embedding of a word based on its contextual information. These learned reordering embeddings are stacked together with self-attention networks to learn sentence representation for machine translation. The reordering mechanism can be easily integrated into both the encoder and the decoder in the Transformer translation system. Experimental results on WMT'14 English-to-German, NIST Chinese-to-English, and WAT Japanese-to-English translation tasks demonstrate that the proposed methods can significantly improve the performance of the Transformer.

Neural Fuzzy Repair: Integrating Fuzzy Matches into Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1175/

We present a simple yet powerful data augmentation method for boosting Neural Machine Translation (NMT) performance by leveraging information retrieved from a Translation Memory (TM). We propose and test two methods for augmenting NMT training data with fuzzy TM matches. Tests on the DGT-TM data set for two language pairs show consistent and substantial improvements over a range of baseline systems. The results suggest that this method is promising for any translation environment in which a sizeable TM is available and a certain amount of repetition across translations is to be expected, especially considering its ease of implementation.

Learning Deep Transformer Models for Machine Translation

- https://www.aclweb.org/anthology/P19-1176/

Transformer is the state-of-the-art model in recent machine translation evaluations. Two strands of research are promising to improve models of this kind: the first uses wide networks (a.k.a. Transformer-Big) and has been the de facto standard for development of the Transformer system, and the other uses deeper language representation but faces the difficulty arising from learning deep networks. Here, we continue the line of research on the latter. We claim that a truly deep Transformer model can surpass the Transformer-Big counterpart by 1) proper use of layer normalization and 2) a novel way of passing the combination of previous layers to the next. On WMT'16 English-German and NIST

OpenMT'12 Chinese-English tasks, our deep system (30/25-layer encoder) outperforms the shallow Transformer-Big/Base baseline (6-layer encoder) by 0.4-2.4 BLEU points. As another bonus, the deep model is 1.6X smaller in size and 3X faster in training than Transformer-Big.

Generating Diverse Translations with Sentence Codes

- https://www.aclweb.org/anthology/P19-1177/

Users of machine translation systems may desire to obtain multiple candidates translated in different ways. In this work, we attempt to obtain diverse translations by using sentence codes to condition the sentence generation. We describe two methods to extract the codes, either with or without the help of syntax information. For diverse generation, we sample multiple candidates, each of which conditioned on a unique code. Experiments show that the sampled translations have much higher diversity scores when using reasonable sentence codes, where the translation quality is still on par with the baselines even under strong constraint imposed by the codes. In qualitative analysis, we show that our method is able to generate paraphrase translations with drastically different structures. The proposed approach can be easily adopted to existing translation systems as no modification to the model is required.

Self-Supervised Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1178/

We present a simple new method where an emergent NMT system is used for simultaneously selecting training data and learning internal NMT representations. This is done in a self-supervised way without parallel data, in such a way that both tasks enhance each other during training. The method is language independent, introduces no additional hyper-parameters, and achieves BLEU scores of 29.21 (en2fr) and 27.36 (fr2en) on newstest2014 using English and French Wikipedia data for training.

Exploring Phoneme-Level Speech Representations for End-to-End Speech Translation

- https://www.aclweb.org/anthology/P19-1179/

Previous work on end-to-end translation from speech has primarily used frame-level features as speech representations, which creates longer, sparser sequences than text. We show that a naive method to create compressed phoneme-like speech representations is far more effective and efficient for translation than traditional frame-level speech features. Specifically, we generate phoneme labels

for speech frames and average consecutive frames with the same label to create shorter, higher-level source sequences for translation. We see improvements of up to 5 BLEU on both our high and low resource language pairs, with a reduction in training time of 60%. Our improvements hold across multiple data sizes and two language pairs.

Visually Grounded Neural Syntax Acquisition

- https://www.aclweb.org/anthology/P19-1180/

We present the Visually Grounded Neural Syntax Learner (VG-NSL), an approach for learning syntactic representations and structures without any explicit supervision. The model learns by looking at natural images and reading paired captions. VG-NSL generates constituency parse trees of texts, recursively composes representations for constituents, and matches them with images. We define concreteness of constituents by their matching scores with images, and use it to guide the parsing of text. Experiments on the MSCOCO data set show that VG-NSL outperforms various unsupervised parsing approaches that do not use visual grounding, in terms of F1 scores against gold parse trees. We find that VGNSL is much more stable with respect to the choice of random initialization and the amount of training data. We also find that the concreteness acquired by VG-NSL correlates well with a similar measure defined by linguists. Finally, we also apply VG-NSL to multiple languages in the Multi30K data set, showing that our model consistently outperforms prior unsupervised approaches.

Stay on the Path: Instruction Fidelity in Vision-and-Language Navigation

- https://www.aclweb.org/anthology/P19-1181/

Advances in learning and representations have reinvigorated work that connects language to other modalities. A particularly exciting direction is Vision-and-Language Navigation(VLN), in which agents interpret natural language instructions and visual scenes to move through environments and reach goals. Despite recent progress, current research leaves unclear how much of a role language under-standing plays in this task, especially because dominant evaluation metrics have focused on goal completion rather than the sequence of actions corresponding to the instructions. Here, we highlight shortcomings of current metrics for the Room-to-Room dataset (Anderson et al.,2018b) and propose a new metric, Coverage weighted by Length Score (CLS). We also show that the existing paths in the dataset are not ideal for evaluating instruction following because they are direct-to-goal shortest paths. We join existing short paths to form more challenging extended paths to create a new data set, Room-for-Room (R4R). Using R4R and CLS, we show that agents that receive rewards for instruction fidelity outperform agents that focus on goal completion.

Expressing Visual Relationships via Language

- https://www.aclweb.org/anthology/P19-1182/

Describing images with text is a fundamental problem in vision-language research. Current studies in this domain mostly focus on single image captioning. However, in various real applications (e.g., image editing, difference interpretation, and retrieval), generating relational captions for two images, can also be very useful. This important problem has not been explored mostly due to lack of datasets and effective models. To push forward the research in this direction, we first introduce a new language-guided image editing dataset that contains a large number of real image pairs with corresponding editing instructions. We then propose a new relational speaker model based on an encoder-decoder architecture with static relational attention and sequential multi-head attention. We also extend the model with dynamic relational attention, which calculates visual alignment while decoding. Our models are evaluated on our newly collected and two public datasets consisting of image pairs annotated with relationship sentences. Experimental results, based on both automatic and human evaluation, demonstrate that our model outperforms all baselines and existing methods on all the datasets.

Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video

- https://www.aclweb.org/anthology/P19-1183/

In this paper, we address a novel task, namely weakly-supervised spatio-temporally grounding natural sentence in video. Specifically, given a natural sentence and a video, we localize a spatio-temporal tube in the video that semantically corresponds to the given sentence, with no reliance on any spatio-temporal annotations during training. First, a set of spatio-temporal tubes, referred to as instances, are extracted from the video. We then encode these instances and the sentence using our newly proposed attentive interactor which can exploit their fine-grained relationships to characterize their matching behaviors. Besides a ranking loss, a novel diversity loss is introduced to train our attentive interactor to strengthen the matching behaviors of reliable instance-sentence pairs and penalize the unreliable ones. We also contribute a dataset, called VID-sentence, based on the ImageNet video object detection dataset, to serve as a benchmark for our task. Results from extensive experiments demonstrate the superiority of our model over the baseline approaches.

The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue

- https://www.aclweb.org/anthology/P19-1184/

This paper introduces the PhotoBook dataset, a large-scale collection of visually-grounded, task-oriented dialogues in English designed to investigate shared dialogue history accumulating during conversation. Taking inspiration from seminal work on dialogue analysis, we propose a data-collection task formulated as a collaborative game prompting two online participants to refer to images utilising both their visual context as well as previously established referring expressions. We provide a detailed description of the task setup and a thorough analysis of the 2,500 dialogues collected. To further illustrate the novel features of the dataset, we propose a baseline model for reference resolution which uses a simple method to take into account shared information accumulated in a reference chain. Our results show that this information is particularly important to resolve later descriptions and underline the need to develop more sophisticated models of common ground in dialogue interaction.

Continual and Multi-Task Architecture Search

- https://www.aclweb.org/anthology/P19-1185/

Architecture search is the process of automatically learning the neural model or cell structure that best suits the given task. Recently, this approach has shown promising performance improvements (on language modeling and image classification) with reasonable training speed, using a weight sharing strategy called Efficient Neural Architecture Search (ENAS). In our work, we first introduce a novel continual architecture search (CAS) approach, so as to continually evolve the model parameters during the sequential training of several tasks, without losing performance on previously learned tasks (via block-sparsity and orthogonality constraints), thus enabling life-long learning. Next, we explore a multi-task architecture search (MAS) approach over ENAS for finding a unified, single cell structure that performs well across multiple tasks (via joint controller rewards), and hence allows more generalizable transfer of the cell structure knowledge to an unseen new task. We empirically show the effectiveness of our sequential continual learning and parallel multi-task learning based architecture search approaches on diverse sentence-pair classification tasks (GLUE) and multimodal-generation based video captioning tasks. Further, we present several ablations and analyses on the learned cell structures.

Semi-supervised Stochastic Multi-Domain Learning using Variational Inference

- https://www.aclweb.org/anthology/P19-1186/

Supervised models of NLP rely on large collections of text which closely resemble the intended testing setting. Unfortunately matching text is often not available in sufficient quantity, and moreover, within any domain of text, data is often highly heterogenous. In this paper we propose a method to distill the important domain signal as part of a multi-domain learning system, using a latent variable model in which parts of a neural model are stochastically gated based on the inferred domain. We compare the use of discrete versus continuous latent variables, operating in a domain-supervised or a domain semi-supervised setting, where the domain is known only for a subset of training inputs. We show that our model leads to substantial performance improvements over competitive benchmark domain adaptation methods, including methods using adversarial learning.

Boosting Entity Linking Performance by Leveraging Unlabeled Documents

- https://www.aclweb.org/anthology/P19-1187/

Modern entity linking systems rely on large collections of documents specifically annotated for the task (e.g., AIDA CoNLL). In contrast, we propose an approach which exploits only naturally occurring information: unlabeled documents and Wikipedia. Our approach consists of two stages. First, we construct a high recall list of candidate entities for each mention in an unlabeled document. Second, we use the candidate lists as weak supervision to constrain our document-level entity linking model. The model treats entities as latent variables and, when estimated on a collection of unlabelled texts, learns to choose entities relying both on local context of each mention and on coherence with other entities in the document. The resulting approach rivals fully-supervised state-of-the-art systems on standard test sets. It also approaches their performance in the very challenging setting: when tested on a test set sampled from the data used to estimate the supervised systems. By comparing to Wikipedia-only training of our model, we demonstrate that modeling unlabeled documents is beneficial.

Pre-Learning Environment Representations for Data-Efficient Neural Instruction Following

- https://www.aclweb.org/anthology/P19-1188/

We consider the problem of learning to map from natural language instructions to state transitions (actions) in a data-efficient manner. Our method takes inspiration from the idea that it should be easier to ground language to concepts that have already been formed through pre-linguistic observation. We augment a baseline instruction-following learner with an initial environment-learning phase

that uses observations of language-free state transitions to induce a suitable latent representation of actions before processing the instruction-following training data. We show that mapping to pre-learned representations substantially improves performance over systems whose representations are learned from limited instructional data alone.

Reinforced Training Data Selection for Domain Adaptation

- https://www.aclweb.org/anthology/P19-1189/

Supervised models suffer from the problem of domain shifting where distribution mismatch in the data across domains greatly affect model performance. To solve the problem, training data selection (TDS) has been proven to be a prospective solution for domain adaptation in leveraging appropriate data. However, conventional TDS methods normally requires a predefined threshold which is neither easy to set nor can be applied across tasks, and models are trained separately with the TDS process. To make TDS self-adapted to data and task, and to combine it with model training, in this paper, we propose a reinforcement learning (RL) framework that synchronously searches for training instances relevant to the target domain and learns better representations for them. A selection distribution generator (SDG) is designed to perform the selection and is updated according to the rewards computed from the selected data, where a predictor is included in the framework to ensure a task-specific model can be trained on the selected data and provides feedback to rewards. Experimental results from part-of-speech tagging, dependency parsing, and sentiment analysis, as well as ablation studies, illustrate that the proposed framework is not only effective in data selection and representation, but also generalized to accommodate different NLP tasks.

Generating Long and Informative Reviews with Aspect-Aware Coarse-to-Fine Decoding

- https://www.aclweb.org/anthology/P19-1190/

Generating long and informative review text is a challenging natural language generation task. Previous work focuses on word-level generation, neglecting the importance of topical and syntactic characteristics from natural languages. In this paper, we propose a novel review generation model by characterizing an elaborately designed aspect-aware coarse-to-fine generation process. First, we model the aspect transitions to capture the overall content flow. Then, to generate a sentence, an aspect-aware sketch will be predicted using an aspect-aware decoder. Finally, another decoder fills in the semantic slots by generating corresponding words. Our approach is able to jointly utilize aspect semantics, syntactic sketch, and context information. Extensive experiments results have demonstrated the effectiveness of the proposed model.

PaperRobot: Incremental Draft Generation of Scientific Ideas

- https://www.aclweb.org/anthology/P19-1191/

We present a PaperRobot who performs as an automatic research assistant by (1) conducting deep understanding of a large collection of human-written papers in a target domain and constructing comprehensive background knowledge graphs (KGs); (2) creating new ideas by predicting links from the background KGs, by combining graph attention and contextual text attention; (3) incrementally writing some key elements of a new paper based on memory-attention networks: from the input title along with predicted related entities to generate a paper , from the to generate conclusion and future work, and finally from future work to generate a title for a follow-on paper. Turing Tests, where a biomedical domain expert is asked to compare a system output and a human-authored string, show PaperRobot generated s, conclusion and future work sections, and new titles are chosen over human-written ones up to 30%, 24% and 12% of the time, respectively.

Rhetorically Controlled Encoder-Decoder for Modern Chinese Poetry Generation

- https://www.aclweb.org/anthology/P19-1192/

Rhetoric is a vital element in modern poetry, and plays an essential role in improving its aesthetics. However, to date, it has not been considered in research on automatic poetry generation. In this paper, we propose a rhetorically controlled encoder-decoder for modern Chinese poetry generation. Our model relies on a continuous latent variable as a rhetoric controller to capture various rhetorical patterns in an encoder, and then incorporates rhetoric-based mixtures while generating modern Chinese poetry. For metaphor and personification, an automated evaluation shows that our model outperforms state-of-the-art baselines by a substantial margin, while human evaluation shows that our model generates better poems than baseline methods in terms of fluency, coherence, meaningfulness, and rhetorical aesthetics.

Enhancing Topic-to-Essay Generation with External Commonsense Knowledge

- https://www.aclweb.org/anthology/P19-1193/

Automatic topic-to-essay generation is a challenging task since it requires generating novel, diverse, and topic-consistent paragraph-level text with a set of topics as input. Previous work tends to perform essay generation based solely on the given topics while ignoring massive commonsense knowledge. However, this

commonsense knowledge provides additional background information, which can help to generate essays that are more novel and diverse. Towards filling this gap, we propose to integrate commonsense from the external knowledge base into the generator through dynamic memory mechanism. Besides, the adversarial training based on a multi-label discriminator is employed to further improve topic-consistency. We also develop a series of automatic evaluation metrics to comprehensively assess the quality of the generated essay. Experiments show that with external commonsense knowledge and adversarial training, the generated essays are more novel, diverse, and topic-consistent than existing methods in terms of both automatic and human evaluation.

Towards Fine-grained Text Sentiment Transfer

- https://www.aclweb.org/anthology/P19-1194/

In this paper, we focus on the task of fine-grained text sentiment transfer (FGST). This task aims to revise an input sequence to satisfy a given sentiment intensity, while preserving the original semantic content. Different from the conventional sentiment transfer task that only reverses the sentiment polarity (positive/negative) of text, the FTST task requires more nuanced and fine-grained control of sentiment. To remedy this, we propose a novel Seq2SentiSeq model. Specifically, the numeric sentiment intensity value is incorporated into the decoder via a Gaussian kernel layer to finely control the sentiment intensity of the output. Moreover, to tackle the problem of lacking parallel data, we propose a cycle reinforcement learning algorithm to guide the model training. In this framework, the elaborately designed rewards can balance both sentiment transformation and content preservation, while not requiring any ground truth output. Experimental results show that our approach can outperform existing methods by a large margin in both automatic evaluation and human evaluation.

Data-to-text Generation with Entity Modeling

- https://www.aclweb.org/anthology/P19-1195/

Recent approaches to data-to-text generation have shown great promise thanks to the use of large-scale datasets and the application of neural network architectures which are trained end-to-end. These models rely on representation learning to select content appropriately, structure it coherently, and verbalize it grammatically, treating entities as nothing more than vocabulary tokens. In this work we propose an entity-centric neural architecture for data-to-text generation. Our model creates entity-specific representations which are dynamically updated. Text is generated conditioned on the data input and entity memory representations using hierarchical attention at each time step. We present experiments on the RotoWire benchmark and a (five times larger) new dataset

on the baseball domain which we create. Our results show that the proposed model outperforms competitive baselines in automatic and human evaluation.

Ensuring Readability and Data-fidelity using Head-modifier Templates in Deep Type Description Generation

- https://www.aclweb.org/anthology/P19-1196/

A type description is a succinct noun compound which helps human and machines to quickly grasp the informative and distinctive information of an entity. Entities in most knowledge graphs (KGs) still lack such descriptions, thus calling for automatic methods to supplement such information. However, existing generative methods either overlook the grammatical structure or make factual mistakes in generated texts. To solve these problems, we propose a head-modifier template based method to ensure the readability and data fidelity of generated type descriptions. We also propose a new dataset and two metrics for this task. Experiments show that our method improves substantially compared with baselines and achieves state-of-the-art performance on both datasets.

Key Fact as Pivot: A Two-Stage Model for Low Resource Table-to-Text Generation

- https://www.aclweb.org/anthology/P19-1197/

Table-to-text generation aims to translate the structured data into the unstructured text. Most existing methods adopt the encoder-decoder framework to learn the transformation, which requires large-scale training samples. However, the lack of large parallel data is a major practical problem for many domains. In this work, we consider the scenario of low resource table-to-text generation, where only limited parallel data is available. We propose a novel model to separate the generation into two stages: key fact prediction and surface realization. It first predicts the key facts from the tables, and then generates the text with the key facts. The training of key fact prediction needs much fewer annotated data, while surface realization can be trained with pseudo parallel corpus. We evaluate our model on a biography generation dataset. Our model can achieve 27.34 BLEU score with only 1,000 parallel data, while the baseline model only obtain the performance of 9.71 BLEU score.

Unsupervised Neural Text Simplification

- https://www.aclweb.org/anthology/P19-1198/

The paper presents a first attempt towards unsupervised neural text simplification that relies only on unlabeled text corpora. The core framework is com-

posed of a shared encoder and a pair of attentional-decoders, crucially assisted by discrimination-based losses and denoising. The framework is trained using unlabeled text collected from en-Wikipedia dump. Our analysis (both quantitative and qualitative involving human evaluators) on public test data shows that the proposed model can perform text-simplification at both lexical and syntactic levels, competitive to existing supervised methods. It also outperforms viable unsupervised baselines. Adding a few labeled pairs helps improve the performance further.

Syntax-Infused Variational Autoencoder for Text Generation

- https://www.aclweb.org/anthology/P19-1199/

We present a syntax-infused variational autoencoder (SIVAE), that integrates sentences with their syntactic trees to improve the grammar of generated sentences. Distinct from existing VAE-based text generative models, SIVAE contains two separate latent spaces, for sentences and syntactic trees. The evidence lower bound objective is redesigned correspondingly, by optimizing a joint distribution that accommodates two encoders and two decoders. SIVAE works with long short-term memory architectures to simultaneously generate sentences and syntactic trees. Two versions of SIVAE are proposed: one captures the dependencies between the latent variables through a conditional prior network, and the other treats the latent variables independently such that syntactically-controlled sentence generation can be performed. Experimental results demonstrate the generative superiority of SIVAE on both reconstruction and targeted syntactic evaluations. Finally, we show that the proposed models can be used for unsupervised paraphrasing given different syntactic tree templates.

Towards Generating Long and Coherent Text with Multi-Level Latent Variable Models

- https://www.aclweb.org/anthology/P19-1200/

Variational autoencoders (VAEs) have received much attention recently as an end-to-end architecture for text generation with latent variables. However, previous works typically focus on synthesizing relatively short sentences (up to 20 words), and the posterior collapse issue has been widely identified in text-VAEs. In this paper, we propose to leverage several multi-level structures to learn a VAE model for generating long, and coherent text. In particular, a hierarchy of stochastic layers between the encoder and decoder networks is employed to more informative and semantic-rich latent codes. Besides, we utilize a multi-level decoder structure to capture the coherent long-term structure inherent in long-form texts, by generating intermediate sentence representations

as high-level plan vectors. Extensive experimental results demonstrate that the proposed multi-level VAE model produces more coherent and less repetitive long text compared to baselines as well as can mitigate the posterior-collapse issue.

Jointly Learning Semantic Parser and Natural Language Generator via Dual Information Maximization

- https://www.aclweb.org/anthology/P19-1201/

Semantic parsing aims to transform natural language (NL) utterances into formal meaning representations (MRs), whereas an NL generator achieves the reverse: producing an NL description for some given MRs. Despite this intrinsic connection, the two tasks are often studied separately in prior work. In this paper, we model the duality of these two tasks via a joint learning framework, and demonstrate its effectiveness of boosting the performance on both tasks. Concretely, we propose a novel method of dual information maximization (DIM) to regularize the learning process, where DIM empirically maximizes the variational lower bounds of expected joint distributions of NL and MRs. We further extend DIM to a semi-supervision setup (SemiDIM), which leverages unlabeled data of both tasks. Experiments on three datasets of dialogue management and code generation (and summarization) show that performance on both semantic parsing and NL generation can be consistently improved by DIM, in both supervised and semi-supervised setups.

Learning to Select, Track, and Generate for Data-to-Text

- https://www.aclweb.org/anthology/P19-1202/

We propose a data-to-text generation model with two modules, one for tracking and the other for text generation. Our tracking module selects and keeps track of salient information and memorizes which record has been mentioned. Our generation module generates a summary conditioned on the state of tracking module. Our proposed model is considered to simulate the human-like writing process that gradually selects the information by determining the intermediate variables while writing the summary. In addition, we also explore the effectiveness of the writer information for generations. Experimental results show that our proposed model outperforms existing models in all evaluation metrics even without writer information. Incorporating writer information further improves the performance, contributing to content planning and surface realization.

Reinforced Dynamic Reasoning for Conversational Question Generation

- https://www.aclweb.org/anthology/P19-1203/

This paper investigates a new task named Conversational Question Generation (CQG) which is to generate a question based on a passage and a conversation history (i.e., previous turns of question-answer pairs). CQG is a crucial task for developing intelligent agents that can drive question-answering style conversations or test user understanding of a given passage. Towards that end, we propose a new approach named Reinforced Dynamic Reasoning network, which is based on the general encoder-decoder framework but incorporates a reasoning procedure in a dynamic manner to better understand what has been asked and what to ask next about the passage into the general encoder-decoder framework. To encourage producing meaningful questions, we leverage a popular question answering (QA) model to provide feedback and fine-tune the question generator using a reinforcement learning mechanism. Empirical results on the recently released CoQA dataset demonstrate the effectiveness of our method in comparison with various baselines and model variants. Moreover, to show the applicability of our method, we also apply it to create multi-turn question-answering conversations for passages in SQuAD.

TalkSumm: A Dataset and Scalable Annotation Method for Scientific Paper Summarization Based on Conference Talks

- https://www.aclweb.org/anthology/P19-1204/

Currently, no large-scale training data is available for the task of scientific paper summarization. In this paper, we propose a novel method that automatically generates summaries for scientific papers, by utilizing videos of talks at scientific conferences. We hypothesize that such talks constitute a coherent and concise description of the papers' content, and can form the basis for good summaries. We collected 1716 papers and their corresponding videos, and created a dataset of paper summaries. A model trained on this dataset achieves similar performance as models trained on a dataset of summaries created manually. In addition, we validated the quality of our summaries by human experts.

Improving ive Document Summarization with Salient Information Modeling

- https://www.aclweb.org/anthology/P19-1205/

Comprehensive document encoding and salient information selection are two major difficulties for generating summaries with adequate salient information. To tackle the above difficulties, we propose a Transformer-based encoder-decoder framework with two novel extensions for ive document summarization. Specifically, (1) to encode the documents comprehensively, we design a focus-attention

mechanism and incorporate it into the encoder. This mechanism models a Gaussian focal bias on attention scores to enhance the perception of local context, which contributes to producing salient and informative summaries. (2) To distinguish salient information precisely, we design an independent saliency-selection network which manages the information flow from encoder to decoder. This network effectively reduces the influences of secondary information on the generated summaries. Experimental results on the popular CNN/Daily Mail benchmark demonstrate that our model outperforms other state-of-the-art baselines on the ROUGE metrics.

Unsupervised Neural Single-Document Summarization of Reviews via Learning Latent Discourse Structure and its Ranking

- https://www.aclweb.org/anthology/P19-1206/

This paper focuses on the end-to-end ive summarization of a single product review without supervision. We assume that a review can be described as a discourse tree, in which the summary is the root, and the child sentences explain their parent in detail. By recursively estimating a parent from its children, our model learns the latent discourse tree without an external parser and generates a concise summary. We also introduce an architecture that ranks the importance of each sentence on the tree to support summary generation focusing on the main review point. The experimental results demonstrate that our model is competitive with or outperforms other unsupervised approaches. In particular, for relatively long reviews, it achieves a competitive or better performance than supervised models. The induced tree shows that the child sentences provide additional information about their parent, and the generated summary s the entire review.

BiSET: Bi-directional Selective Encoding with Template for ive Summarization

- https://www.aclweb.org/anthology/P19-1207/

The success of neural summarization models stems from the meticulous encodings of source articles. To overcome the impediments of limited and sometimes noisy training data, one promising direction is to make better use of the available training data by applying filters during summarization. In this paper, we propose a novel Bi-directional Selective Encoding with Template (BiSET) model, which leverages template discovered from training data to softly select key information from each source article to guide its summarization process. Extensive experiments on a standard summarization dataset are conducted and the results show that the template-equipped BiSET model manages to improve the

summarization performance significantly with a new state of the art.

Neural Keyphrase Generation via Reinforcement Learning with Adaptive Rewards

- https://www.aclweb.org/anthology/P19-1208/

Generating keyphrases that summarize the main points of a document is a fundamental task in natural language processing. Although existing generative models are capable of predicting multiple keyphrases for an input document as well as determining the number of keyphrases to generate, they still suffer from the problem of generating too few keyphrases. To address this problem, we propose a reinforcement learning (RL) approach for keyphrase generation, with an adaptive reward function that encourages a model to generate both sufficient and accurate keyphrases. Furthermore, we introduce a new evaluation method that incorporates name variations of the ground-truth keyphrases using the Wikipedia knowledge base. Thus, our evaluation method can more robustly evaluate the quality of predicted keyphrases. Extensive experiments on five real-world datasets of different scales demonstrate that our RL approach consistently and significantly improves the performance of the state-of-the-art generative models with both conventional and new evaluation methods.

Scoring Sentence Singletons and Pairs for ive Summarization

- https://www.aclweb.org/anthology/P19-1209/

When writing a summary, humans tend to choose content from one or two sentences and merge them into a single summary sentence. However, the mechanisms behind the selection of one or multiple source sentences remain poorly understood. Sentence fusion assumes multi-sentence input; yet sentence selection methods only work with single sentences and not combinations of them. There is thus a crucial gap between sentence selection and fusion to support summarizing by both compressing single sentences and fusing pairs. This paper attempts to bridge the gap by ranking sentence singletons and pairs together in a unified space. Our proposed framework attempts to model human methodology by selecting either a single sentence or a pair of sentences, then compressing or fusing the sentence(s) to produce a summary sentence. We conduct extensive experiments on both single- and multi-document summarization datasets and report findings on sentence selection and ion.

Keep Meeting Summaries on Topic: ive Multi-Modal Meeting Summarization

- https://www.aclweb.org/anthology/P19-1210/

Transcripts of natural, multi-person meetings differ significantly from documents like news articles, which can make Natural Language Generation models for generating summaries unfocused. We develop an ive meeting summarizer from both videos and audios of meeting recordings. Specifically, we propose a multi-modal hierarchical attention across three levels: segment, utterance and word. To narrow down the focus into topically-relevant segments, we jointly model topic segmentation and summarization. In addition to traditional text features, we introduce new multi-modal features derived from visual focus of attention, based on the assumption that the utterance is more important if the speaker receives more attention. Experiments show that our model significantly outperforms the state-of-the-art with both BLEU and ROUGE measures.

Adversarial Domain Adaptation Using Artificial Titles for ive Title Generation

- https://www.aclweb.org/anthology/P19-1211/

A common issue in training a deep learning, ive summarization model is lack of a large set of training summaries. This paper examines techniques for adapting from a labeled source domain to an unlabeled target domain in the context of an encoder-decoder model for text generation. In addition to adversarial domain adaptation (ADA), we introduce the use of artificial titles and sequential training to capture the grammatical style of the unlabeled target domain. Evaluation on adapting to/from news articles and Stack Exchange posts indicates that the use of these techniques can boost performance for both unsupervised adaptation as well as fine-tuning with limited target data.

BIGPATENT: A Large-Scale Dataset for ive and Coherent Summarization

- https://www.aclweb.org/anthology/P19-1212/

Most existing text summarization datasets are compiled from the news domain, where summaries have a flattened discourse structure. In such datasets, summary-worthy content often appears in the beginning of input articles. Moreover, large segments from input articles are present verbatim in their respective summaries. These issues impede the learning and evaluation of systems that can understand an article's global content structure as well as produce ive summaries with high compression ratio. In this work, we present a novel dataset, BIGPATENT, consisting of 1.3 million records of U.S. patent documents along with

human written ive summaries. Compared to existing summarization datasets, BIGPATENT has the following properties: i) summaries contain a richer discourse structure with more recurring entities, ii) salient content is evenly distributed in the input, and iii) lesser and shorter extractive fragments are present in the summaries. Finally, we train and evaluate baselines and popular learning models on BIGPATENT to shed light on new challenges and motivate future directions for summarization research.

Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference

- https://www.aclweb.org/anthology/P19-1213/

While recent progress on ive summarization has led to remarkably fluent summaries, factual errors in generated summaries still severely limit their use in practice. In this paper, we evaluate summaries produced by state-of-the-art models via crowdsourcing and show that such errors occur frequently, in particular with more ive models. We study whether textual entailment predictions can be used to detect such errors and if they can be reduced by reranking alternative predicted summaries. That leads to an interesting downstream application for entailment models. In our experiments, we find that out-of-the-box entailment models trained on NLI datasets do not yet offer the desired performance for the downstream task and we therefore release our annotations as additional test data for future extrinsic evaluations of NLI.

Self-Supervised Learning for Contextualized Extractive Summarization

- https://www.aclweb.org/anthology/P19-1214/

Existing models for extractive summarization are usually trained from scratch with a cross-entropy loss, which does not explicitly capture the global context at the document level. In this paper, we aim to improve this task by introducing three auxiliary pre-training tasks that learn to capture the document-level context in a self-supervised fashion. Experiments on the widely-used CNN/DM dataset validate the effectiveness of the proposed auxiliary tasks. Furthermore, we show that after pre-training, a clean model with simple building blocks is able to outperform previous state-of-the-art that are carefully designed.

On the Summarization of Consumer Health Questions

- https://www.aclweb.org/anthology/P19-1215/

Question understanding is one of the main challenges in question answering. In real world applications, users often submit natural language questions that are longer than needed and include peripheral information that increases the complexity of the question, leading to substantially more false positives in answer retrieval. In this paper, we study neural ive models for medical question summarization. We introduce the MeQSum corpus of 1,000 summarized consumer health questions. We explore data augmentation methods and evaluate state-of-the-art neural ive models on this new task. In particular, we show that semantic augmentation from question datasets improves the overall performance, and that pointer-generator networks outperform sequence-to-sequence attentional models on this task, with a ROUGE-1 score of 44.16%. We also present a detailed error analysis and discuss directions for improvement that are specific to question summarization.

Unsupervised Rewriter for Multi-Sentence Compression

- https://www.aclweb.org/anthology/P19-1216/

Multi-sentence compression (MSC) aims to generate a grammatical but reduced compression from multiple input sentences while retaining their key information. Previous dominating approach for MSC is the extraction-based word graph approach. A few variants further leveraged lexical substitution to yield more ive compression. However, two limitations exist. First, the word graph approach that simply concatenates fragments from multiple sentences may yield non-fluent or ungrammatical compression. Second, lexical substitution is often inappropriate without the consideration of context information. To tackle the above-mentioned issues, we present a neural rewriter for multi-sentence compression that does not need any parallel corpus. Empirical studies have shown that our approach achieves comparable results upon automatic evaluation and improves the grammaticality of compression based on human evaluation. A parallel corpus with more than 140,000 (sentence group, compression) pairs is also constructed as a by-product for future research.

Inferential Machine Comprehension: Answering Questions by Recursively Deducing the Evidence Chain from Text

- https://www.aclweb.org/anthology/P19-1217/

This paper focuses on the topic of inferential machine comprehension, which aims to fully understand the meanings of given text to answer generic questions, especially the ones needed reasoning skills. In particular, we first encode the given document, question and options in a context aware way. We then propose a new network to solve the inference problem by decomposing it into a series of attention-based reasoning steps. The result of the previous step acts as the context of next step. To make each step can be directly inferred from the text,

we design an operational cell with prior structure. By recursively linking the cells, the inferred results are synthesized together to form the evidence chain for reasoning, where the reasoning direction can be guided by imposing structural constraints to regulate interactions on the cells. Moreover, a termination mechanism is introduced to dynamically determine the uncertain reasoning depth, and the network is trained by reinforcement learning. Experimental results on 3 popular data sets, including MCTest, RACE and MultiRC, demonstrate the effectiveness of our approach.

Token-level Dynamic Self-Attention Network for Multi-Passage Reading Comprehension

- https://www.aclweb.org/anthology/P19-1218/

Multi-passage reading comprehension requires the ability to combine cross-passage information and reason over multiple passages to infer the answer. In this paper, we introduce the Dynamic Self-attention Network (DynSAN) for multi-passage reading comprehension task, which processes cross-passage information at token-level and meanwhile avoids substantial computational costs. The core module of the dynamic self-attention is a proposed gated token selection mechanism, which dynamically selects important tokens from a sequence. These chosen tokens will attend to each other via a self-attention mechanism to model long-range dependencies. Besides, convolutional layers are combined with the dynamic self-attention to enhance the model's capacity of extracting local semantic. The experimental results show that the proposed DynSAN achieves new state-of-the-art performance on the SearchQA, Quasar-T and WikiHop datasets. Further ablation study also validates the effectiveness of our model components.

Explicit Utilization of General Knowledge in Machine Reading Comprehension

- https://www.aclweb.org/anthology/P19-1219/

To bridge the gap between Machine Reading Comprehension (MRC) models and human beings, which is mainly reflected in the hunger for data and the robustness to noise, in this paper, we explore how to integrate the neural networks of MRC models with the general knowledge of human beings. On the one hand, we propose a data enrichment method, which uses WordNet to extract inter-word semantic connections as general knowledge from each given passage-question pair. On the other hand, we propose an end-to-end MRC model named as Knowledge Aided Reader (KAR), which explicitly uses the above extracted general knowledge to assist its attention mechanisms. Based on the data enrichment method, KAR is comparable in performance with the state-of-the-art

MRC models, and significantly more robust to noise than them. When only a subset (20%-80%) of the training examples are available, KAR outperforms the state-of-the-art MRC models by a large margin, and is still reasonably robust to noise.

Multi-style Generative Reading Comprehension

- https://www.aclweb.org/anthology/P19-1220/

This study tackles generative reading comprehension (RC), which consists of answering questions based on textual evidence and natural language generation (NLG). We propose a multi-style ive summarization model for question answering, called Masque. The proposed model has two key characteristics. First, unlike most studies on RC that have focused on extracting an answer span from the provided passages, our model instead focuses on generating a summary from the question and multiple passages. This serves to cover various answer styles required for real-world applications. Second, whereas previous studies built a specific model for each answer style because of the difficulty of acquiring one general model, our approach learns multi-style answers within a model to improve the NLG capability for all styles involved. This also enables our model to give an answer in the target style. Experiments show that our model achieves state-of-the-art performance on the Q&A task and the Q&A + NLG task of MS MARCO 2.1 and the summary task of NarrativeQA. We observe that the transfer of the style-independent NLG capability to the target style is the key to its success.

Retrieve, Read, Rerank: Towards End-to-End Multi-Document Reading Comprehension

- https://www.aclweb.org/anthology/P19-1221/

This paper considers the reading comprehension task in which multiple documents are given as input. Prior work has shown that a pipeline of retriever, reader, and reranker can improve the overall performance. However, the pipeline system is inefficient since the input is re-encoded within each module, and is unable to leverage upstream components to help downstream training. In this work, we present RE3QA, a unified question answering model that combines context retrieving, reading comprehension, and answer reranking to predict the final answer. Unlike previous pipelined approaches, RE3QA shares contextualized text representation across different components, and is carefully designed to use high-quality upstream outputs (e.g., retrieved context or candidate answers) for directly supervising downstream modules (e.g., the reader or the reranker). As a result, the whole network can be trained end-to-end to avoid the context inconsistency problem. Experiments show that our model outper-

forms the pipelined baseline and achieves state-of-the-art results on two versions of TriviaQA and two variants of SQuAD.

Multi-Hop Paragraph Retrieval for Open-Domain Question Answering

- https://www.aclweb.org/anthology/P19-1222/

This paper is concerned with the task of multi-hop open-domain Question Answering (QA). This task is particularly challenging since it requires the simultaneous performance of textual reasoning and efficient searching. We present a method for retrieving multiple supporting paragraphs, nested amidst a large knowledge base, which contain the necessary evidence to answer a given question. Our method iteratively retrieves supporting paragraphs by forming a joint vector representation of both a question and a paragraph. The retrieval is performed by considering contextualized sentence-level representations of the paragraphs in the knowledge source. Our method achieves state-of-the-art performance over two well-known datasets, SQuAD-Open and HotpotQA, which serve as our single- and multi-hop open-domain QA benchmarks, respectively.

E3: Entailment-driven Extracting and Editing for Conversational Machine Reading

- https://www.aclweb.org/anthology/P19-1223/

Conversational machine reading systems help users answer high-level questions (e.g. determine if they qualify for particular government benefits) when they do not know the exact rules by which the determination is made (e.g. whether they need certain income levels or veteran status). The key challenge is that these rules are only provided in the form of a procedural text (e.g. guidelines from government website) which the system must read to figure out what to ask the user. We present a new conversational machine reading model that jointly extracts a set of decision rules from the procedural text while reasoning about which are entailed by the conversational history and which still need to be edited to create questions for the user. On the recently introduced ShARC conversational machine reading dataset, our Entailment-driven Extract and Edit network (E3) achieves a new state-of-the-art, outperforming existing systems as well as a new BERT-based baseline. In addition, by explicitly highlighting which information still needs to be gathered, E3 provides a more explainable alternative to prior work. We release source code for our models and experiments at https://github.com/vzhong/e3.

Generating Question-Answer Hierarchies

- https://www.aclweb.org/anthology/P19-1224/

The process of knowledge acquisition can be viewed as a question-answer game between a student and a teacher in which the student typically starts by asking broad, open-ended questions before drilling down into specifics (Hintikka, 1981; Hakkarainen and Sintonen, 2002). This pedagogical perspective motivates a new way of representing documents. In this paper, we present SQUASH (Specificity-controlled Question-Answer Hierarchies), a novel and challenging text generation task that converts an input document into a hierarchy of question-answer pairs. Users can click on high-level questions (e.g., "Why did Frodo leave the Fellowship?") to reveal related but more specific questions (e.g., "Who did Frodo leave with?"). Using a question taxonomy loosely based on Lehnert (1978), we classify questions in existing reading comprehension datasets as either GENERAL or SPECIFIC . We then use these labels as input to a pipelined system centered around a conditional neural language model. We extensively evaluate the quality of the generated QA hierarchies through crowdsourced experiments and report strong empirical results.

Answering while Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction

- https://www.aclweb.org/anthology/P19-1225/

Question answering (QA) using textual sources for purposes such as reading comprehension (RC) has attracted much attention. This study focuses on the task of explainable multi-hop QA, which requires the system to return the answer with evidence sentences by reasoning and gathering disjoint pieces of the reference texts. It proposes the Query Focused Extractor (QFE) model for evidence extraction and uses multi-task learning with the QA model. QFE is inspired by extractive summarization models; compared with the existing method, which extracts each evidence sentence independently, it sequentially extracts evidence sentences by using an RNN with an attention mechanism on the question sentence. It enables QFE to consider the dependency among the evidence sentences and cover important information in the question sentence. Experimental results show that QFE with a simple RC baseline model achieves a state-of-the-art evidence extraction score on HotpotQA. Although designed for RC, it also achieves a state-of-the-art evidence extraction score on FEVER, which is a recognizing textual entailment task on a large textual database.

Enhancing Pre-Trained Language Representations with Rich Knowledge for Machine Reading Comprehension

- https://www.aclweb.org/anthology/P19-1226/

Machine reading comprehension (MRC) is a crucial and challenging task in NLP. Recently, pre-trained language models (LMs), especially BERT, have achieved remarkable success, presenting new state-of-the-art results in MRC. In this work, we investigate the potential of leveraging external knowledge bases (KBs) to further improve BERT for MRC. We introduce KT-NET, which employs an attention mechanism to adaptively select desired knowledge from KBs, and then fuses selected knowledge with BERT to enable context- and knowledge-aware predictions. We believe this would combine the merits of both deep LMs and curated KBs towards better MRC. Experimental results indicate that KT-NET offers significant and consistent improvements over BERT, outperforming competitive baselines on ReCoRD and SQuAD1.1 benchmarks. Notably, it ranks the 1st place on the ReCoRD leaderboard, and is also the best single model on the SQuAD1.1 leaderboard at the time of submission (March 4th, 2019).

XQA: A Cross-lingual Open-domain Question Answering Dataset

- https://www.aclweb.org/anthology/P19-1227/

Open-domain question answering (OpenQA) aims to answer questions through text retrieval and reading comprehension. Recently, lots of neural network-based models have been proposed and achieved promising results in OpenQA. However, the success of these models relies on a massive volume of training data (usually in English), which is not available in many other languages, especially for those low-resource languages. Therefore, it is essential to investigate cross-lingual OpenQA. In this paper, we construct a novel dataset XQA for cross-lingual OpenQA research. It consists of a training set in English as well as development and test sets in eight other languages. Besides, we provide several baseline systems for cross-lingual OpenQA, including two machine translation-based methods and one zero-shot cross-lingual method (multilingual BERT). Experimental results show that the multilingual BERT model achieves the best results in almost all target languages, while the performance of cross-lingual OpenQA is still much lower than that of English. Our analysis indicates that the performance of cross-lingual OpenQA is related to not only how similar the target language and English are, but also how difficult the question set of the target language is. The XQA dataset is publicly available at http://github.com/thunlp/XQA.

Compound Probabilistic Context-Free Grammars for Grammar Induction

- https://www.aclweb.org/anthology/P19-1228/

We study a formalization of the grammar induction problem that models sentences as being generated by a compound probabilistic context free grammar.

In contrast to traditional formulations which learn a single stochastic grammar, our context-free rule probabilities are modulated by a per-sentence continuous latent variable, which induces marginal dependencies beyond the traditional context-free assumptions. Inference in this context-dependent grammar is performed by collapsed variational inference, in which an amortized variational posterior is placed on the continuous variable, and the latent trees are marginalized with dynamic programming. Experiments on English and Chinese show the effectiveness of our approach compared to recent state-of-the-art methods for grammar induction from words with neural language models.

Semi-supervised Domain Adaptation for Dependency Parsing

- https://www.aclweb.org/anthology/P19-1229/

During the past decades, due to the lack of sufficient labeled data, most studies on cross-domain parsing focus on unsupervised domain adaptation, assuming there is no target-domain training data. However, unsupervised approaches make limited progress so far due to the intrinsic difficulty of both domain adaptation and parsing. This paper tackles the semi-supervised domain adaptation problem for Chinese dependency parsing, based on two newly-annotated large-scale domain-aware datasets. We propose a simple domain embedding approach to merge the source- and target-domain training data, which is shown to be more effective than both direct corpus concatenation and multi-task learning. In order to utilize unlabeled target-domain data, we employ the recent contextualized word representations and show that a simple fine-tuning procedure can further boost cross-domain parsing accuracy by large margin.

Head-Driven Phrase Structure Grammar Parsing on Penn Treebank

- https://www.aclweb.org/anthology/P19-1230/

Head-driven phrase structure grammar (HPSG) enjoys a uniform formalism representing rich contextual syntactic and even semantic meanings. This paper makes the first attempt to formulate a simplified HPSG by integrating constituent and dependency formal representations into head-driven phrase structure. Then two parsing algorithms are respectively proposed for two converted tree representations, division span and joint span. As HPSG encodes both constituent and dependency structure information, the proposed HPSG parsers may be regarded as a sort of joint decoder for both types of structures and thus are evaluated in terms of extracted or converted constituent and dependency parsing trees. Our parser achieves new state-of-the-art performance for both parsing tasks on Penn Treebank (PTB) and Chinese Penn Treebank, verifying

the effectiveness of joint learning constituent and dependency structures. In details, we report 95.84 F1 of constituent parsing and 97.00% UAS of dependency parsing on PTB.

Distantly Supervised Named Entity Recognition using Positive-Unlabeled Learning

- https://www.aclweb.org/anthology/P19-1231/

In this work, we explore the way to perform named entity recognition (NER) using only unlabeled data and named entity dictionaries. To this end, we formulate the task as a positive-unlabeled (PU) learning problem and accordingly propose a novel PU learning algorithm to perform the task. We prove that the proposed algorithm can unbiasedly and consistently estimate the task loss as if there is fully labeled data. A key feature of the proposed method is that it does not require the dictionaries to label every entity within a sentence, and it even does not require the dictionaries to label all of the words constituting an entity. This greatly reduces the requirement on the quality of the dictionaries and makes our method generalize well with quite simple dictionaries. Empirical studies on four public NER datasets demonstrate the effectiveness of our proposed method. We have published the source code at https://github.com/v-mipeng/LexiconNER.

Multi-Task Semantic Dependency Parsing with Policy Gradient for Learning Easy-First Strategies

- https://www.aclweb.org/anthology/P19-1232/

In Semantic Dependency Parsing (SDP), semantic relations form directed acyclic graphs, rather than trees. We propose a new iterative predicate selection (IPS) algorithm for SDP. Our IPS algorithm combines the graph-based and transition-based parsing approaches in order to handle multiple semantic head words. We train the IPS model using a combination of multi-task learning and task-specific policy gradient training. Trained this way, IPS achieves a new state of the art on the SemEval 2015 Task 18 datasets. Furthermore, we observe that policy gradient training learns an easy-first strategy.

GCDT: A Global Context Enhanced Deep Transition Architecture for Sequence Labeling

- https://www.aclweb.org/anthology/P19-1233/

Current state-of-the-art systems for sequence labeling are typically based on the family of Recurrent Neural Networks (RNNs). However, the shallow connections

between consecutive hidden states of RNNs and insufficient modeling of global information restrict the potential performance of those models. In this paper, we try to address these issues, and thus propose a Global Context enhanced Deep Transition architecture for sequence labeling named GCDT. We deepen the state transition path at each position in a sentence, and further assign every token with a global representation learned from the entire sentence. Experiments on two standard sequence labeling tasks show that, given only training data and the ubiquitous word embeddings (Glove), our GCDT achieves 91.96 F1 on the CoNLL03 NER task and 95.43 F1 on the CoNLL2000 Chunking task, which outperforms the best reported results under the same settings. Furthermore, by leveraging BERT as an additional resource, we establish new state-of-the-art results with 93.47 F1 on NER and 97.30 F1 on Chunking.

Unsupervised Learning of PCFGs with Normalizing Flow

- https://www.aclweb.org/anthology/P19-1234/

Unsupervised PCFG inducers hypothesize sets of compact context-free rules as explanations for sentences. PCFG induction not only provides tools for low-resource languages, but also plays an important role in modeling language acquisition (Bannard et al., 2009; Abend et al. 2017). However, current PCFG induction models, using word tokens as input, are unable to incorporate semantics and morphology into induction, and may encounter issues of sparse vocabulary when facing morphologically rich languages. This paper describes a neural PCFG inducer which employs context embeddings (Peters et al., 2018) in a normalizing flow model (Dinh et al., 2015) to extend PCFG induction to use semantic and morphological information. Linguistically motivated sparsity and categorical distance constraints are imposed on the inducer as regularization. Experiments show that the PCFG induction model with normalizing flow produces grammars with state-of-the-art accuracy on a variety of different languages. Ablation further shows a positive effect of normalizing flow, context embeddings and proposed regularizers.

Variance of Average Surprisal: A Better Predictor for Quality of Grammar from Unsupervised PCFG Induction

- https://www.aclweb.org/anthology/P19-1235/

In unsupervised grammar induction, data likelihood is known to be only weakly correlated with parsing accuracy, especially at convergence after multiple runs. In order to find a better indicator for quality of induced grammars, this paper correlates several linguistically- and psycholinguistically-motivated predictors to parsing accuracy on a large multilingual grammar induction evaluation data set. Results show that variance of average surprisal (VAS) better correlates with parsing accuracy than data likelihood and that using VAS instead of data

likelihood for model selection provides a significant accuracy boost. Further evidence shows VAS to be a better candidate than data likelihood for predicting word order typology classification. Analyses show that VAS seems to separate content words from function words in natural language grammars, and to better arrange words with different frequencies into separate classes that are more consistent with linguistic theory.

Cross-Domain NER using Cross-Domain Language Modeling

- https://www.aclweb.org/anthology/P19-1236/

Due to limitation of labeled resources, cross-domain named entity recognition (NER) has been a challenging task. Most existing work considers a supervised setting, making use of labeled data for both the source and target domains. A disadvantage of such methods is that they cannot train for domains without NER data. To address this issue, we consider using cross-domain LM as a bridge cross-domains for NER domain adaptation, performing cross-domain and cross-task knowledge transfer by designing a novel parameter generation network. Results show that our method can effectively extract domain differences from cross-domain LM contrast, allowing unsupervised domain adaptation while also giving state-of-the-art results among supervised domain adaptation methods.

Graph-based Dependency Parsing with Graph Neural Networks

- https://www.aclweb.org/anthology/P19-1237/

We investigate the problem of efficiently incorporating high-order features into neural graph-based dependency parsing. Instead of explicitly extracting high-order features from intermediate parse trees, we develop a more powerful dependency tree node representation which captures high-order information concisely and efficiently. We use graph neural networks (GNNs) to learn the representations and discuss several new configurations of GNN's updating and aggregation functions. Experiments on PTB show that our parser achieves the best UAS and LAS on PTB (96.0%, 94.3%) among systems without using any external resources.

Wide-Coverage Neural A* Parsing for Minimalist Grammars

- https://www.aclweb.org/anthology/P19-1238/

Minimalist Grammars (Stabler, 1997) are a computationally oriented, and rigorous formalisation of many aspects of Chomsky's (1995) Minimalist Program. This paper presents the first ever application of this formalism to the task of realistic wide-coverage parsing. The parser uses a linguistically expressive yet highly constrained grammar, together with an adaptation of the A* search algorithm currently used in CCG parsing (Lewis and Steedman, 2014; Lewis et al., 2016), with supertag probabilities provided by a bi-LSTM neural network supertagger trained on MGbank, a corpus of MG derivation trees. We report on some promising initial experimental results for overall dependency recovery as well as on the recovery of certain unbounded long distance dependencies. Finally, although like other MG parsers, ours has a high order polynomial worst case time complexity, we show that in practice its expected time complexity is cubic in the length of the sentence. The parser is publicly available.

Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model

- https://www.aclweb.org/anthology/P19-1239/

Sarcasm is a subtle form of language in which people express the opposite of what is implied. Previous works of sarcasm detection focused on texts. However, more and more social media platforms like Twitter allow users to create multi-modal messages, including texts, images, and videos. It is insufficient to detect sarcasm from multi-model messages based only on texts. In this paper, we focus on multi-modal sarcasm detection for tweets consisting of texts and images in Twitter. We treat text features, image features and image attributes as three modalities and propose a multi-modal hierarchical fusion model to address this task. Our model first extracts image features and attribute features, and then leverages attribute features and bidirectional LSTM network to extract text features. Features of three modalities are then reconstructed and fused into one feature vector for prediction. We create a multi-modal sarcasm detection dataset based on Twitter. Evaluation results on the dataset demonstrate the efficacy of our proposed model and the usefulness of the three modalities.

Topic-Aware Neural Keyphrase Generation for Social Media Language

- https://www.aclweb.org/anthology/P19-1240/

A huge volume of user-generated content is daily produced on social media. To facilitate automatic language understanding, we study keyphrase prediction, distilling salient information from massive posts. While most existing methods extract words from source posts to form keyphrases, we propose a sequence-to-sequence (seq2seq) based neural keyphrase generation framework, enabling absent keyphrases to be created. Moreover, our model, being topic-aware, allows

joint modeling of corpus-level latent topic representations, which helps alleviate data sparsity widely exhibited in social media language. Experiments on three datasets collected from English and Chinese social media platforms show that our model significantly outperforms both extraction and generation models without exploiting latent topics. Further discussions show that our model learns meaningful topics, which interprets its superiority in social media keyphrase generation.

#YouToo? Detection of Personal Recollections of Sexual Harassment on Social Media

- https://www.aclweb.org/anthology/P19-1241/

The availability of large-scale online social data, coupled with computational methods can help us answer fundamental questions relat- ing to our social lives, particularly our health and well-being. The #MeToo trend has led to people talking about personal experiences of harassment more openly. This work attempts to aggregate such experiences of sex- ual abuse to facilitate a better understanding of social media constructs and to bring about social change. It has been found that disclo- sure of abuse has positive psychological im- pacts. Hence, we contend that such informa- tion can leveraged to create better campaigns for social change by analyzing how users react to these stories and to obtain a better insight into the consequences of sexual abuse. We use a three part Twitter-Specific Social Media Lan- guage Model to segregate personal recollections of sexual harassment from Twitter posts. An extensive comparison with state-of-the-art generic and specific models along with a de- tailed error analysis explores the merit of our proposed model.

Multi-task Pairwise Neural Ranking for Hashtag Segmentation

- https://www.aclweb.org/anthology/P19-1242/

Hashtags are often employed on social media and beyond to add metadata to a textual utterance with the goal of increasing discoverability, aiding search, or providing additional semantics. However, the semantic content of hashtags is not straightforward to infer as these represent ad-hoc conventions which frequently include multiple words joined together and can include abbreviations and unorthodox spellings. We build a dataset of 12,594 hashtags split into individual segments and propose a set of approaches for hashtag segmentation by framing it as a pairwise ranking problem between candidate segmentations. Our novel neural approaches demonstrate 24.6% error reduction in hashtag segmentation accuracy compared to the current state-of-the-art method. Finally, we demonstrate that a deeper understanding of hashtag semantics obtained

through segmentation is useful for downstream applications such as sentiment analysis, for which we achieved a 2.6% increase in average recall on the SemEval 2017 sentiment analysis dataset.

Entity-Centric Contextual Affective Analysis

- https://www.aclweb.org/anthology/P19-1243/

While contextualized word representations have improved state-of-the-art benchmarks in many NLP tasks, their potential usefulness for social-oriented tasks remains largely unexplored. We show how contextualized word embeddings can be used to capture affect dimensions in portrayals of people. We evaluate our methodology quantitatively, on held-out affect lexicons, and qualitatively, through case examples. We find that contextualized word representations do encode meaningful affect information, but they are heavily biased towards their training data, which limits their usefulness to in-domain analyses. We ultimately use our method to examine differences in portrayals of men and women.

Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks

- https://www.aclweb.org/anthology/P19-1244/

Claim verification is generally a task of verifying the veracity of a given claim, which is critical to many downstream applications. It is cumbersome and inefficient for human fact-checkers to find consistent pieces of evidence, from which solid verdict could be inferred against the claim. In this paper, we propose a novel end-to-end hierarchical attention network focusing on learning to represent coherent evidence as well as their semantic relatedness with the claim. Our model consists of three main components: 1) A coherence-based attention layer embeds coherent evidence considering the claim and sentences from relevant articles; 2) An entailment-based attention layer attends on sentences that can semantically infer the claim on top of the first attention; and 3) An output layer predicts the verdict based on the embedded evidence. Experimental results on three public benchmark datasets show that our proposed model outperforms a set of state-of-the-art baselines.

Predicting Human Activities from User-Generated Content

- https://www.aclweb.org/anthology/P19-1245/

The activities we do are linked to our interests, personality, political preferences, and decisions we make about the future. In this paper, we explore the task of predicting human activities from user-generated content. We collect a dataset

containing instances of social media users writing about a range of everyday activities. We then use a state-of-the-art sentence embedding framework tailored to recognize the semantics of human activities and perform an automatic clustering of these activities. We train a neural network model to make predictions about which clusters contain activities that were performed by a given user based on the text of their previous posts and self-description. Additionally, we explore the degree to which incorporating inferred user traits into our model helps with this prediction task.

You Write like You Eat: Stylistic Variation as a Predictor of Social Stratification

- https://www.aclweb.org/anthology/P19-1246/

Inspired by Labov's seminal work on stylisticvariation as a function of social stratification,we develop and compare neural models thatpredict a person's presumed socio-economicstatus, obtained through distant supervision,from their writing style on social media. Thefocus of our work is on identifying the mostimportant stylistic parameters to predict socio-economic group. In particular, we show theeffectiveness of morpho-syntactic features aspredictors of style, in contrast to lexical fea-tures, which are good predictors of topic

Encoding Social Information with Graph Convolutional Networks forPolitical Perspective Detection in News Media

- https://www.aclweb.org/anthology/P19-1247/

Identifying the political perspective shaping the way news events are discussed in the media is an important and challenging task. In this paper, we highlight the importance of contextualizing social information, capturing how this information is disseminated in social networks. We use Graph Convolutional Networks, a recently proposed neural architecture for representing relational information, to capture the documents' social context. We show that social information can be used effectively as a source of distant supervision, and when direct supervision is available, even little social information can significantly improve performance.

Fine-Grained Spoiler Detection from Large-Scale Review Corpora

- https://www.aclweb.org/anthology/P19-1248/

This paper presents computational approaches for automatically detecting critical plot twists in reviews of media products. First, we created a large-scale book review dataset that includes fine-grained spoiler annotations at the sentence-level, as well as book and (anonymized) user information. Second, we carefully analyzed this dataset, and found that: spoiler language tends to be book-specific; spoiler distributions vary greatly across books and review authors; and spoiler sentences tend to jointly appear in the latter part of reviews. Third, inspired by these findings, we developed an end-to-end neural network architecture to detect spoiler sentences in review corpora. Quantitative and qualitative results demonstrate that the proposed method substantially outperforms existing baselines.

Celebrity Profiling

- https://www.aclweb.org/anthology/P19-1249/

Celebrities are among the most prolific users of social media, promoting their personas and rallying followers. This activity is closely tied to genuine writing samples, which makes them worthy research subjects in many respects, not least profiling. With this paper we introduce the Webis Celebrity Corpus 2019. For its construction the Twitter feeds of 71,706 verified accounts have been carefully linked with their respective Wikidata items, crawling both. After cleansing, the resulting profiles contain an average of 29,968 words per profile and up to 239 pieces of personal information. A cross-evaluation that checked the correct association of Twitter account and Wikidata item revealed an error rate of only 0.6%, rendering the profiles highly reliable. Our corpus comprises a wide cross-section of local and global celebrities, forming a unique combination of scale, profile comprehensiveness, and label reliability. We further establish the state of the art's profiling performance by evaluating the winning approaches submitted to the PAN gender prediction tasks in a transfer learning experiment. They are only outperformed by our own deep learning approach, which we also use to exemplify celebrity occupation prediction for the first time.

Dataset Creation for Ranking Constructive News Comments

- https://www.aclweb.org/anthology/P19-1250/

Ranking comments on an online news service is a practically important task for the service provider, and thus there have been many studies on this task. However, most of them considered users' positive feedback, such as "Like"-button clicks, as a quality measure. In this paper, we address directly evaluating the quality of comments on the basis of "constructiveness," separately from user feedback. To this end, we create a new dataset including 100K+ Japanese comments with constructiveness scores (C-scores). Our experiments clarify that

C-scores are not always related to users' positive feedback, and the performance of pairwise ranking models tends to be enhanced by the variation of comments rather than articles.

Enhancing Air Quality Prediction with Social Media and Natural Language Processing

- https://www.aclweb.org/anthology/P19-1251/

Accompanied by modern industrial developments, air pollution has already become a major concern for human health. Hence, air quality measures, such as the concentration of PM2.5, have attracted increasing attention. Even some studies apply historical measurements into air quality forecast, the changes of air quality conditions are still hard to monitor. In this paper, we propose to exploit social media and natural language processing techniques to enhance air quality prediction. Social media users are treated as social sensors with their findings and locations. After filtering noisy tweets using word selection and topic modeling, a deep learning model based on convolutional neural networks and over-tweet-pooling is proposed to enhance air quality prediction. We conduct experiments on 7-month real-world Twitter datasets in the five most heavily polluted states in the USA. The results show that our approach significantly improves air quality prediction over the baseline that does not use social media by 6.9% to 17.7% in macro-F1 scores.

Twitter Homophily: Network Based Prediction of User's Occupation

- https://www.aclweb.org/anthology/P19-1252/

In this paper, we investigate the importance of social network information compared to content information in the prediction of a Twitter user's occupational class. We show that the content information of a user's tweets, the profile descriptions of a user's follower/following community, and the user's social network provide useful information for classifying a user's occupational group. In our study, we extend an existing data set for this problem, and we achieve significantly better performance by using social network homophily that has not been fully exploited in previous work. In our analysis, we found that by using the graph convolutional network to exploit social homophily, we can achieve competitive performance on this data set with just a small fraction of the training data.

Domain Adaptive Dialog Generation via Meta Learning

- https://www.aclweb.org/anthology/P19-1253/

Domain adaptation is an essential task in dialog system building because there are so many new dialog tasks created for different needs every day. Collecting and annotating training data for these new tasks is costly since it involves real user interactions. We propose a domain adaptive dialog generation method based on meta-learning (DAML). DAML is an end-to-end trainable dialog system model that learns from multiple rich-resource tasks and then adapts to new domains with minimal training samples. We train a dialog system model using multiple rich-resource single-domain dialog data by applying the model-agnostic meta-learning algorithm to dialog domain. The model is capable of learning a competitive dialog system on a new domain with only a few training examples in an efficient manner. The two-step gradient updates in DAML enable the model to learn general features across multiple tasks. We evaluate our method on a simulated dialog dataset and achieve state-of-the-art performance, which is generalizable to new tasks.

Strategies for Structuring Story Generation

- https://www.aclweb.org/anthology/P19-1254/

Writers often rely on plans or sketches to write long stories, but most current language models generate word by word from left to right. We explore coarse-to-fine models for creating narrative texts of several hundred words, and introduce new models which decompose stories by ing over actions and entities. The model first generates the predicate-argument structure of the text, where different mentions of the same entity are marked with placeholder tokens. It then generates a surface realization of the predicate-argument structure, and finally replaces the entity placeholders with context-sensitive names and references. Human judges prefer the stories from our models to a wide range of previous approaches to hierarchical text generation. Extensive analysis shows that our methods can help improve the diversity and coherence of events and entities in generated stories.

Argument Generation with Retrieval, Planning, and Realization

- https://www.aclweb.org/anthology/P19-1255/

Automatic argument generation is an appealing but challenging task. In this paper, we study the specific problem of counter-argument generation, and present a novel framework, CANDELA. It consists of a powerful retrieval system and a novel two-step generation model, where a text planning decoder first decides on the main talking points and a proper language style for each sentence, then a content realization decoder reflects the decisions and constructs an informative paragraph-level argument. Furthermore, our generation model is empowered by a retrieval system indexed with 12 million articles collected from Wikipedia and popular English news media, which provides access to high-quality content

with diversity. Automatic evaluation on a large-scale dataset collected from Reddit shows that our model yields significantly higher BLEU, ROUGE, and METEOR scores than the state-of-the-art and non-trivial comparisons. Human evaluation further indicates that our system arguments are more appropriate for refutation and richer in content.

A Simple Recipe towards Reducing Hallucination in Neural Surface Realisation

- https://www.aclweb.org/anthology/P19-1256/

Recent neural language generation systems often hallucinate contents (i.e., producing irrelevant or contradicted facts), especially when trained on loosely corresponding pairs of the input structure and text. To mitigate this issue, we propose to integrate a language understanding module for data refinement with self-training iterations to effectively induce strong equivalence between the input data and the paired text. Experiments on the E2E challenge dataset show that our proposed framework can reduce more than 50% relative unaligned noise from the original data-text pairs. A vanilla sequence-to-sequence neural NLG model trained on the refined data has improved on content correctness compared with the current state-of-the-art ensemble generator.

Cross-Modal Commentator: Automatic Machine Commenting Based on Cross-Modal Information

- https://www.aclweb.org/anthology/P19-1257/

Automatic commenting of online articles can provide additional opinions and facts to the reader, which improves user experience and engagement on social media platforms. Previous work focuses on automatic commenting based solely on textual content. However, in real-scenarios, online articles usually contain multiple modal contents. For instance, graphic news contains plenty of images in addition to text. Contents other than text are also vital because they are not only more attractive to the reader but also may provide critical information. To remedy this, we propose a new task: cross-model automatic commenting (CMAC), which aims to make comments by integrating multiple modal contents. We construct a large-scale dataset for this task and explore several representative methods. Going a step further, an effective co-attention model is presented to capture the dependency between textual and visual information. Evaluation results show that our proposed model can achieve better performance than competitive baselines.

A Working Memory Model for Task-oriented Dialog Response Generation

- https://www.aclweb.org/anthology/P19-1258/

Recently, to incorporate external Knowledge Base (KB) information, one form of world knowledge, several end-to-end task-oriented dialog systems have been proposed. These models, however, tend to confound the dialog history with KB tuples and simply store them into one memory. Inspired by the psychological studies on working memory, we propose a working memory model (WMM2Seq) for dialog response generation. Our WMM2Seq adopts a working memory to interact with two separated long-term memories, which are the episodic memory for memorizing dialog history and the semantic memory for storing KB tuples. The working memory consists of a central executive to attend to the aforementioned memories, and a short-term storage system to store the "activated" contents from the long-term memories. Furthermore, we introduce a context-sensitive perceptual process for the token representations of dialog history, and then feed them into the episodic memory. Extensive experiments on two task-oriented dialog datasets demonstrate that our WMM2Seq significantly outperforms the state-of-the-art results in several evaluation metrics.

Cognitive Graph for Multi-Hop Reading Comprehension at Scale

- https://www.aclweb.org/anthology/P19-1259/

We propose a new CogQA framework for multi-hop reading comprehension question answering in web-scale documents. Founded on the dual process theory in cognitive science, the framework gradually builds a cognitive graph in an iterative process by coordinating an implicit extraction module (System 1) and an explicit reasoning module (System 2). While giving accurate answers, our framework further provides explainable reasoning paths. Specifically, our implementation based on BERT and graph neural network efficiently handles millions of documents for multi-hop reasoning questions in the HotpotQA fullwiki dataset, achieving a winning joint F1 score of 34.9 on the leaderboard, compared to 23.1 of the best competitor.

Multi-hop Reading Comprehension across Multiple Documents by Reasoning over Heterogeneous Graphs

- https://www.aclweb.org/anthology/P19-1260/

Multi-hop reading comprehension (RC) across documents poses new challenge over single-document RC because it requires reasoning over multiple documents to reach the final answer. In this paper, we propose a new model to tackle the

multi-hop RC problem. We introduce a heterogeneous graph with different types of nodes and edges, which is named as Heterogeneous Document-Entity (HDE) graph. The advantage of HDE graph is that it contains different granularity levels of information including candidates, documents and entities in specific document contexts. Our proposed model can do reasoning over the HDE graph with nodes representation initialized with co-attention and self-attention based context encoders. We employ Graph Neural Networks (GNN) based message passing algorithms to accumulate evidences on the proposed HDE graph. Evaluated on the blind test set of the Qangaroo WikiHop data set, our HDE graph based single model delivers competitive result, and the ensemble model achieves the state-of-the-art performance.

Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension

- https://www.aclweb.org/anthology/P19-1261/

Multi-hop reading comprehension requires the model to explore and connect relevant information from multiple sentences/documents in order to answer the question about the context. To achieve this, we propose an interpretable 3-module system called Explore-Propose-Assemble reader (EPAr). First, the Document Explorer iteratively selects relevant documents and represents divergent reasoning chains in a tree structure so as to allow assimilating information from all chains. The Answer Proposer then proposes an answer from every root-to-leaf path in the reasoning tree. Finally, the Evidence Assembler extracts a key sentence containing the proposed answer from every path and combines them to predict the final answer. Intuitively, EPAr approximates the coarse-to-fine-grained comprehension behavior of human readers when facing multiple long documents. We jointly optimize our 3 modules by minimizing the sum of losses from each stage conditioned on the previous stage's output. On two multi-hop reading comprehension datasets WikiHop and MedHop, our EPAr model achieves significant improvements over the baseline and competitive results compared to the state-of-the-art model. We also present multiple reasoning-chain-recovery tests and ablation studies to demonstrate our system's ability to perform interpretable and accurate reasoning.

Avoiding Reasoning Shortcuts: Adversarial Evaluation, Training, and Model Development for Multi-Hop QA

- https://www.aclweb.org/anthology/P19-1262/

Multi-hop question answering requires a model to connect multiple pieces of evidence scattered in a long context to answer the question. In this paper, we show that in the multi-hop HotpotQA (Yang et al., 2018) dataset, the examples often contain reasoning shortcuts through which models can directly locate

the answer by word-matching the question with a sentence in the context. We demonstrate this issue by constructing adversarial documents that create contradicting answers to the shortcut but do not affect the validity of the original answer. The performance of strong baseline models drops significantly on our adversarial test, indicating that they are indeed exploiting the shortcuts rather than performing multi-hop reasoning. After adversarial training, the baseline's performance improves but is still limited on the adversarial test. Hence, we use a control unit that dynamically attends to the question at different reasoning hops to guide the model's multi-hop reasoning. We show that our 2-hop model trained on the regular data is more robust to the adversaries than the baseline. After adversarial training, it not only achieves significant improvements over its counterpart trained on regular data, but also outperforms the adversarially-trained baseline significantly. Finally, we sanity-check that these improvements are not obtained by exploiting potential new shortcuts in the adversarial data, but indeed due to robust multi-hop reasoning skills of the models.

Exploiting Explicit Paths for Multi-hop Reading Comprehension

- https://www.aclweb.org/anthology/P19-1263/

We propose a novel, path-based reasoning approach for the multi-hop reading comprehension task where a system needs to combine facts from multiple passages to answer a question. Although inspired by multi-hop reasoning over knowledge graphs, our proposed approach operates directly over unstructured text. It generates potential paths through passages and scores them without any direct path supervision. The proposed model, named PathNet, attempts to extract implicit relations from text through entity pair representations, and compose them to encode each path. To capture additional context, PathNet also composes the passage representations along each path to compute a passage-based representation. Unlike previous approaches, our model is then able to explain its reasoning via these explicit paths through the passages. We show that our approach outperforms prior models on the multi-hop Wikihop dataset, and also can be generalized to apply to the OpenBookQA dataset, matching state-of-the-art performance.

Sentence Mover's Similarity: Automatic Evaluation for Multi-Sentence Texts

- https://www.aclweb.org/anthology/P19-1264/

For evaluating machine-generated texts, automatic methods hold the promise of avoiding collection of human judgments, which can be expensive and time-consuming. The most common automatic metrics, like BLEU and ROUGE,

depend on exact word matching, an inflexible approach for measuring semantic similarity. We introduce methods based on sentence mover's similarity; our automatic metrics evaluate text in a continuous space using word and sentence embeddings. We find that sentence-based metrics correlate with human judgments significantly better than ROUGE, both on machine-generated summaries (average length of 3.4 sentences) and human-authored essays (average length of 7.5). We also show that sentence mover's similarity can be used as a reward when learning a generation model via reinforcement learning; we present both automatic and human evaluations of summaries learned in this way, finding that our approach outperforms ROUGE.

Analysis of Automatic Annotation Suggestions for Hard Discourse-Level Tasks in Expert Domains

- https://www.aclweb.org/anthology/P19-1265/

Many complex discourse-level tasks can aid domain experts in their work but require costly expert annotations for data creation. To speed up and ease annotations, we investigate the viability of automatically generated annotation suggestions for such tasks. As an example, we choose a task that is particularly hard for both humans and machines: the segmentation and classification of epistemic activities in diagnostic reasoning texts. We create and publish a new dataset covering two domains and carefully analyse the suggested annotations. We find that suggestions have positive effects on annotation speed and performance, while not introducing noteworthy biases. Envisioning suggestion models that improve with newly annotated texts, we contrast methods for continuous model adjustment and suggest the most effective setup for suggestions in future expert tasks.

Deep Dominance - How to Properly Compare Deep Neural Models

- https://www.aclweb.org/anthology/P19-1266/

Comparing between Deep Neural Network (DNN) models based on their performance on unseen data is crucial for the progress of the NLP field. However, these models have a large number of hyper-parameters and, being non-convex, their convergence point depends on the random values chosen at initialization and during training. Proper DNN comparison hence requires a comparison between their empirical score distributions on unseen data, rather than between single evaluation scores as is standard for more simple, convex models. In this paper, we propose to adapt to this problem a recently proposed test for the Almost Stochastic Dominance relation between two distributions. We define the criteria for a high quality comparison method between DNNs, and show,

both theoretically and through analysis of extensive experimental results with leading DNN models for sequence tagging tasks, that the proposed test meets all criteria while previously proposed methods fail to do so. We hope the test we propose here will set a new working practice in the NLP community.

We Need to Talk about Standard Splits

- https://www.aclweb.org/anthology/P19-1267/

It is standard practice in speech & language technology to rank systems according to their performance on a test set held out for evaluation. However, few researchers apply statistical tests to determine whether differences in performance are likely to arise by chance, and few examine the stability of system ranking across multiple training-testing splits. We conduct replication and reproduction experiments with nine part-of-speech taggers published between 2000 and 2018, each of which claimed state-of-the-art performance on a widely-used "standard split". While we replicate results on the standard split, we fail to reliably reproduce some rankings when we repeat this analysis with randomly generated training-testing splits. We argue that randomly generated splits should be used in system evaluation.

Aiming beyond the Obvious: Identifying Non-Obvious Cases in Semantic Similarity Datasets

- https://www.aclweb.org/anthology/P19-1268/

Existing datasets for scoring text pairs in terms of semantic similarity contain instances whose resolution differs according to the degree of difficulty. This paper proposes to distinguish obvious from non-obvious text pairs based on superficial lexical overlap and ground-truth labels. We characterise existing datasets in terms of containing difficult cases and find that recently proposed models struggle to capture the non-obvious cases of semantic similarity. We describe metrics that emphasise cases of similarity which require more complex inference and propose that these are used for evaluating systems for semantic similarity.

Putting Evaluation in Context: Contextual Embeddings Improve Machine Translation Evaluation

- https://www.aclweb.org/anthology/P19-1269/

Accurate, automatic evaluation of machine translation is critical for system tuning, and evaluating progress in the field. We proposed a simple unsupervised

metric, and additional supervised metrics which rely on contextual word embeddings to encode the translation and reference sentences. We find that these models rival or surpass all existing metrics in the WMT 2017 sentence-level and system-level tracks, and our trained model has a substantially higher correlation with human judgements than all existing metrics on the WMT 2017 to-English sentence level dataset.

Joint Effects of Context and User History for Predicting Online Conversation Re-entries

- https://www.aclweb.org/anthology/P19-1270/

As the online world continues its exponential growth, interpersonal communication has come to play an increasingly central role in opinion formation and change. In order to help users better engage with each other online, we study a challenging problem of re-entry prediction foreseeing whether a user will come back to a conversation they once participated in. We hypothesize that both the context of the ongoing conversations and the users' previous chatting history will affect their continued interests in future engagement. Specifically, we propose a neural framework with three main layers, each modeling context, user history, and interactions between them, to explore how the conversation context and user chatting history jointly result in their re-entry behavior. We experiment with two large-scale datasets collected from Twitter and Reddit. Results show that our proposed framework with bi-attention achieves an F1 score of 61.1 on Twitter conversations, outperforming the state-of-the-art methods from previous work.

CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech

- https://www.aclweb.org/anthology/P19-1271/

Although there is an unprecedented effort to provide adequate responses in terms of laws and policies to hate content on social media platforms, dealing with hatred online is still a tough problem. Tackling hate speech in the standard way of content deletion or user suspension may be charged with censorship and overblocking. One alternate strategy, that has received little attention so far by the research community, is to actually oppose hate content with counter-narratives (i.e. informed textual responses). In this paper, we describe the creation of the first large-scale, multilingual, expert-based dataset of hate-speech/counter-narrative pairs. This dataset has been built with the effort of more than 100 operators from three different NGOs that applied their training

and expertise to the task. Together with the collected data we also provide additional annotations about expert demographics, hate and response type, and data augmentation through translation and paraphrasing. Finally, we provide initial experiments to assess the quality of our data.

Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts

- https://www.aclweb.org/anthology/P19-1272/

Text in social media posts is frequently accompanied by images in order to provide content, supply context, or to express feelings. This paper studies how the meaning of the entire tweet is composed through the relationship between its textual content and its image. We build and release a data set of image tweets annotated with four classes which express whether the text or the image provides additional information to the other modality. We show that by combining the text and image information, we can build a machine learning approach that accurately distinguishes between the relationship types. Further, we derive insights into how these relationships are materialized through text and image content analysis and how they are impacted by user demographic traits. These methods can be used in several downstream applications including pre-training image tagging models, collecting distantly supervised data for image captioning, and can be directly used in end-user applications to optimize screen estate.

Who Sides with Whom? Towards Computational Construction of Discourse Networks for Political Debates

- https://www.aclweb.org/anthology/P19-1273/

Understanding the structures of political debates (which actors make what claims) is essential for understanding democratic political decision making. The vision of computational construction of such discourse networks from newspaper reports brings together political science and natural language processing. This paper presents three contributions towards this goal: (a) a requirements analysis, linking the task to knowledge base population; (b) an annotated pilot corpus of migration claims based on German newspaper reports; (c) initial modeling results.

Analyzing Linguistic Differences between Owner and Staff Attributed Tweets

- https://www.aclweb.org/anthology/P19-1274/

Research on social media has to date assumed that all posts from an account are authored by the same person. In this study, we challenge this assumption and study the linguistic differences between posts signed by the account owner or attributed to their staff. We introduce a novel data set of tweets posted by U.S. politicians who self-reported their tweets using a signature. We analyze the linguistic topics and style features that distinguish the two types of tweets. Predictive results show that we are able to predict owner and staff attributed tweets with good accuracy, even when not using any training data from that account.

Exploring Author Context for Detecting Intended vs Perceived Sarcasm

- https://www.aclweb.org/anthology/P19-1275/

We investigate the impact of using author context on textual sarcasm detection. We define author context as the embedded representation of their historical posts on Twitter and suggest neural models that extract these representations. We experiment with two tweet datasets, one labelled manually for sarcasm, and the other via tag-based distant supervision. We achieve state-of-the-art performance on the second dataset, but not on the one labelled manually, indicating a difference between intended sarcasm, captured by distant supervision, and perceived sarcasm, captured by manual labelling.

Open Domain Event Extraction Using Neural Latent Variable Models

- https://www.aclweb.org/anthology/P19-1276/

We consider open domain event extraction, the task of extracting unconstraint types of events from news clusters. A novel latent variable neural model is constructed, which is scalable to very large corpus. A dataset is collected and manually annotated, with task-specific evaluation metrics being designed. Results show that the proposed unsupervised model gives better performance compared to the state-of-the-art method for event schema induction.

Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification

- https://www.aclweb.org/anthology/P19-1277/

This paper presents a multi-level matching and aggregation network (MLMAN) for few-shot relation classification. Previous studies on this topic adopt prototypical networks, which calculate the embedding vector of a query instance

and the prototype vector of the support set for each relation candidate independently. On the contrary, our proposed MLMAN model encodes the query instance and each support set in an interactive way by considering their matching information at both local and instance levels. The final class prototype for each support set is obtained by attentive aggregation over the representations of support instances, where the weights are calculated using the query instance. Experimental results demonstrate the effectiveness of our proposed methods, which achieve a new state-of-the-art performance on the FewRel dataset.

Quantifying Similarity between Relations with Fact Distribution

- https://www.aclweb.org/anthology/P19-1278/

We introduce a conceptually simple and effective method to quantify the similarity between relations in knowledge bases. Specifically, our approach is based on the divergence between the conditional probability distributions over entity pairs. In this paper, these distributions are parameterized by a very simple neural network. Although computing the exact similarity is in-tractable, we provide a sampling-based method to get a good approximation. We empirically show the outputs of our approach significantly correlate with human judgments. By applying our method to various tasks, we also find that (1) our approach could effectively detect redundant relations extracted by open information extraction (Open IE) models, that (2) even the most competitive models for relational classification still make mistakes among very similar relations, and that (3) our approach could be incorporated into negative sampling and softmax classification to alleviate these mistakes.

Matching the Blanks: Distributional Similarity for Relation Learning

- https://www.aclweb.org/anthology/P19-1279/

General purpose relation extractors, which can model arbitrary relations, are a core aspiration in information extraction. Efforts have been made to build general purpose extractors that represent relations with their surface forms, or which jointly embed surface forms with relations from an existing knowledge graph. However, both of these approaches are limited in their ability to generalize. In this paper, we build on extensions of Harris' distributional hypothesis to relations, as well as recent advances in learning text representations (specifically, BERT), to build task agnostic relation representations solely from entity-linked text. We show that these representations significantly outperform previous work on exemplar based relation extraction (FewRel) even without using any of that task's training data. We also show that models initialized with our task agnostic representations, and then tuned on supervised relation extraction datasets,

significantly outperform the previous methods on SemEval 2010 Task 8, KBP37, and TACRED

Fine-Grained Temporal Relation Extraction

- https://www.aclweb.org/anthology/P19-1280/

We present a novel semantic framework for modeling temporal relations and event durations that maps pairs of events to real-valued scales. We use this framework to construct the largest temporal relations dataset to date, covering the entirety of the Universal Dependencies English Web Treebank. We use this dataset to train models for jointly predicting fine-grained temporal relations and event durations. We report strong results on our data and show the efficacy of a transfer-learning approach for predicting categorical relations.

FIESTA: Fast IdEntification of State-of-The-Art models using adaptive bandit algorithms

- https://www.aclweb.org/anthology/P19-1281/

We present FIESTA, a model selection approach that significantly reduces the computational resources required to reliably identify state-of-the-art performance from large collections of candidate models. Despite being known to produce unreliable comparisons, it is still common practice to compare model evaluations based on single choices of random seeds. We show that reliable model selection also requires evaluations based on multiple train-test splits (contrary to common practice in many shared tasks). Using bandit theory from the statistics literature, we are able to adaptively determine appropriate numbers of data splits and random seeds used to evaluate each model, focusing computational resources on the evaluation of promising models whilst avoiding wasting evaluations on models with lower performance. Furthermore, our user-friendly Python implementation produces confidence guarantees of correctly selecting the optimal model. We evaluate our algorithms by selecting between 8 target-dependent sentiment analysis methods using dramatically fewer model evaluations than current model selection approaches.

Is Attention Interpretable?

- https://www.aclweb.org/anthology/P19-1282/

Attention mechanisms have recently boosted performance on a range of NLP tasks. Because attention layers explicitly weight input components' representations, it is also often assumed that attention can be used to identify information that models found important (e.g., specific contextualized word tokens). We test

whether that assumption holds by manipulating attention weights in already-trained text classification models and analyzing the resulting differences in their predictions. While we observe some ways in which higher attention weights correlate with greater impact on model predictions, we also find many ways in which this does not hold, i.e., where gradient-based rankings of attention weights better predict their effects than their magnitudes. We conclude that while attention noisily predicts input components' overall importance to a model, it is by no means a fail-safe indicator.

Correlating Neural and Symbolic Representations of Language

- https://www.aclweb.org/anthology/P19-1283/

Analysis methods which enable us to better understand the representations and functioning of neural models of language are increasingly needed as deep learning becomes the dominant approach in NLP. Here we present two methods based on Representational Similarity Analysis (RSA) and Tree Kernels (TK) which allow us to directly quantify how strongly the information encoded in neural activation patterns corresponds to information represented by symbolic structures such as syntax trees. We first validate our methods on the case of a simple synthetic language for arithmetic expressions with clearly defined syntax and semantics, and show that they exhibit the expected pattern of results. We then our methods to correlate neural representations of English sentences with their constituency parse trees.

Interpretable Neural Predictions with Differentiable Binary Variables

- https://www.aclweb.org/anthology/P19-1284/

The success of neural networks comes hand in hand with a desire for more interpretability. We focus on text classifiers and make them more interpretable by having them provide a justification–a rationale–for their predictions. We approach this problem by jointly training two neural network models: a latent model that selects a rationale (i.e. a short and informative part of the input text), and a classifier that learns from the words in the rationale alone. Previous work proposed to assign binary latent masks to input positions and to promote short selections via sparsity-inducing penalties such as L0 regularisation. We propose a latent model that mixes discrete and continuous behaviour allowing at the same time for binary selections and gradient-based training without REINFORCE. In our formulation, we can tractably compute the expected value of penalties such as L0, which allows us to directly optimise the model

towards a pre-specified text selection rate. We show that our approach is competitive with previous work on rationale extraction, and explore further uses in attention mechanisms.

Transformer-XL: Attentive Language Models beyond a Fixed-Length Context

- https://www.aclweb.org/anthology/P19-1285/

Transformers have a potential of learning longer-term dependency, but are limited by a fixed-length context in the setting of language modeling. We propose a novel neural architecture Transformer-XL that enables learning dependency beyond a fixed length without disrupting temporal coherence. It consists of a segment-level recurrence mechanism and a novel positional encoding scheme. Our method not only enables capturing longer-term dependency, but also resolves the context fragmentation problem. As a result, Transformer-XL learns dependency that is 80% longer than RNNs and 450% longer than vanilla Transformers, achieves better performance on both short and long sequences, and is up to 1,800+ times faster than vanilla Transformers during evaluation. Notably, we improve the state-of-the-art results of bpc/perplexity to 0.99 on enwiki8, 1.08 on text8, 18.3 on WikiText-103, 21.8 on One Billion Word, and 54.5 on Penn Treebank (without finetuning). When trained only on WikiText-103, Transformer-XL manages to generate reasonably coherent, novel text articles with thousands of tokens. Our code, pretrained models, and hyperparameters are available in both Tensorflow and PyTorch.

Domain Adaptation of Neural Machine Translation by Lexicon Induction

- https://www.aclweb.org/anthology/P19-1286/

It has been previously noted that neural machine translation (NMT) is very sensitive to domain shift. In this paper, we argue that this is a dual effect of the highly lexicalized nature of NMT, resulting in failure for sentences with large numbers of unknown words, and lack of supervision for domain-specific words. To remedy this problem, we propose an unsupervised adaptation method which fine-tunes a pre-trained out-of-domain NMT model using a pseudo-in-domain corpus. Specifically, we perform lexicon induction to extract an in-domain lexicon, and construct a pseudo-parallel in-domain corpus by performing word-for-word back-translation of monolingual in-domain target sentences. In five domains over twenty pairwise adaptation settings and two model architectures, our method achieves consistent improvements without using any in-domain parallel sentences, improving up to 14 BLEU over unadapted models, and up to 2 BLEU over strong back-translation baselines.

Reference Network for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1287/

Neural Machine Translation (NMT) has achieved notable success in recent years. Such a framework usually generates translations in isolation. In contrast, human translators often refer to reference data, either rephrasing the intricate sentence fragments with common terms in source language, or just accessing to the golden translation directly. In this paper, we propose a Reference Network to incorporate referring process into translation decoding of NMT. To construct a reference book, an intuitive way is to store the detailed translation history with extra memory, which is computationally expensive. Instead, we employ Local Coordinates Coding (LCC) to obtain global context vectors containing monolingual and bilingual contextual information for NMT decoding. Experimental results on Chinese-English and English-German tasks demonstrate that our proposed model is effective in improving the translation quality with lightweight computation cost.

Retrieving Sequential Information for Non-Autoregressive Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1288/

Non-Autoregressive Transformer (NAT) aims to accelerate the Transformer model through discarding the autoregressive mechanism and generating target words independently, which fails to exploit the target sequential information. Over-translation and under-translation errors often occur for the above reason, especially in the long sentence translation scenario. In this paper, we propose two approaches to retrieve the target sequential information for NAT to enhance its translation ability while preserving the fast-decoding property. Firstly, we propose a sequence-level training method based on a novel reinforcement algorithm for NAT (Reinforce-NAT) to reduce the variance and stabilize the training procedure. Secondly, we propose an innovative Transformer decoder named FS-decoder to fuse the target sequential information into the top layer of the decoder. Experimental results on three translation tasks show that the Reinforce-NAT surpasses the baseline NAT system by a significant margin on BLEU without decelerating the decoding speed and the FS-decoder achieves comparable translation performance to the autoregressive Transformer with considerable speedup.

STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework

- https://www.aclweb.org/anthology/P19-1289/

Simultaneous translation, which translates sentences before they are finished, is use- ful in many scenarios but is notoriously dif- ficult due to word-order differences. While the conventional seq-to-seq framework is only suitable for full-sentence translation, we pro- pose a novel prefix-to-prefix framework for simultaneous translation that implicitly learns to anticipate in a single translation model. Within this framework, we present a very sim- ple yet surprisingly effective "wait-k" policy trained to generate the target sentence concur- rently with the source sentence, but always k words behind. Experiments show our strategy achieves low latency and reasonable qual- ity (compared to full-sentence translation) on 4 directions: zh en and de en.

Look Harder: A Neural Machine Translation Model with Hard Attention

- https://www.aclweb.org/anthology/P19-1290/

Soft-attention based Neural Machine Translation (NMT) models have achieved promising results on several translation tasks. These models attend all the words in the source sequence for each target token, which makes them ineffective for long sequence translation. In this work, we propose a hard-attention based NMT model which selects a subset of source tokens for each target token to effectively handle long sequence translation. Due to the discrete nature of the hard-attention mechanism, we design a reinforcement learning algorithm coupled with reward shaping strategy to efficiently train it. Experimental results show that the proposed model performs better on long sequences and thereby achieves significant BLEU score improvement on English-German (EN-DE) and English-French (ENFR) translation tasks compared to the soft attention based NMT.

Robust Neural Machine Translation with Joint Textual and Phonetic Embedding

- https://www.aclweb.org/anthology/P19-1291/

Neural machine translation (NMT) is notoriously sensitive to noises, but noises are almost inevitable in practice. One special kind of noise is the homophone noise, where words are replaced by other words with similar pronunciations. We propose to improve the robustness of NMT to homophone noises by 1) jointly embedding both textual and phonetic information of source sentences,

and 2) augmenting the training dataset with homophone noises. Interestingly, to achieve better translation quality and more robustness, we found that most (though not all) weights should be put on the phonetic rather than textual information. Experiments show that our method not only significantly improves the robustness of NMT to homophone noises, but also surprisingly improves the translation quality on some clean test sets.

A Simple and Effective Approach to Automatic Post-Editing with Transfer Learning

- https://www.aclweb.org/anthology/P19-1292/

Automatic post-editing (APE) seeks to automatically refine the output of a black-box machine translation (MT) system through human post-edits. APE systems are usually trained by complementing human post-edited data with large, artificial data generated through back-translations, a time-consuming process often no easier than training a MT system from scratch. in this paper, we propose an alternative where we fine-tune pre-trained BERT models on both the encoder and decoder of an APE system, exploring several parameter sharing strategies. By only training on a dataset of 23K sentences for 3 hours on a single GPU we obtain results that are competitive with systems that were trained on 5M artificial sentences. When we add this artificial data our method obtains state-of-the-art results.

Translating Translationese: A Two-Step Approach to Unsupervised Machine Translation

- https://www.aclweb.org/anthology/P19-1293/

Given a rough, word-by-word gloss of a source language sentence, target language natives can uncover the latent, fully-fluent rendering of the translation. In this work we explore this intuition by breaking translation into a two step process: generating a rough gloss by means of a dictionary and then 'translating' the resulting pseudo-translation, or 'Translationese' into a fully fluent translation. We build our Translationese decoder once from a mish-mash of parallel data that has the target language in common and then can build dictionaries on demand using unsupervised techniques, resulting in rapidly generated unsupervised neural MT systems for many source languages. We apply this process to 14 test languages, obtaining better or comparable translation results on high-resource languages than previously published unsupervised MT studies, and obtaining good quality results for low-resource languages that have never been used in an unsupervised MT scenario.

Training Neural Machine Translation to Apply Terminology Constraints

- https://www.aclweb.org/anthology/P19-1294/

This paper proposes a novel method to inject custom terminology into neural machine translation at run time. Previous works have mainly proposed modifications to the decoding algorithm in order to constrain the output to include run-time-provided target terms. While being effective, these constrained decoding methods add, however, significant computational overhead to the inference step, and, as we show in this paper, can be brittle when tested in realistic conditions. In this paper we approach the problem by training a neural MT system to learn how to use custom terminology when provided with the input. Comparative experiments show that our method is not only more effective than a state-of-the-art implementation of constrained decoding, but is also as fast as constraint-free decoding.

Leveraging Local and Global Patterns for Self-Attention Networks

- https://www.aclweb.org/anthology/P19-1295/

Self-attention networks have received increasing research attention. By default, the hidden states of each word are hierarchically calculated by attending to all words in the sentence, which assembles global information. However, several studies pointed out that taking all signals into account may lead to overlooking neighboring information (e.g. phrase pattern). To address this argument, we propose a hybrid attention mechanism to dynamically leverage both of the local and global information. Specifically, our approach uses a gating scalar for integrating both sources of the information, which is also convenient for quantifying their contributions. Experiments on various neural machine translation tasks demonstrate the effectiveness of the proposed method. The extensive analyses verify that the two types of contexts are complementary to each other, and our method gives highly effective improvements in their integration.

Sentence-Level Agreement for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1296/

The training objective of neural machine translation (NMT) is to minimize the loss between the words in the translated sentences and those in the references. In NMT, there is a natural correspondence between the source sentence and the target sentence. However, this relationship has only been represented using the entire neural network and the training objective is computed in word-level. In this paper, we propose a sentence-level agreement module to directly minimize

the difference between the representation of source and target sentence. The proposed agreement module can be integrated into NMT as an additional training objective function and can also be used to enhance the representation of the source sentences. Empirical results on the NIST Chinese-to-English and WMT English-to-German tasks show the proposed agreement module can significantly improve the NMT performance.

Multilingual Unsupervised NMT using Shared Encoder and Language-Specific Decoders

- https://www.aclweb.org/anthology/P19-1297/

In this paper, we propose a multilingual unsupervised NMT scheme which jointly trains multiple languages with a shared encoder and multiple decoders. Our approach is based on denoising autoencoding of each language and back-translating between English and multiple non-English languages. This results in a universal encoder which can encode any language participating in training into an interlingual representation, and language-specific decoders. Our experiments using only monolingual corpora show that multilingual unsupervised model performs better than the separately trained bilingual models achieving improvement of up to 1.48 BLEU points on WMT test sets. We also observe that even if we do not train the network for all possible translation directions, the network is still able to translate in a many-to-many fashion leveraging encoder's ability to generate interlingual representation.

Lattice-Based Transformer Encoder for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1298/

Neural machine translation (NMT) takes deterministic sequences for source representations. However, either word-level or subword-level segmentations have multiple choices to split a source sequence with different word segmentors or different subword vocabulary sizes. We hypothesize that the diversity in segmentations may affect the NMT performance. To integrate different segmentations with the state-of-the-art NMT model, Transformer, we propose lattice-based encoders to explore effective word or subword representation in an automatic way during training. We propose two methods: 1) lattice positional encoding and 2) lattice-aware self-attention. These two methods can be used together and show complementary to each other to further improve translation performance. Experiment results show superiorities of lattice-based encoders in word-level and subword-level representations over conventional Transformer encoder.

Multi-Source Cross-Lingual Model Transfer: Learning What to Share

- https://www.aclweb.org/anthology/P19-1299/

Modern NLP applications have enjoyed a great boost utilizing neural networks models. Such deep neural models, however, are not applicable to most human languages due to the lack of annotated training data for various NLP tasks. Cross-lingual transfer learning (CLTL) is a viable method for building NLP models for a low-resource target language by leveraging labeled data from other (source) languages. In this work, we focus on the multilingual transfer setting where training data in multiple source languages is leveraged to further boost target language performance. Unlike most existing methods that rely only on language-invariant features for CLTL, our approach coherently utilizes both language-invariant and language-specific features at instance level. Our model leverages adversarial networks to learn language-invariant features, and mixture-of-experts models to dynamically exploit the similarity between the target language and each individual source language. This enables our model to learn effectively what to share between various languages in the multilingual setup. Moreover, when coupled with unsupervised multilingual embeddings, our model can operate in a zero-resource setting where neither target language training data nor cross-lingual resources are available. Our model achieves significant performance gains over prior art, as shown in an extensive set of experiments over multiple text classification and sequence tagging tasks including a large-scale industry dataset.

Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models

- https://www.aclweb.org/anthology/P19-1300/

Recently, a variety of unsupervised methods have been proposed that map pre-trained word embeddings of different languages into the same space without any parallel data. These methods aim to find a linear transformation based on the assumption that monolingual word embeddings are approximately isomorphic between languages. However, it has been demonstrated that this assumption holds true only on specific conditions, and with limited resources, the performance of these methods decreases drastically. To overcome this problem, we propose a new unsupervised multilingual embedding method that does not rely on such assumption and performs well under resource-poor scenarios, namely when only a small amount of monolingual data (i.e., 50k sentences) are available, or when the domains of monolingual data are different across languages. Our proposed model, which we call 'Multilingual Neural Language Models', shares some of the network parameters among multiple languages, and encodes sentences of multiple languages into the same space. The model jointly learns

word embeddings of different languages in the same space, and generates multilingual embeddings without any parallel data or pre-training. Our experiments on word alignment tasks have demonstrated that, on the low-resource condition, our model substantially outperforms existing unsupervised and even supervised methods trained with 500 bilingual pairs of words. Our model also outperforms unsupervised methods given different-domain corpora across languages. Our code is publicly available.

Choosing Transfer Languages for Cross-Lingual Learning

- https://www.aclweb.org/anthology/P19-1301/

Cross-lingual transfer, where a high-resource transfer language is used to improve the accuracy of a low-resource task language, is now an invaluable tool for improving performance of natural language processing (NLP) on low-resource languages. However, given a particular task language, it is not clear which language to transfer from, and the standard strategy is to select languages based on ad hoc criteria, usually the intuition of the experimenter. Since a large number of features contribute to the success of cross-lingual transfer (including phylogenetic similarity, typological properties, lexical overlap, or size of available data), even the most enlightened experimenter rarely considers all these factors for the particular task at hand. In this paper, we consider this task of automatically selecting optimal transfer languages as a ranking problem, and build models that consider the aforementioned features to perform this prediction. In experiments on representative NLP tasks, we demonstrate that our model predicts good transfer languages much better than ad hoc baselines considering single features in isolation, and glean insights on what features are most informative for each different NLP tasks, which may inform future ad hoc selection even without use of our method.

CogNet: A Large-Scale Cognate Database

- https://www.aclweb.org/anthology/P19-1302/

This paper introduces CogNet, a new, large-scale lexical database that provides cognates -words of common origin and meaning- across languages. The database currently contains 3.1 million cognate pairs across 338 languages using 35 writing systems. The paper also describes the automated method by which cognates were computed from publicly available wordnets, with an accuracy evaluated to 94%. Finally, it presents statistics about the cognate data and some initial insights into it, hinting at a possible future exploitation of the resource by various fields of lingustics.

Neural Decipherment via Minimum-Cost Flow: From Ugaritic to Linear B

- https://www.aclweb.org/anthology/P19-1303/

In this paper we propose a novel neural approach for automatic decipherment of lost languages. To compensate for the lack of strong supervision signal, our model design is informed by patterns in language change documented in historical linguistics. The model utilizes an expressive sequence-to-sequence model to capture character-level correspondences between cognates. To effectively train the model in unsupervised manner, we innovate the training procedure by formalizing it as a minimum-cost flow problem. When applied to decipherment of Ugaritic, we achieve 5% absolute improvement over state-of-the-art results. We also report first automatic results in deciphering Linear B, a syllabic language related to ancient Greek, where our model correctly translates 67.3% of cognates.

Cross-lingual Knowledge Graph Alignment via Graph Matching Neural Network

- https://www.aclweb.org/anthology/P19-1304/

Previous cross-lingual knowledge graph (KG) alignment studies rely on entity embeddings derived only from monolingual KG structural information, which may fail at matching entities that have different facts in two KGs. In this paper, we introduce the topic entity graph, a local sub-graph of an entity, to represent entities with their contextual information in KG. From this view, the KB-alignment task can be formulated as a graph matching problem; and we further propose a graph-attention based solution, which first matches all entities in two topic entity graphs, and then jointly model the local matching information to derive a graph-level matching vector. Experiments show that our model outperforms previous state-of-the-art methods by a large margin.

Zero-Shot Cross-Lingual ive Sentence Summarization through Teaching Generation and Attention

- https://www.aclweb.org/anthology/P19-1305/

ive Sentence Summarization (ASSUM) targets at grasping the core idea of the source sentence and presenting it as the summary. It is extensively studied using statistical models or neural models based on the large-scale monolingual source-summary parallel corpus. But there is no cross-lingual parallel corpus, whose source sentence language is different to the summary language, to directly train a cross-lingual ASSUM system. We propose to solve this zero-shot problem by using resource-rich monolingual ASSUM system to teach zero-shot

cross-lingual ASSUM system on both summary word generation and attention. This teaching process is along with a back-translation process which simulates source-summary pairs. Experiments on cross-lingual ASSUM task show that our proposed method is significantly better than pipeline baselines and previous works, and greatly enhances the cross-lingual performances closer to the monolingual performances.

Improving Low-Resource Cross-lingual Document Retrieval by Reranking with Deep Bilingual Representations

- https://www.aclweb.org/anthology/P19-1306/

In this paper, we propose to boost low-resource cross-lingual document retrieval performance with deep bilingual query-document representations. We match queries and documents in both source and target languages with four components, each of which is implemented as a term interaction-based deep neural network with cross-lingual word embeddings as input. By including query likelihood scores as extra features, our model effectively learns to rerank the retrieved documents by using a small number of relevance labels for low-resource language pairs. Due to the shared cross-lingual word embedding space, the model can also be directly applied to another language pair without any training label. Experimental results on the Material dataset show that our model outperforms the competitive translation-based baselines on English-Swahili, English-Tagalog, and English-Somali cross-lingual information retrieval tasks.

Are Girls Neko or Shōjo? Cross-Lingual Alignment of Non-Isomorphic Embeddings with Iterative Normalization

- https://www.aclweb.org/anthology/P19-1307/

Cross-lingual word embeddings (CLWE) underlie many multilingual natural language processing systems, often through orthogonal transformations of pretrained monolingual embeddings. However, orthogonal mapping only works on language pairs whose embeddings are naturally isomorphic. For non-isomorphic pairs, our method (Iterative Normalization) transforms monolingual embeddings to make orthogonal alignment easier by simultaneously enforcing that (1) individual word vectors are unit length, and (2) each language's average vector is zero. Iterative Normalization consistently improves word translation accuracy of three CLWE methods, with the largest improvement observed on English-Japanese (from 2% to 44% test accuracy).

MAAM: A Morphology-Aware Alignment Model for Unsupervised Bilingual Lexicon Induction

- https://www.aclweb.org/anthology/P19-1308/

The task of unsupervised bilingual lexicon induction (UBLI) aims to induce word translations from monolingual corpora in two languages. Previous work has shown that morphological variation is an intractable challenge for the UBLI task, where the induced translation in failure case is usually morphologically related to the correct translation. To tackle this challenge, we propose a morphology-aware alignment model for the UBLI task. The proposed model aims to alleviate the adverse effect of morphological variation by introducing grammatical information learned by the pre-trained denoising language model. Results show that our approach can substantially outperform several state-of-the-art unsupervised systems, and even achieves competitive performance compared to supervised methods.

Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings

- https://www.aclweb.org/anthology/P19-1309/

Machine translation is highly sensitive to the size and quality of the training data, which has led to an increasing interest in collecting and filtering large parallel corpora. In this paper, we propose a new method for this task based on multilingual sentence embeddings. In contrast to previous approaches, which rely on nearest neighbor retrieval with a hard threshold over cosine similarity, our proposed method accounts for the scale inconsistencies of this measure, considering the margin between a given sentence pair and its closest candidates instead. Our experiments show large improvements over existing methods. We outperform the best published results on the BUCC mining task and the UN reconstruction task by more than 10 F1 and 30 precision points, respectively. Filtering the English-German ParaCrawl corpus with our approach, we obtain 31.2 BLEU points on newstest2014, an improvement of more than one point over the best official filtered version.

JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages

- https://www.aclweb.org/anthology/P19-1310/

Viable cross-lingual transfer critically depends on the availability of parallel texts. Shortage of such resources imposes a development and evaluation bottleneck in multilingual processing. We introduce JW300, a parallel corpus of over 300 languages with around 100 thousand parallel sentences per language

pair on average. In this paper, we present the resource and showcase its utility in experiments with cross-lingual word embedding induction and multi-source part-of-speech projection.

Cross-Lingual Syntactic Transfer through Unsupervised Adaptation of Invertible Projections

- https://www.aclweb.org/anthology/P19-1311/

Cross-lingual transfer is an effective way to build syntactic analysis tools in low-resource languages. However, transfer is difficult when transferring to typologically distant languages, especially when neither annotated target data nor parallel corpora are available. In this paper, we focus on methods for cross-lingual transfer to distant languages and propose to learn a generative model with a structured prior that utilizes labeled source data and unlabeled target data jointly. The parameters of source model and target model are softly shared through a regularized log likelihood objective. An invertible projection is employed to learn a new interlingual latent embedding space that compensates for imperfect cross-lingual word embedding input. We evaluate our method on two syntactic tasks: part-of-speech (POS) tagging and dependency parsing. On the Universal Dependency Treebanks, we use English as the only source corpus and transfer to a wide range of target languages. On the 10 languages in this dataset that are distant from English, our method yields an average of 5.2% absolute improvement on POS tagging and 8.3% absolute improvement on dependency parsing over a direct transfer method using state-of-the-art discriminative models.

Unsupervised Joint Training of Bilingual Word Embeddings

- https://www.aclweb.org/anthology/P19-1312/

State-of-the-art methods for unsupervised bilingual word embeddings (BWE) train a mapping function that maps pre-trained monolingual word embeddings into a bilingual space. Despite its remarkable results, unsupervised mapping is also well-known to be limited by the original dissimilarity between the word embedding spaces to be mapped. In this work, we propose a new approach that trains unsupervised BWE jointly on synthetic parallel data generated through unsupervised machine translation. We demonstrate that existing algorithms that jointly train BWE are very robust to noisy training data and show that unsupervised BWE jointly trained significantly outperform unsupervised mapped BWE in several cross-lingual NLP tasks.

Inferring Concept Hierarchies from Text Corpora via Hyperbolic Embeddings

- https://www.aclweb.org/anthology/P19-1313/

We consider the task of inferring "is-a" relationships from large text corpora. For this purpose, we propose a new method combining hyperbolic embeddings and Hearst patterns. This approach allows us to set appropriate constraints for inferring concept hierarchies from distributional contexts while also being able to predict missing "is-a"-relationships and to correct wrong extractions. Moreover – and in contrast with other methods – the hierarchical nature of hyperbolic space allows us to learn highly efficient representations and to improve the taxonomic consistency of the inferred hierarchies. Experimentally, we show that our approach achieves state-of-the-art performance on several commonly-used benchmarks.

Is Word Segmentation Necessary for Deep Learning of Chinese Representations?

- https://www.aclweb.org/anthology/P19-1314/

Segmenting a chunk of text into words is usually the first step of processing Chinese text, but its necessity has rarely been explored. In this paper, we ask the fundamental question of whether Chinese word segmentation (CWS) is necessary for deep learning-based Chinese Natural Language Processing. We benchmark neural word-based models which rely on word segmentation against neural char-based models which do not involve word segmentation in four end-to-end NLP benchmark tasks: language modeling, machine translation, sentence matching/paraphrase and text classification. Through direct comparisons between these two types of models, we find that char-based models consistently outperform word-based models. Based on these observations, we conduct comprehensive experiments to study why word-based models underperform char-based models in these deep learning-based NLP tasks. We show that it is because word-based models are more vulnerable to data sparsity and the presence of out-of-vocabulary (OOV) words, and thus more prone to overfitting. We hope this paper could encourage researchers in the community to rethink the necessity of word segmentation in deep learning-based Chinese Natural Language Processing.

Towards Understanding Linear Word Analogies

- https://www.aclweb.org/anthology/P19-1315/

A surprising property of word vectors is that word analogies can often be solved with vector arithmetic. However, it is unclear why arithmetic operators corre-

spond to non-linear embedding models such as skip-gram with negative sampling (SGNS). We provide a formal explanation of this phenomenon without making the strong assumptions that past theories have made about the vector space and word distribution. Our theory has several implications. Past work has conjectured that linear substructures exist in vector spaces because relations can be represented as ratios; we prove that this holds for SGNS. We provide novel justification for the addition of SGNS word vectors by showing that it automatically down-weights the more frequent word, as weighting schemes do ad hoc. Lastly, we offer an information theoretic interpretation of Euclidean distance in vector spaces, justifying its use in capturing word dissimilarity.

On the Compositionality Prediction of Noun Phrases using Poincaré Embeddings

- https://www.aclweb.org/anthology/P19-1316/

The compositionality degree of multiword expressions indicates to what extent the meaning of a phrase can be derived from the meaning of its constituents and their grammatical relations. Prediction of (non)-compositionality is a task that has been frequently addressed with distributional semantic models. We introduce a novel technique to blend hierarchical information with distributional information for predicting compositionality. In particular, we use hypernymy information of the multiword and its constituents encoded in the form of the recently introduced Poincaré embeddings in addition to the distributional information to detect compositionality for noun phrases. Using a weighted average of the distributional similarity and a Poincaré similarity function, we obtain consistent and substantial, statistically significant improvement across three gold standard datasets over state-of-the-art models based on distributional information only. Unlike traditional approaches that solely use an unsupervised setting, we have also framed the problem as a supervised task, obtaining comparable improvements. Further, we publicly release our Poincaré embeddings, which are trained on the output of handcrafted lexical-syntactic patterns on a large corpus.

Robust Representation Learning of Biomedical Names

- https://www.aclweb.org/anthology/P19-1317/

Biomedical concepts are often mentioned in medical documents under different name variations (synonyms). This mismatch between surface forms is problematic, resulting in difficulties pertaining to learning effective representations. Consequently, this has tremendous implications such as rendering downstream applications inefficacious and/or potentially unreliable. This paper proposes a new framework for learning robust representations of biomedical names and terms. The idea behind our approach is to consider and encode contextual

meaning, conceptual meaning, and the similarity between synonyms during the representation learning process. Via extensive experiments, we show that our proposed method outperforms other baselines on a battery of retrieval, similarity and relatedness benchmarks. Moreover, our proposed method is also able to compute meaningful representations for unseen names, resulting in high practical utility in real-world applications.

Relational Word Embeddings

- https://www.aclweb.org/anthology/P19-1318/

While word embeddings have been shown to implicitly encode various forms of attributional knowledge, the extent to which they capture relational information is far more limited. In previous work, this limitation has been addressed by incorporating relational knowledge from external knowledge bases when learning the word embedding. Such strategies may not be optimal, however, as they are limited by the coverage of available resources and conflate similarity with other forms of relatedness. As an alternative, in this paper we propose to encode relational knowledge in a separate word embedding, which is aimed to be complementary to a given standard word embedding. This relational word embedding is still learned from co-occurrence statistics, and can thus be used even when no external knowledge base is available. Our analysis shows that relational word vectors do indeed capture information that is complementary to what is encoded in standard word embeddings.

Unraveling Antonym's Word Vectors through a Siamese-like Network

- https://www.aclweb.org/anthology/P19-1319/

Discriminating antonyms and synonyms is an important NLP task that has the difficulty that both, antonyms and synonyms, contains similar distributional information. Consequently, pairs of antonyms and synonyms may have similar word vectors. We present an approach to unravel antonymy and synonymy from word vectors based on a siamese network inspired approach. The model consists of a two-phase training of the same base network: a pre-training phase according to a siamese model supervised by synonyms and a training phase on antonyms through a siamese-like model that supports the antitransitivity present in antonymy. The approach makes use of the claim that the antonyms in common of a word tend to be synonyms. We show that our approach outperforms distributional and pattern-based approaches, relaying on a simple feed forward network as base network of the training phases.

Incorporating Syntactic and Semantic Information in Word Embeddings using Graph Convolutional Networks

- https://www.aclweb.org/anthology/P19-1320/

Word embeddings have been widely adopted across several NLP applications. Most existing word embedding methods utilize sequential context of a word to learn its embedding. While there have been some attempts at utilizing syntactic context of a word, such methods result in an explosion of the vocabulary size. In this paper, we overcome this problem by proposing SynGCN, a flexible Graph Convolution based method for learning word embeddings. SynGCN utilizes the dependency context of a word without increasing the vocabulary size. Word embeddings learned by SynGCN outperform existing methods on various intrinsic and extrinsic tasks and provide an advantage when used with ELMo. We also propose SemGCN, an effective framework for incorporating diverse semantic knowledge for further enhancing learned word representations. We make the source code of both models available to encourage reproducible research.

Word and Document Embedding with vMF-Mixture Priors on Context Word Vectors

- https://www.aclweb.org/anthology/P19-1321/

Word embedding models typically learn two types of vectors: target word vectors and context word vectors. These vectors are normally learned such that they are predictive of some word co-occurrence statistic, but they are otherwise unconstrained. However, the words from a given language can be organized in various natural groupings, such as syntactic word classes (e.g. nouns, adjectives, verbs) and semantic themes (e.g. sports, politics, sentiment). Our hypothesis in this paper is that embedding models can be improved by explicitly imposing a cluster structure on the set of context word vectors. To this end, our model relies on the assumption that context word vectors are drawn from a mixture of von Mises-Fisher (vMF) distributions, where the parameters of this mixture distribution are jointly optimized with the word vectors. We show that this results in word vectors which are qualitatively different from those obtained with existing word embedding models. We furthermore show that our embedding model can also be used to learn high-quality document representations.

Delta Embedding Learning

- https://www.aclweb.org/anthology/P19-1322/

Unsupervised word embeddings have become a popular approach of word representation in NLP tasks. However there are limitations to the semantics represented by unsupervised embeddings, and inadequate fine-tuning of embeddings

can lead to suboptimal performance. We propose a novel learning technique called Delta Embedding Learning, which can be applied to general NLP tasks to improve performance by optimized tuning of the word embeddings. A structured regularization is applied to the embeddings to ensure they are tuned in an incremental way. As a result, the tuned word embeddings become better word representations by absorbing semantic information from supervision without "forgetting." We apply the method to various NLP tasks and see a consistent improvement in performance. Evaluation also confirms the tuned word embeddings have better semantic properties.

Annotation and Automatic Classification of Aspectual Categories

- https://www.aclweb.org/anthology/P19-1323/

We present the first annotated resource for the aspectual classification of German verb tokens in their clausal context. We use aspectual features compatible with the plurality of aspectual classifications in previous work and treat aspectual ambiguity systematically. We evaluate our corpus by using it to train supervised classifiers to automatically assign aspectual categories to verbs in context, permitting favourable comparisons to previous work.

Putting Words in Context: LSTM Language Models and Lexical Ambiguity

- https://www.aclweb.org/anthology/P19-1324/

In neural network models of language, words are commonly represented using context-invariant representations (word embeddings) which are then put in context in the hidden layers. Since words are often ambiguous, representing the contextually relevant information is not trivial. We investigate how an LSTM language model deals with lexical ambiguity in English, designing a method to probe its hidden representations for lexical and contextual information about words. We find that both types of information are represented to a large extent, but also that there is room for improvement for contextual information.

Making Fast Graph-based Algorithms with Graph Metric Embeddings

- https://www.aclweb.org/anthology/P19-1325/

Graph measures, such as node distances, are inefficient to compute. We explore dense vector representations as an effective way to approximate the same information. We introduce a simple yet efficient and effective approach for learning

graph embeddings. Instead of directly operating on the graph structure, our method takes structural measures of pairwise node similarities into account and learns dense node representations reflecting user-defined graph distance measures, such as e.g. the shortest path distance or distance measures that take information beyond the graph structure into account. We demonstrate a speed-up of several orders of magnitude when predicting word similarity by vector operations on our embeddings as opposed to directly computing the respective path-based measures, while outperforming various other graph embeddings on semantic similarity and word sense disambiguation tasks.

Embedding Imputation with Grounded Language Information

- https://www.aclweb.org/anthology/P19-1326/

Due to the ubiquitous use of embeddings as input representations for a wide range of natural language tasks, imputation of embeddings for rare and unseen words is a critical problem in language processing. Embedding imputation involves learning representations for rare or unseen words during the training of an embedding model, often in a post-hoc manner. In this paper, we propose an approach for embedding imputation which uses grounded information in the form of a knowledge graph. This is in contrast to existing approaches which typically make use of vector space properties or subword information. We propose an online method to construct a graph from grounded information and design an algorithm to map from the resulting graphical structure to the space of the pre-trained embeddings. Finally, we evaluate our approach on a range of rare and unseen word tasks across various domains and show that our model can learn better representations. For example, on the Card-660 task our method improves Pearson's and Spearman's correlation coefficients upon the state-of-the-art by 11% and 17.8% respectively using GloVe embeddings.

The Effectiveness of Simple Hybrid Systems for Hypernym Discovery

- https://www.aclweb.org/anthology/P19-1327/

Hypernymy modeling has largely been separated according to two paradigms, pattern-based methods and distributional methods. However, recent works utilizing a mix of these strategies have yielded state-of-the-art results. This paper evaluates the contribution of both paradigms to hybrid success by evaluating the benefits of hybrid treatment of baseline models from each paradigm. Even with a simple methodology for each individual system, utilizing a hybrid approach establishes new state-of-the-art results on two domain-specific English hypernym discovery tasks and outperforms all non-hybrid approaches in a general English hypernym discovery task.

BERT-based Lexical Substitution

- https://www.aclweb.org/anthology/P19-1328/

Previous studies on lexical substitution tend to obtain substitute candidates by finding the target word's synonyms from lexical resources (e.g., WordNet) and then rank the candidates based on its contexts. These approaches have two limitations: (1) They are likely to overlook good substitute candidates that are not the synonyms of the target words in the lexical resources; (2) They fail to take into account the substitution's influence on the global context of the sentence. To address these issues, we propose an end-to-end BERT-based lexical substitution approach which can propose and validate substitute candidates without using any annotated data or manually curated resources. Our approach first applies dropout to the target word's embedding for partially masking the word, allowing BERT to take balanced consideration of the target word's semantics and contexts for proposing substitute candidates, and then validates the candidates based on their substitution's influence on the global contextualized representation of the sentence. Experiments show our approach performs well in both proposing and ranking substitute candidates, achieving the state-of-the-art results in both LS07 and LS14 benchmarks.

Exploring Numeracy in Word Embeddings

- https://www.aclweb.org/anthology/P19-1329/

Word embeddings are now pervasive across NLP subfields as the de-facto method of forming text representataions. In this work, we show that existing embedding models are inadequate at constructing representations that capture salient aspects of mathematical meaning for numbers, which is important for language understanding. Numbers are ubiquitous and frequently appear in text. Inspired by cognitive studies on how humans perceive numbers, we develop an analysis framework to test how well word embeddings capture two essential properties of numbers: magnitude (e.g. 3<4) and numeration (e.g. 3=three). Our experiments reveal that most models capture an approximate notion of magnitude, but are inadequate at capturing numeration. We hope that our observations provide a starting point for the development of methods which better capture numeracy in NLP systems.

HighRES: Highlight-based Reference-less Evaluation of Summarization

- https://www.aclweb.org/anthology/P19-1330/

There has been substantial progress in summarization research enabled by the availability of novel, often large-scale, datasets and recent advances on neural

network-based approaches. However, manual evaluation of the system generated summaries is inconsistent due to the difficulty the task poses to human non-expert readers. To address this issue, we propose a novel approach for manual evaluation, Highlight-based Reference-less Evaluation of Summarization (HighRES), in which summaries are assessed by multiple annotators against the source document via manually highlighted salient content in the latter. Thus summary assessment on the source document by human judges is facilitated, while the highlights can be used for evaluating multiple systems. To validate our approach we employ crowd-workers to augment with highlights a recently proposed dataset and compare two state-of-the-art systems. We demonstrate that HighRES improves inter-annotator agreement in comparison to using the source document directly, while they help emphasize differences among systems that would be ignored under other evaluation approaches.

EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing

- https://www.aclweb.org/anthology/P19-1331/

We present the first sentence simplification model that learns explicit edit operations (ADD, DELETE, and KEEP) via a neural programmer-interpreter approach. Most current neural sentence simplification systems are variants of sequence-to-sequence models adopted from machine translation. These methods learn to simplify sentences as a byproduct of the fact that they are trained on complex-simple sentence pairs. By contrast, our neural programmer-interpreter is directly trained to predict explicit edit operations on targeted parts of the input sentence, resembling the way that humans perform simplification and revision. Our model outperforms previous state-of-the-art neural sentence simplification models (without external knowledge) by large margins on three benchmark text simplification corpora in terms of SARI (+0.95 WikiLarge, +1.89 WikiSmall, +1.41 Newsela), and is judged by humans to produce overall better and simpler output sentences.

Decomposable Neural Paraphrase Generation

- https://www.aclweb.org/anthology/P19-1332/

Paraphrasing exists at different granularity levels, such as lexical level, phrasal level and sentential level. This paper presents Decomposable Neural Paraphrase Generator (DNPG), a Transformer-based model that can learn and generate paraphrases of a sentence at different levels of granularity in a disentangled way. Specifically, the model is composed of multiple encoders and decoders with different structures, each of which corresponds to a specific granularity. The empirical study shows that the decomposition mechanism of DNPG makes paraphrase generation more interpretable and controllable. Based on DNPG,

we further develop an unsupervised domain adaptation method for paraphrase generation. Experimental results show that the proposed model achieves competitive in-domain performance compared to state-of-the-art neural models, and significantly better performance when adapting to a new domain.

Transforming Complex Sentences into a Semantic Hierarchy

- https://www.aclweb.org/anthology/P19-1333/

We present an approach for recursively splitting and rephrasing complex English sentences into a novel semantic hierarchy of simplified sentences, with each of them presenting a more regular structure that may facilitate a wide variety of artificial intelligence tasks, such as machine translation (MT) or information extraction (IE). Using a set of hand-crafted transformation rules, input sentences are recursively transformed into a two-layered hierarchical representation in the form of core sentences and accompanying contexts that are linked via rhetorical relations. In this way, the semantic relationship of the decomposed constituents is preserved in the output, maintaining its interpretability for downstream applications. Both a thorough manual analysis and automatic evaluation across three datasets from two different domains demonstrate that the proposed syntactic simplification approach outperforms the state of the art in structural text simplification. Moreover, an extrinsic evaluation shows that when applying our framework as a preprocessing step the performance of state-of-the-art Open IE systems can be improved by up to 346% in precision and 52% in recall. To enable reproducible research, all code is provided online.

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

- https://www.aclweb.org/anthology/P19-1334/

A machine learning system can score well on a given test set by relying on heuristics that are effective for frequent example types but break down in more challenging cases. We study this issue within natural language inference (NLI), the task of determining whether one sentence entails another. We hypothesize that statistical NLI models may adopt three fallible syntactic heuristics: the lexical overlap heuristic, the subsequence heuristic, and the constituent heuristic. To determine whether models have adopted these heuristics, we introduce a controlled evaluation set called HANS (Heuristic Analysis for NLI Systems), which contains many examples where the heuristics fail. We find that models trained on MNLI, including BERT, a state-of-the-art model, perform very poorly on HANS, suggesting that they have indeed adopted these heuristics. We conclude that there is substantial room for improvement in NLI systems, and that the HANS dataset can motivate and measure progress in this area.

Zero-Shot Entity Linking by Reading Entity Descriptions

- https://www.aclweb.org/anthology/P19-1335/

We present the zero-shot entity linking task, where mentions must be linked to unseen entities without in-domain labeled data. The goal is to enable robust transfer to highly specialized domains, and so no metadata or alias tables are assumed. In this setting, entities are only identified by text descriptions, and models must rely strictly on language understanding to resolve the new entities. First, we show that strong reading comprehension models pre-trained on large unlabeled data can be used to generalize to unseen entities. Second, we propose a simple and effective adaptive pre-training strategy, which we term domain-adaptive pre-training (DAP), to address the domain shift problem associated with linking unseen entities in a new domain. We present experiments on a new dataset that we construct for this task and show that DAP improves over strong pre-training baselines, including BERT. The data and code are available at https://github.com/lajanugen/zeshel.

Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition

- https://www.aclweb.org/anthology/P19-1336/

We propose a new neural transfer method termed Dual Adversarial Transfer Network (DATNet) for addressing low-resource Named Entity Recognition (NER). Specifically, two variants of DATNet, i.e., DATNet-F and DATNet-P, are investigated to explore effective feature fusion between high and low resource. To address the noisy and imbalanced training data, we propose a novel Generalized Resource-Adversarial Discriminator (GRAD). Additionally, adversarial training is adopted to boost model generalization. In experiments, we examine the effects of different components in DATNet across domains and languages and show that significant improvement can be obtained especially for low-resource data, without augmenting any additional hand-crafted features and pre-trained language model.

Scalable Syntax-Aware Language Models Using Knowledge Distillation

- https://www.aclweb.org/anthology/P19-1337/

Prior work has shown that, on small amounts of training data, syntactic neural language models learn structurally sensitive generalisations more successfully than sequential language models. However, their computational complexity renders scaling difficult, and it remains an open question whether structural biases are still necessary when sequential models have access to ever larger amounts

of training data. To answer this question, we introduce an efficient knowledge distillation (KD) technique that transfers knowledge from a syntactic language model trained on a small corpus to an LSTM language model, hence enabling the LSTM to develop a more structurally sensitive representation of the larger training data it learns from. On targeted syntactic evaluations, we find that, while sequential LSTMs perform much better than previously reported, our proposed technique substantially improves on this baseline, yielding a new state of the art. Our findings and analysis affirm the importance of structural biases, even in models that learn from large amounts of data.

An Imitation Learning Approach to Unsupervised Parsing

- https://www.aclweb.org/anthology/P19-1338/

Recently, there has been an increasing interest in unsupervised parsers that optimize semantically oriented objectives, typically using reinforcement learning. Unfortunately, the learned trees often do not match actual syntax trees well. Shen et al. (2018) propose a structured attention mechanism for language modeling (PRPN), which induces better syntactic structures but relies on ad hoc heuristics. Also, their model lacks interpretability as it is not grounded in parsing actions. In our work, we propose an imitation learning approach to unsupervised parsing, where we transfer the syntactic knowledge induced by PRPN to a Tree-LSTM model with discrete parsing actions. Its policy is then refined by Gumbel-Softmax training towards a semantically oriented objective. We evaluate our approach on the All Natural Language Inference dataset and show that it achieves a new state of the art in terms of parsing F-score, outperforming our base models, including PRPN.

Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing

- https://www.aclweb.org/anthology/P19-1339/

Several linguistic studies have shown the prevalence of various lexical and grammatical patterns in texts authored by a person of a particular gender, but models for part-of-speech tagging and dependency parsing have still not adapted to account for these differences. To address this, we annotate the Wall Street Journal part of the Penn Treebank with the gender information of the articles' authors, and build taggers and parsers trained on this data that show performance differences in text written by men and women. Further analyses reveal numerous part-of-speech tags and syntactic relations whose prediction performances benefit from the prevalence of a specific gender in the training data. The results underscore the importance of accounting for gendered differences in

syntactic tasks, and outline future venues for developing more accurate taggers and parsers. We release our data to the research community.

Multilingual Constituency Parsing with Self-Attention and Pre-Training

- https://www.aclweb.org/anthology/P19-1340/

We show that constituency parsing benefits from unsupervised pre-training across a variety of languages and a range of pre-training conditions. We first compare the benefits of no pre-training, fastText, ELMo, and BERT for English and find that BERT outperforms ELMo, in large part due to increased model capacity, whereas ELMo in turn outperforms the non-contextual fastText embeddings. We also find that pre-training is beneficial across all 11 languages tested; however, large model sizes (more than 100 million parameters) make it computationally expensive to train separate models for each language. To address this shortcoming, we show that joint multilingual pre-training and fine-tuning allows sharing all but a small number of parameters between ten languages in the final model. The 10x reduction in model size compared to fine-tuning one model per language causes only a 3.2% relative error increase in aggregate. We further explore the idea of joint fine-tuning and show that it gives low-resource languages a way to benefit from the larger datasets of other languages. Finally, we demonstrate new state-of-the-art results for 11 languages, including English (95.8 F1) and Chinese (91.8 F1).

A Multilingual BPE Embedding Space for Universal Sentiment Lexicon Induction

- https://www.aclweb.org/anthology/P19-1341/

We present a new method for sentiment lexicon induction that is designed to be applicable to the entire range of typological diversity of the world's languages. We evaluate our method on Parallel Bible Corpus+ (PBC+), a parallel corpus of 1593 languages. The key idea is to use Byte Pair Encodings (BPEs) as basic units for multilingual embeddings. Through zero-shot transfer from English sentiment, we learn a seed lexicon for each language in the domain of PBC+. Through domain adaptation, we then generalize the domain-specific lexicon to a general one. We show – across typologically diverse languages in PBC+ – good quality of seed and general-domain sentiment lexicons by intrinsic and extrinsic and by automatic and human evaluation. We make freely available our code, seed sentiment lexicons for all 1593 languages and induced general-domain sentiment lexicons for 200 languages.

Tree Communication Models for Sentiment Analysis

- https://www.aclweb.org/anthology/P19-1342/

Tree-LSTMs have been used for tree-based sentiment analysis over Stanford Sentiment Treebank, which allows the sentiment signals over hierarchical phrase structures to be calculated simultaneously. However, traditional tree-LSTMs capture only the bottom-up dependencies between constituents. In this paper, we propose a tree communication model using graph convolutional neural network and graph recurrent neural network, which allows rich information exchange between phrases constituent tree. Experiments show that our model outperforms existing work on bidirectional tree-LSTMs in both accuracy and efficiency, providing more consistent predictions on phrase-level sentiments.

Improved Sentiment Detection via Label Transfer from Monolingual to Synthetic Code-Switched Text

- https://www.aclweb.org/anthology/P19-1343/

Multilingual writers and speakers often alternate between two languages in a single discourse. This practice is called "code-switching". Existing sentiment detection methods are usually trained on sentiment-labeled monolingual text. Manually labeled code-switched text, especially involving minority languages, is extremely rare. Consequently, the best monolingual methods perform relatively poorly on code-switched text. We present an effective technique for synthesizing labeled code-switched text from labeled monolingual text, which is relatively readily available. The idea is to replace carefully selected subtrees of constituency parses of sentences in the resource-rich language with suitable token spans selected from automatic translations to the resource-poor language. By augmenting the scarce labeled code-switched text with plentiful synthetic labeled code-switched text, we achieve significant improvements in sentiment labeling accuracy (1.5%, 5.11% 7.20%) for three different language pairs (English-Hindi, English-Spanish and English-Bengali). The improvement is even significant in hatespeech detection whereby we achieve a 4% improvement using only synthetic code-switched data (6% with data augmentation).

Exploring Sequence-to-Sequence Learning in Aspect Term Extraction

- https://www.aclweb.org/anthology/P19-1344/

Aspect term extraction (ATE) aims at identifying all aspect terms in a sentence and is usually modeled as a sequence labeling problem. However, sequence labeling based methods cannot make full use of the overall meaning of the whole sentence and have the limitation in processing dependencies between labels.

To tackle these problems, we first explore to formalize ATE as a sequence-to-sequence (Seq2Seq) learning task where the source sequence and target sequence are composed of words and labels respectively. At the same time, to make Seq2Seq learning suit to ATE where labels correspond to words one by one, we design the gated unit networks to incorporate corresponding word representation into the decoder, and position-aware attention to pay more attention to the adjacent words of a target word. The experimental results on two datasets show that Seq2Seq learning is effective in ATE accompanied with our proposed gated unit networks and position-aware attention mechanism.

Aspect Sentiment Classification Towards Question-Answering with Reinforced Bidirectional Attention Network

- https://www.aclweb.org/anthology/P19-1345/

In the literature, existing studies on aspect sentiment classification (ASC) focus on individual non-interactive reviews. This paper extends the research to interactive reviews and proposes a new research task, namely Aspect Sentiment Classification towards Question-Answering (ASC-QA), for real-world applications. This new task aims to predict sentiment polarities for specific aspects from interactive QA style reviews. In particular, a high-quality annotated corpus is constructed for ASC-QA to facilitate corresponding research. On this basis, a Reinforced Bidirectional Attention Network (RBAN) approach is proposed to address two inherent challenges in ASC-QA, i.e., semantic matching between question and answer, and data noise. Experimental results demonstrate the great advantage of the proposed approach to ASC-QA against several state-of-the-art baselines.

ELI5: Long Form Question Answering

- https://www.aclweb.org/anthology/P19-1346/

We introduce the first large-scale corpus for long form question answering, a task requiring elaborate and in-depth answers to open-ended questions. The dataset comprises 270K threads from the Reddit forum "Explain Like I'm Five" (ELI5) where an online community provides answers to questions which are comprehensible by five year olds. Compared to existing datasets, ELI5 comprises diverse questions requiring multi-sentence answers. We provide a large set of web documents to help answer the question. Automatic and human evaluations show that an ive model trained with a multi-task objective outperforms conventional Seq2Seq, language modeling, as well as a strong extractive baseline.However, our best model is still far from human performance since raters prefer gold responses in over 86% of cases, leaving ample opportunity for future improvement.

Textbook Question Answering with Multi-modal Context Graph Understanding and Self-supervised Open-set Comprehension

- https://www.aclweb.org/anthology/P19-1347/

In this work, we introduce a novel algorithm for solving the textbook question answering (TQA) task which describes more realistic QA problems compared to other recent tasks. We mainly focus on two related issues with analysis of the TQA dataset. First, solving the TQA problems requires to comprehend multi-modal contexts in complicated input data. To tackle this issue of extracting knowledge features from long text lessons and merging them with visual features, we establish a context graph from texts and images, and propose a new module f-GCN based on graph convolutional networks (GCN). Second, scientific terms are not spread over the chapters and subjects are split in the TQA dataset. To overcome this so called 'out-of-domain' issue, before learning QA problems, we introduce a novel self-supervised open-set learning process without any annotations. The experimental results show that our model significantly outperforms prior state-of-the-art methods. Moreover, ablation studies validate that both methods of incorporating f-GCN for extracting knowledge from multi-modal contexts and our newly proposed self-supervised learning process are effective for TQA problems.

Generating Question Relevant Captions to Aid Visual Question Answering

- https://www.aclweb.org/anthology/P19-1348/

Visual question answering (VQA) and image captioning require a shared body of general knowledge connecting language and vision. We present a novel approach to better VQA performance that exploits this connection by jointly generating captions that are targeted to help answer a specific visual question. The model is trained using an existing caption dataset by automatically determining question-relevant captions using an online gradient-based method. Experimental results on the VQA v2 challenge demonstrates that our approach obtains state-of-the-art VQA performance (e.g. 68.4% in the Test-standard set using a single model) by simultaneously generating question-relevant captions.

Multi-grained Attention with Object-level Grounding for Visual Question Answering

- https://www.aclweb.org/anthology/P19-1349/

Attention mechanisms are widely used in Visual Question Answering (VQA) to search for visual clues related to the question. Most approaches train attention

models from a coarse-grained association between sentences and images, which tends to fail on small objects or uncommon concepts. To address this problem, this paper proposes a multi-grained attention method. It learns explicit word-object correspondence by two types of word-level attention complementary to the sentence-image association. Evaluated on the VQA benchmark, the multi-grained attention model achieves competitive performance with state-of-the-art models. And the visualized attention maps demonstrate that addition of object-level groundings leads to a better understanding of the images and locates the attended objects more precisely.

Psycholinguistics Meets Continual Learning: Measuring Catastrophic Forgetting in Visual Question Answering

- https://www.aclweb.org/anthology/P19-1350/

We study the issue of catastrophic forgetting in the context of neural multimodal approaches to Visual Question Answering (VQA). Motivated by evidence from psycholinguistics, we devise a set of linguistically-informed VQA tasks, which differ by the types of questions involved (Wh-questions and polar questions). We test what impact task difficulty has on continual learning, and whether the order in which a child acquires question types facilitates computational models. Our results show that dramatic forgetting is at play and that task difficulty and order matter. Two well-known current continual learning methods mitigate the problem only to a limiting degree.

Improving Visual Question Answering by Referring to Generated Paragraph Captions

- https://www.aclweb.org/anthology/P19-1351/

Paragraph-style image captions describe diverse aspects of an image as opposed to the more common single-sentence captions that only provide an description of the image. These paragraph captions can hence contain substantial information of the image for tasks such as visual question answering. Moreover, this textual information is complementary with visual information present in the image because it can discuss both more concepts and more explicit, intermediate symbolic information about objects, events, and scenes that can directly be matched with the textual question and copied into the textual answer (i.e., via easier modality match). Hence, we propose a combined Visual and Textual Question Answering (VTQA) model which takes as input a paragraph caption as well as the corresponding image, and answers the given question based on both inputs. In our model, the inputs are fused to extract related information by cross-attention (early fusion), then fused again in the form of consensus (late fusion), and finally expected answers are given an extra score to enhance the

chance of selection (later fusion). Empirical results show that paragraph captions, even when automatically generated (via an RL-based encoder-decoder model), help correctly answer more visual questions. Overall, our joint model, when trained on the Visual Genome dataset, significantly improves the VQA performance over a strong baseline model.

Shared-Private Bilingual Word Embeddings for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1352/

Word embedding is central to neural machine translation (NMT), which has attracted intensive research interest in recent years. In NMT, the source embedding plays the role of the entrance while the target embedding acts as the terminal. These layers occupy most of the model parameters for representation learning. Furthermore, they indirectly interface via a soft-attention mechanism, which makes them comparatively isolated. In this paper, we propose shared-private bilingual word embeddings, which give a closer relationship between the source and target embeddings, and which also reduce the number of model parameters. For similar source and target words, their embeddings tend to share a part of the features and they cooperatively learn these common representation units. Experiments on 5 language pairs belonging to 6 different language families and written in 5 different alphabets demonstrate that the proposed model provides a significant performance boost over the strong baselines with dramatically fewer model parameters.

Literary Event Detection

- https://www.aclweb.org/anthology/P19-1353/

In this work we present a new dataset of literary events—events that are depicted as taking place within the imagined space of a novel. While previous work has focused on event detection in the domain of contemporary news, literature poses a number of complications for existing systems, including complex narration, the depiction of a broad array of mental states, and a strong emphasis on figurative language. We outline the annotation decisions of this new dataset and compare several models for predicting events; the best performing model, a bidirectional LSTM with BERT token representations, achieves an F1 score of 73.9. We then apply this model to a corpus of novels split across two dimensions—prestige and popularity—and demonstrate that there are statistically significant differences in the distribution of events for prestige.

Assessing the Ability of Self-Attention Networks to Learn Word Order

- https://www.aclweb.org/anthology/P19-1354/

Self-attention networks (SAN) have attracted a lot of interests due to their high parallelization and strong performance on a variety of NLP tasks, e.g. machine translation. Due to the lack of recurrence structure such as recurrent neural networks (RNN), SAN is ascribed to be weak at learning positional information of words for sequence modeling. However, neither this speculation has been empirically confirmed, nor explanations for their strong performances on machine translation tasks when "lacking positional information" have been explored. To this end, we propose a novel word reordering detection task to quantify how well the word order information learned by SAN and RNN. Specifically, we randomly move one word to another position, and examine whether a trained model can detect both the original and inserted positions. Experimental results reveal that: 1) SAN trained on word reordering detection indeed has difficulty learning the positional information even with the position embedding; and 2) SAN trained on machine translation learns better positional information than its RNN counterpart, in which position embedding plays a critical role. Although recurrence structure make the model more universally-effective on learning word order, learning objectives matter more in the downstream tasks such as machine translation.

Energy and Policy Considerations for Deep Learning in NLP

- https://www.aclweb.org/anthology/P19-1355/

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware. In this paper we bring this issue to the attention of NLP researchers by quantifying the approximate financial and environmental costs of training a variety of recently successful neural network models for NLP. Based on these findings, we propose actionable recommendations to reduce costs and improve equity in NLP research and practice.

What Does BERT Learn about the Structure of Language?

- https://www.aclweb.org/anthology/P19-1356/

BERT is a recent language representation model that has surprisingly performed well in diverse language understanding benchmarks. This result indicates the possibility that BERT networks capture structural information about language. In this work, we provide novel support for this claim by performing a series of experiments to unpack the elements of English language structure learned by BERT. Our findings are fourfold. BERT's phrasal representation captures the phrase-level information in the lower layers. The intermediate layers of BERT compose a rich hierarchy of linguistic information, starting with surface features at the bottom, syntactic features in the middle followed by semantic features at the top. BERT requires deeper layers while tracking subject-verb agreement to handle long-term dependency problem. Finally, the compositional scheme underlying BERT mimics classical, tree-like structures.

A Just and Comprehensive Strategy for Using NLP to Address Online Abuse

- https://www.aclweb.org/anthology/P19-1357/

Online abusive behavior affects millions and the NLP community has attempted to mitigate this problem by developing technologies to detect abuse. However, current methods have largely focused on a narrow definition of abuse to detriment of victims who seek both validation and solutions. In this position paper, we argue that the community needs to make three substantive changes: (1) expanding our scope of problems to tackle both more subtle and more serious forms of abuse, (2) developing proactive technologies that counter or inhibit abuse before it harms, and (3) reframing our effort within a framework of justice to promote healthy communities.

Learning from Dialogue after Deployment: Feed Yourself, Chatbot!

- https://www.aclweb.org/anthology/P19-1358/

The majority of conversations a dialogue agent sees over its lifetime occur after it has already been trained and deployed, leaving a vast store of potential training signal untapped. In this work, we propose the self-feeding chatbot, a dialogue agent with the ability to extract new training examples from the conversations it participates in. As our agent engages in conversation, it also estimates user satisfaction in its responses. When the conversation appears to be going well, the user's responses become new training examples to imitate. When the agent believes it has made a mistake, it asks for feedback; learning to predict the

feedback that will be given improves the chatbot's dialogue abilities further. On the PersonaChat chit-chat dataset with over 131k training examples, we find that learning from dialogue with a self-feeding chatbot significantly improves performance, regardless of the amount of traditional supervision.

Generating Responses with a Specific Emotion in Dialog

- https://www.aclweb.org/anthology/P19-1359/

It is desirable for dialog systems to have capability to express specific emotions during a conversation, which has a direct, quantifiable impact on improvement of their usability and user satisfaction. After a careful investigation of real-life conversation data, we found that there are at least two ways to express emotions with language. One is to describe emotional states by explicitly using strong emotional words; another is to increase the intensity of the emotional experiences by implicitly combining neutral words in distinct ways. We propose an emotional dialogue system (EmoDS) that can generate the meaningful responses with a coherent structure for a post, and meanwhile express the desired emotion explicitly or implicitly within a unified framework. Experimental results showed EmoDS performed better than the baselines in BLEU, diversity and the quality of emotional expression.

Semantically Conditioned Dialog Response Generation via Hierarchical Disentangled Self-Attention

- https://www.aclweb.org/anthology/P19-1360/

Semantically controlled neural response generation on limited-domain has achieved great performance. However, moving towards multi-domain large-scale scenarios are shown to be difficult because the possible combinations of semantic inputs grow exponentially with the number of domains. To alleviate such scalability issue, we exploit the structure of dialog acts to build a multi-layer hierarchical graph, where each act is represented as a root-to-leaf route on the graph. Then, we incorporate such graph structure prior as an inductive bias to build a hierarchical disentangled self-attention network, where we disentangle attention heads to model designated nodes on the dialog act graph. By activating different (disentangled) heads at each layer, combinatorially many dialog act semantics can be modeled to control the neural response generation. On the large-scale Multi-Domain-WOZ dataset, our model can yield a significant improvement over the baselines on various automatic and human evaluation metrics.

Incremental Learning from Scratch for Task-Oriented Dialogue Systems

- https://www.aclweb.org/anthology/P19-1361/

Clarifying user needs is essential for existing task-oriented dialogue systems. However, in real-world applications, developers can never guarantee that all possible user demands are taken into account in the design phase. Consequently, existing systems will break down when encountering unconsidered user needs. To address this problem, we propose a novel incremental learning framework to design task-oriented dialogue systems, or for short Incremental Dialogue System (IDS), without pre-defining the exhaustive list of user needs. Specifically, we introduce an uncertainty estimation module to evaluate the confidence of giving correct responses. If there is high confidence, IDS will provide responses to users. Otherwise, humans will be involved in the dialogue process, and IDS can learn from human intervention through an online learning module. To evaluate our method, we propose a new dataset which simulates unanticipated user needs in the deployment stage. Experiments show that IDS is robust to unconsidered user actions, and can update itself online by smartly selecting only the most effective training data, and hence attains better performance with less annotation cost.

ReCoSa: Detecting the Relevant Contexts with Self-Attention for Multi-turn Dialogue Generation

- https://www.aclweb.org/anthology/P19-1362/

In multi-turn dialogue generation, response is usually related with only a few contexts. Therefore, an ideal model should be able to detect these relevant contexts and produce a suitable response accordingly. However, the widely used hierarchical recurrent encoder-decoder models just treat all the contexts indiscriminately, which may hurt the following response generation process. Some researchers try to use the cosine similarity or the traditional attention mechanism to find the relevant contexts, but they suffer from either insufficient relevance assumption or position bias problem. In this paper, we propose a new model, named ReCoSa, to tackle this problem. Firstly, a word level LSTM encoder is conducted to obtain the initial representation of each context. Then, the self-attention mechanism is utilized to update both the context and masked response representation. Finally, the attention weights between each context and response representations are computed and used in the further decoding process. Experimental results on both Chinese customer services dataset and English Ubuntu dialogue dataset show that ReCoSa significantly outperforms baseline models, in terms of both metric-based and human evaluations. Further analysis on attention shows that the detected relevant contexts by ReCoSa are highly coherent with human's understanding, validating the correctness and interpretability of ReCoSa.

Dialogue Natural Language Inference

- https://www.aclweb.org/anthology/P19-1363/

Consistency is a long standing issue faced by dialogue models. In this paper, we frame the consistency of dialogue agents as natural language inference (NLI) and create a new natural language inference dataset called Dialogue NLI. We propose a method which demonstrates that a model trained on Dialogue NLI can be used to improve the consistency of a dialogue model, and evaluate the method with human evaluation and with automatic metrics on a suite of evaluation sets designed to measure a dialogue model's consistency.

Budgeted Policy Learning for Task-Oriented Dialogue Systems

- https://www.aclweb.org/anthology/P19-1364/

This paper presents a new approach that extends Deep Dyna-Q (DDQ) by incorporating a Budget-Conscious Scheduling (BCS) to best utilize a fixed, small amount of user interactions (budget) for learning task-oriented dialogue agents. BCS consists of (1) a Poisson-based global scheduler to allocate budget over different stages of training; (2) a controller to decide at each training step whether the agent is trained using real or simulated experiences; (3) a user goal sampling module to generate the experiences that are most effective for policy learning. Experiments on a movie-ticket booking task with simulated and real users show that our approach leads to significant improvements in success rate over the state-of-the-art baselines given the fixed budget.

Comparison of Diverse Decoding Methods from Conditional Language Models

- https://www.aclweb.org/anthology/P19-1365/

While conditional language models have greatly improved in their ability to output high quality natural language, many NLP applications benefit from being able to generate a diverse set of candidate sequences. Diverse decoding strategies aim to, within a given-sized candidate list, cover as much of the space of high-quality outputs as possible, leading to improvements for tasks that rerank and combine candidate outputs. Standard decoding methods, such as beam search, optimize for generating high likelihood sequences rather than diverse ones, though recent work has focused on increasing diversity in these methods. In this work, we perform an extensive survey of decoding-time strategies for generating diverse outputs from a conditional language model. In addition, we present a novel method where we over-sample candidates, then use clustering

to remove similar sequences, thus achieving high diversity without sacrificing quality.

Retrieval-Enhanced Adversarial Training for Neural Response Generation

- https://www.aclweb.org/anthology/P19-1366/

Dialogue systems are usually built on either generation-based or retrieval-based approaches, yet they do not benefit from the advantages of different models. In this paper, we propose a Retrieval-Enhanced Adversarial Training (REAT) method for neural response generation. Distinct from existing approaches, the REAT method leverages an encoder-decoder framework in terms of an adversarial training paradigm, while taking advantage of N-best response candidates from a retrieval-based system to construct the discriminator. An empirical study on a large scale public available benchmark dataset shows that the REAT method significantly outperforms the vanilla Seq2Seq model as well as the conventional adversarial training approach.

Vocabulary Pyramid Network: Multi-Pass Encoding and Decoding with Multi-Level Vocabularies for Response Generation

- https://www.aclweb.org/anthology/P19-1367/

We study the task of response generation. Conventional methods employ a fixed vocabulary and one-pass decoding, which not only make them prone to safe and general responses but also lack further refining to the first generated raw sequence. To tackle the above two problems, we present a Vocabulary Pyramid Network (VPN) which is able to incorporate multi-pass encoding and decoding with multi-level vocabularies into response generation. Specifically, the dialogue input and output are represented by multi-level vocabularies which are obtained from hierarchical clustering of raw words. Then, multi-pass encoding and decoding are conducted on the multi-level vocabularies. Since VPN is able to leverage rich encoding and decoding information with multi-level vocabularies, it has the potential to generate better responses. Experiments on English Twitter and Chinese Weibo datasets demonstrate that VPN remarkably outperforms strong baselines.

On-device Structured and Context Partitioned Projection Networks

- https://www.aclweb.org/anthology/P19-1368/

A challenging problem in on-device text classification is to build highly accurate neural models that can fit in small memory footprint and have low latency. To address this challenge, we propose an on-device neural network SGNN++ which dynamically learns compact projection vectors from raw text using structured and context-dependent partition projections. We show that this results in accelerated inference and performance improvements. We conduct extensive evaluation on multiple conversational tasks and languages such as English, Japanese, Spanish and French. Our SGNN++ model significantly outperforms all baselines, improves upon existing on-device neural models and even surpasses RNN, CNN and BiLSTM models on dialog act and intent prediction. Through a series of ablation studies we show the impact of the partitioned projections and structured information leading to 10% improvement. We study the impact of the model size on accuracy and introduce quatization-aware training for SGNN++ to further reduce the model size while preserving the same quality. Finally, we show fast inference on mobile phones.

Proactive Human-Machine Conversation with Explicit Conversation Goal

- https://www.aclweb.org/anthology/P19-1369/

Though great progress has been made for human-machine conversation, current dialogue system is still in its infancy: it usually converses passively and utters words more as a matter of response, rather than on its own initiatives. In this paper, we take a radical step towards building a human-like conversational agent: endowing it with the ability of proactively leading the conversation (introducing a new topic or maintaining the current topic). To facilitate the development of such conversation systems, we create a new dataset named Konv where one acts as a conversation leader and the other acts as the follower. The leader is provided with a knowledge graph and asked to sequentially change the discussion topics, following the given conversation goal, and meanwhile keep the dialogue as natural and engaging as possible. Konv enables a very challenging task as the model needs to both understand dialogue and plan over the given knowledge graph. We establish baseline results on this dataset (about 270K utterances and 30k dialogues) using several state-of-the-art models. Experimental results show that dialogue models that plan over the knowledge graph can make full use of related knowledge to generate more diverse multi-turn conversations. The baseline systems along with the dataset are publicly available.

Learning a Matching Model with Co-teaching for Multi-turn Response Selection in Retrieval-based Dialogue Systems

- https://www.aclweb.org/anthology/P19-1370/

We study learning of a matching model for response selection in retrieval-based dialogue systems. The problem is equally important with designing the architecture of a model, but is less explored in existing literature. To learn a robust matching model from noisy training data, we propose a general co-teaching framework with three specific teaching strategies that cover both teaching with loss functions and teaching with data curriculum. Under the framework, we simultaneously learn two matching models with independent training sets. In each iteration, one model transfers the knowledge learned from its training set to the other model, and at the same time receives the guide from the other model on how to overcome noise in training. Through being both a teacher and a student, the two models learn from each other and get improved together. Evaluation results on two public data sets indicate that the proposed learning approach can generally and significantly improve the performance of existing matching models.

Learning to for Memory-augmented Conversational Response Generation

- https://www.aclweb.org/anthology/P19-1371/

Neural generative models for open-domain chit-chat conversations have become an active area of research in recent years. A critical issue with most existing generative models is that the generated responses lack informativeness and diversity. A few researchers attempt to leverage the results of retrieval models to strengthen the generative models, but these models are limited by the quality of the retrieval results. In this work, we propose a memory-augmented generative model, which learns to from the training corpus and saves the useful information to the memory to assist the response generation. Our model clusters query-response samples, extracts characteristics of each cluster, and learns to utilize these characteristics for response generation. Experimental results show that our model outperforms other competitive baselines.

Are Training Samples Correlated? Learning to Generate Dialogue Responses with Multiple References

- https://www.aclweb.org/anthology/P19-1372/

Due to its potential applications, open-domain dialogue generation has become popular and achieved remarkable progress in recent years, but sometimes suffers from generic responses. Previous models are generally trained based on 1-to-1 mapping from an input query to its response, which actually ignores the nature of 1-to-n mapping in dialogue that there may exist multiple valid responses corresponding to the same query. In this paper, we propose to utilize the multiple references by considering the correlation of different valid responses and modeling the 1-to-n mapping with a novel two-step generation architecture. The first

generation phase extracts the common features of different responses which, combined with distinctive features obtained in the second phase, can generate multiple diverse and appropriate responses. Experimental results show that our proposed model can effectively improve the quality of response and outperform existing neural dialogue models on both automatic and human evaluations.

Pretraining Methods for Dialog Context Representation Learning

- https://www.aclweb.org/anthology/P19-1373/

This paper examines various unsupervised pretraining objectives for learning dialog context representations. Two novel methods of pretraining dialog context encoders are proposed, and a total of four methods are examined. Each pretraining objective is fine-tuned and evaluated on a set of downstream dialog tasks using the MultiWoz dataset and strong performance improvement is observed. Further evaluation shows that our pretraining objectives result in not only better performance, but also better convergence, models that are less data hungry and have better domain generalizability.

A Large-Scale Corpus for Conversation Disentanglement

- https://www.aclweb.org/anthology/P19-1374/

Disentangling conversations mixed together in a single stream of messages is a difficult task, made harder by the lack of large manually annotated datasets. We created a new dataset of 77,563 messages manually annotated with reply-structure graphs that both disentangle conversations and define internal conversation structure. Our data is 16 times larger than all previously released datasets combined, the first to include adjudication of annotation disagreements, and the first to include context. We use our data to re-examine prior work, in particular, finding that 89% of conversations in a widely used dialogue corpus are either missing messages or contain extra messages. Our manually-annotated data presents an opportunity to develop robust data-driven methods for conversation disentanglement, which will help advance dialogue research.

Self-Supervised Dialogue Learning

- https://www.aclweb.org/anthology/P19-1375/

The sequential order of utterances is often meaningful in coherent dialogues, and the order changes of utterances could lead to low-quality and incoherent conversations. We consider the order information as a crucial supervised signal for dialogue learning, which, however, has been neglected by many previous dialogue

systems. Therefore, in this paper, we introduce a self-supervised learning task, inconsistent order detection, to explicitly capture the flow of conversation in dialogues. Given a sampled utterance pair triple, the task is to predict whether it is ordered or misordered. Then we propose a sampling-based self-supervised network SSN to perform the prediction with sampled triple references from previous dialogue history. Furthermore, we design a joint learning framework where SSN can guide the dialogue systems towards more coherent and relevant dialogue learning through adversarial training. We demonstrate that the proposed methods can be applied to both open-domain and task-oriented dialogue scenarios, and achieve the new state-of-the-art performance on the OpenSubtitiles and Movie-Ticket Booking datasets.

Are we there yet? Encoder-decoder neural networks as cognitive models of English past tense inflection

- https://www.aclweb.org/anthology/P19-1376/

The cognitive mechanisms needed to account for the English past tense have long been a subject of debate in linguistics and cognitive science. Neural network models were proposed early on, but were shown to have clear flaws. Recently, however, Kirov and Cotterell (2018) showed that modern encoder-decoder (ED) models overcome many of these flaws. They also presented evidence that ED models demonstrate humanlike performance in a nonce-word task. Here, we look more closely at the behaviour of their model in this task. We find that (1) the model exhibits instability across multiple simulations in terms of its correlation with human data, and (2) even when results are aggregated across simulations (treating each simulation as an individual human participant), the fit to the human data is not strong—worse than an older rule-based model. These findings hold up through several alternative training regimes and evaluation measures. Although other neural architectures might do better, we conclude that there is still insufficient evidence to claim that neural nets are a good cognitive model for this task.

A Spreading Activation Framework for Tracking Conceptual Complexity of Texts

- https://www.aclweb.org/anthology/P19-1377/

We propose an unsupervised approach for assessing conceptual complexity of texts, based on spreading activation. Using DBpedia knowledge graph as a proxy to long-term memory, mentioned concepts become activated and trigger further activation as the text is sequentially traversed. Drawing inspiration from psycholinguistic theories of reading comprehension, we model memory processes such as semantic priming, sentence wrap-up, and forgetting. We show that our

models capture various aspects of conceptual text complexity and significantly outperform current state of the art.

End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories

- https://www.aclweb.org/anthology/P19-1378/

End-to-end training with Deep Neural Networks (DNN) is a currently popular method for metaphor identification. However, standard sequence tagging models do not explicitly take advantage of linguistic theories of metaphor identification. We experiment with two DNN models which are inspired by two human metaphor identification procedures. By testing on three public datasets, we find that our models achieve state-of-the-art performance in end-to-end metaphor identification.

Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View

- https://www.aclweb.org/anthology/P19-1379/

Diachronic word embeddings have been widely used in detecting temporal changes. However, existing methods face the meaning conflation deficiency by representing a word as a single vector at each time period. To address this issue, this paper proposes a sense representation and tracking framework based on deep contextualized embeddings, aiming at answering not only what and when, but also how the word meaning changes. The experiments show that our framework is effective in representing fine-grained word senses, and it brings a significant improvement in word change detection task. Furthermore, we model the word change from an ecological viewpoint, and sketch two interesting sense behaviors in the process of language evolution, i.e. sense competition and sense cooperation.

Miss Tools and Mr Fruit: Emergent Communication in Agents Learning about Object Affordances

- https://www.aclweb.org/anthology/P19-1380/

Recent research studies communication emergence in communities of deep network agents assigned a joint task, hoping to gain insights on human language evolution. We propose here a new task capturing crucial aspects of the human environment, such as natural object affordances, and of human conversation, such as full symmetry among the participants. By conducting a thorough pragmatic and semantic analysis of the emergent protocol, we show that the agents

solve the shared task through genuine bilateral, referential communication. However, the agents develop multiple idiolects, which makes us conclude that full symmetry is not a sufficient condition for a common language to emerge.

CNNs found to jump around more skillfully than RNNs: Compositional Generalization in Seq2seq Convolutional Networks

- https://www.aclweb.org/anthology/P19-1381/

Lake and Baroni (2018) introduced the SCAN dataset probing the ability of seq2seq models to capture compositional generalizations, such as inferring the meaning of "jump around" 0-shot from the component words. Recurrent networks (RNNs) were found to completely fail the most challenging generalization cases. We test here a convolutional network (CNN) on these tasks, reporting hugely improved performance with respect to RNNs. Despite the big improvement, the CNN has however not induced systematic rules, suggesting that the difference between compositional and non-compositional behaviour is not clear-cut.

Uncovering Probabilistic Implications in Typological Knowledge Bases

- https://www.aclweb.org/anthology/P19-1382/

The study of linguistic typology is rooted in the implications we find between linguistic features, such as the fact that languages with object-verb word ordering tend to have postpositions. Uncovering such implications typically amounts to time-consuming manual processing by trained and experienced linguists, which potentially leaves key linguistic universals unexplored. In this paper, we present a computational model which successfully identifies known universals, including Greenberg universals, but also uncovers new ones, worthy of further linguistic investigation. Our approach outperforms baselines previously used for this problem, as well as a strong baseline from knowledge base population.

Is Word Segmentation Child's Play in All Languages?

- https://www.aclweb.org/anthology/P19-1383/

When learning language, infants need to break down the flow of input speech into minimal word-like units, a process best described as unsupervised bottom-up segmentation. Proposed strategies include several segmentation algorithms, but only cross-linguistically robust algorithms could be plausible candidates for human word learning, since infants have no initial knowledge of the ambient

language. We report on the stability in performance of 11 conceptually diverse algorithms on a selection of 8 typologically distinct languages. The results consist evidence that some segmentation algorithms are cross-linguistically valid, thus could be considered as potential strategies employed by all infants.

On the Distribution of Deep Clausal Embeddings: A Large Cross-linguistic Study

- https://www.aclweb.org/anthology/P19-1384/

Embedding a clause inside another ("the girl [who likes cars [that run fast]] has arrived") is a fundamental resource that has been argued to be a key driver of linguistic expressiveness. As such, it plays a central role in fundamental debates on what makes human language unique, and how they might have evolved. Empirical evidence on the prevalence and the limits of embeddings has however been based on either laboratory setups or corpus data of relatively limited size. We introduce here a collection of large, dependency-parsed written corpora in 17 languages, that allow us, for the first time, to capture clausal embedding through dependency graphs and assess their distribution. Our results indicate that there is no evidence for hard constraints on embedding depth: the tail of depth distributions is heavy. Moreover, although deeply embedded clauses tend to be shorter, suggesting processing load issues, complex sentences with many embeddings do not display a bias towards less deep embeddings. Taken together, the results suggest that deep embeddings are not disfavoured in written language. More generally, our study illustrates how resources and methods from latest-generation big-data NLP can provide new perspectives on fundamental questions in theoretical linguistics.

Attention-based Conditioning Methods for External Knowledge Integration

- https://www.aclweb.org/anthology/P19-1385/

In this paper, we present a novel approach for incorporating external knowledge in Recurrent Neural Networks (RNNs). We propose the integration of lexicon features into the self-attention mechanism of RNN-based architectures. This form of conditioning on the attention distribution, enforces the contribution of the most salient words for the task at hand. We introduce three methods, namely attentional concatenation, feature-based gating and affine transformation. Experiments on six benchmark datasets show the effectiveness of our methods. Attentional feature-based gating yields consistent performance improvement across tasks. Our approach is implemented as a simple add-on module for RNN-based models with minimal computational overhead and can be adapted to any deep neural architecture.

The KnowRef Coreference Corpus: Removing Gender and Number Cues for Difficult Pronominal Anaphora Resolution

- https://www.aclweb.org/anthology/P19-1386/

We introduce a new benchmark for coreference resolution and NLI, KnowRef, that targets common-sense understanding and world knowledge. Previous coreference resolution tasks can largely be solved by exploiting the number and gender of the antecedents, or have been handcrafted and do not reflect the diversity of naturally occurring text. We present a corpus of over 8,000 annotated text passages with ambiguous pronominal anaphora. These instances are both challenging and realistic. We show that various coreference systems, whether rule-based, feature-rich, or neural, perform significantly worse on the task than humans, who display high inter-annotator agreement. To explain this performance gap, we show empirically that state-of-the art models often fail to capture context, instead relying on the gender or number of candidate antecedents to make a decision. We then use problem-specific insights to propose a data-augmentation trick called antecedent switching to alleviate this tendency in models. Finally, we show that antecedent switching yields promising results on other tasks as well: we use it to achieve state-of-the-art results on the GAP coreference task.

StRE: Self Attentive Edit Quality Prediction in Wikipedia

- https://www.aclweb.org/anthology/P19-1387/

Wikipedia can easily be justified as a behemoth, considering the sheer volume of content that is added or removed every minute to its several projects. This creates an immense scope, in the field of natural language processing toward developing automated tools for content moderation and review. In this paper we propose Self Attentive Revision Encoder (StRE) which leverages orthographic similarity of lexical units toward predicting the quality of new edits. In contrast to existing propositions which primarily employ features like page reputation, editor activity or rule based heuristics, we utilize the textual content of the edits which, we believe contains superior signatures of their quality. More specifically, we deploy deep encoders to generate representations of the edits from its text content, which we then leverage to infer quality. We further contribute a novel dataset containing 21M revisions across 32K Wikipedia pages and demonstrate that StRE outperforms existing methods by a significant margin – at least 17% and at most 103%. Our pre-trained model achieves such result after retraining on a set as small as 20% of the edits in a wikipage. This, to the best of our knowledge, is also the first attempt towards employing deep language models to the enormous domain of automated content moderation and review in Wikipedia.

How Large Are Lions? Inducing Distributions over Quantitative Attributes

- https://www.aclweb.org/anthology/P19-1388/

Most current NLP systems have little knowledge about quantitative attributes of objects and events. We propose an unsupervised method for collecting quantitative information from large amounts of web data, and use it to create a new, very large resource consisting of distributions over physical quantities associated with objects, adjectives, and verbs which we call Distributions over Quantitative (DoQ). This contrasts with recent work in this area which has focused on making only relative comparisons such as "Is a lion bigger than a wolf?". Our evaluation shows that DoQ compares favorably with state of the art results on existing datasets for relative comparisons of nouns and adjectives, and on a new dataset we introduce.

Fine-Grained Sentence Functions for Short-Text Conversation

- https://www.aclweb.org/anthology/P19-1389/

Sentence function is an important linguistic feature referring to a user's purpose in uttering a specific sentence. The use of sentence function has shown promising results to improve the performance of conversation models. However, there is no large conversation dataset annotated with sentence functions. In this work, we collect a new Short-Text Conversation dataset with manually annotated SEntence FUNctions (STC-Sefun). Classification models are trained on this dataset to (i) recognize the sentence function of new data in a large corpus of short-text conversations; (ii) estimate a proper sentence function of the response given a test query. We later train conversation models conditioned on the sentence functions, including information retrieval-based and neural generative models. Experimental results demonstrate that the use of sentence functions can help improve the quality of the returned responses.

Give Me More Feedback II: Annotating Thesis Strength and Related Attributes in Student Essays

- https://www.aclweb.org/anthology/P19-1390/

While the vast majority of existing work on automated essay scoring has focused on holistic scoring, researchers have recently begun work on scoring specific dimensions of essay quality. Nevertheless, progress on dimension-specific essay scoring is limited in part by the lack of annotated corpora. To facilitate advances in this area, we design a scoring rubric for scoring a core, yet unexplored dimension of persuasive essay quality, thesis strength, and annotate a corpus of

essays with thesis strength scores. We additionally identify the attributes that could impact thesis strength and annotate the essays with the values of these attributes, which, when predicted by computational models, could provide further feedback to students on why her essay receives a particular thesis strength score.

Crowdsourcing and Validating Event-focused Emotion Corpora for German and English

- https://www.aclweb.org/anthology/P19-1391/

Sentiment analysis has a range of corpora available across multiple languages. For emotion analysis, the situation is more limited, which hinders potential research on crosslingual modeling and the development of predictive models for other languages. In this paper, we fill this gap for German by constructing deISEAR, a corpus designed in analogy to the well-established English ISEAR emotion dataset. Motivated by Scherer's appraisal theory, we implement a crowdsourcing experiment which consists of two steps. In step 1, participants create descriptions of emotional events for a given emotion. In step 2, five annotators assess the emotion expressed by the texts. We show that transferring an emotion classification model from the original English ISEAR to the German crowdsourced deISEAR via machine translation does not, on average, cause a performance drop.

Pay Attention when you Pay the Bills. A Multilingual Corpus with Dependency-based and Semantic Annotation of Collocations.

- https://www.aclweb.org/anthology/P19-1392/

This paper presents a new multilingual corpus with semantic annotation of collocations in English, Portuguese, and Spanish. The whole resource contains 155k tokens and 1,526 collocations labeled in context. The annotated examples belong to three syntactic relations (adjective-noun, verb-object, and nominal compounds), and represent 58 lexical functions in the Meaning-Text Theory (e.g., Oper, Magn, Bon, etc.). Each collocation was annotated by three linguists and the final resource was revised by a team of experts. The resulting corpus can serve as a basis to evaluate different approaches for collocation identification, which in turn can be useful for different NLP tasks such as natural language understanding or natural language generation.

Does it Make Sense? And Why? A Pilot Study for Sense Making and Explanation

- https://www.aclweb.org/anthology/P19-1393/

Introducing common sense to natural language understanding systems has received increasing research attention. It remains a fundamental question on how to evaluate whether a system has the sense-making capability. Existing benchmarks measure common sense knowledge indirectly or without reasoning. In this paper, we release a benchmark to directly test whether a system can differentiate natural language statements that make sense from those that do not make sense. In addition, a system is asked to identify the most crucial reason why a statement does not make sense. We evaluate models trained over large-scale language modeling tasks as well as human performance, showing that there are different challenges for system sense-making.

Large Dataset and Language Model Fun-Tuning for Humor Recognition

- https://www.aclweb.org/anthology/P19-1394/

The task of humor recognition has attracted a lot of attention recently due to the urge to process large amounts of user-generated texts and rise of conversational agents. We collected a dataset of jokes and funny dialogues in Russian from various online resources and complemented them carefully with unfunny texts with similar lexical properties. The dataset comprises of more than 300,000 short texts, which is significantly larger than any previous humor-related corpus. Manual annotation of 2,000 items proved the reliability of the corpus construction approach. Further, we applied language model fine-tuning for text classification and obtained an F1 score of 0.91 on a test set, which constitutes a considerable gain over baseline methods. The dataset is freely available for research community.

Towards Language Agnostic Universal Representations

- https://www.aclweb.org/anthology/P19-1395/

When a bilingual student learns to solve word problems in math, we expect the student to be able to solve these problem in both languages the student is fluent in, even if the math lessons were only taught in one language. However, current representations in machine learning are language dependent. In this work, we present a method to decouple the language from the problem by learning language agnostic representations and therefore allowing training a model in one language and applying to a different one in a zero shot fashion. We learn

these representations by taking inspiration from linguistics, specifically the Universal Grammar hypothesis and learn universal latent representations that are language agnostic. We demonstrate the capabilities of these representations by showing that models trained on a single language using language agnostic representations achieve very similar accuracies in other languages.

Leveraging Meta Information in Short Text Aggregation

- https://www.aclweb.org/anthology/P19-1396/

Short texts such as tweets often contain insufficient word co-occurrence information for training conventional topic models. To deal with the insufficiency, we propose a generative model that aggregates short texts into clusters by leveraging the associated meta information. Our model can generate more interpretable topics as well as document clusters. We develop an effective Gibbs sampling algorithm favoured by the fully local conjugacy in the model. Extensive experiments demonstrate that our model achieves better performance in terms of document clustering and topic coherence.

Exploiting Invertible Decoders for Unsupervised Sentence Representation Learning

- https://www.aclweb.org/anthology/P19-1397/

Encoder-decoder models for unsupervised sentence representation learning using the distributional hypothesis effectively constrain the learnt representation of a sentence to only that needed to reproduce the next sentence. While the decoder is important to constrain the representation, these models tend to discard the decoder after training since only the encoder is needed to map the input sentence into a vector representation. However, parameters learnt in the decoder also contain useful information about the language. In order to utilise the decoder after learning, we present two types of decoding functions whose inverse can be easily derived without expensive inverse calculation. Therefore, the inverse of the decoding function serves as another encoder that produces sentence representations. We show that, with careful design of the decoding functions, the model learns good sentence representations, and the ensemble of the representations produced from the encoder and the inverse of the decoder demonstrate even better generalisation ability and solid transferability.

Self-Attentive, Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text

- https://www.aclweb.org/anthology/P19-1398/

There exist few text-specific methods for unsupervised anomaly detection, and for those that do exist, none utilize pre-trained models for distributed vector representations of words. In this paper we introduce a new anomaly detection method—Context Vector Data Description (CVDD)—which builds upon word embedding models to learn multiple sentence representations that capture multiple semantic contexts via the self-attention mechanism. Modeling multiple contexts enables us to perform contextual anomaly detection of sentences and phrases with respect to the multiple themes and concepts present in an unlabeled text corpus. These contexts in combination with the self-attention weights make our method highly interpretable. We demonstrate the effectiveness of CVDD quantitatively as well as qualitatively on the well-known Reuters, 20 Newsgroups, and IMDB Movie Reviews datasets.

Hubless Nearest Neighbor Search for Bilingual Lexicon Induction

- https://www.aclweb.org/anthology/P19-1399/

Bilingual Lexicon Induction (BLI) is the task of translating words from corpora in two languages. Recent advances in BLI work by aligning the two word embedding spaces. Following that, a key step is to retrieve the nearest neighbor (NN) in the target space given the source word. However, a phenomenon called hubness often degrades the accuracy of NN. Hubness appears as some data points, called hubs, being extra-ordinarily close to many of the other data points. Reducing hubness is necessary for retrieval tasks. One successful example is Inverted SoFtmax (ISF), recently proposed to improve NN. This work proposes a new method, Hubless Nearest Neighbor (HNN), to mitigate hubness. HNN differs from NN by imposing an additional equal preference assumption. Moreover, the HNN formulation explains why ISF works as well as it does. Empirical results demonstrate that HNN outperforms NN, ISF and other state-of-the-art. For reproducibility and follow-ups, we have published all code.

Distant Learning for Entity Linking with Automatic Noise Detection

- https://www.aclweb.org/anthology/P19-1400/

Accurate entity linkers have been produced for domains and languages where annotated data (i.e., texts linked to a knowledge base) is available. However, little progress has been made for the settings where no or very limited amounts of labeled data are present (e.g., legal or most scientific domains). In this work, we show how we can learn to link mentions without having any labeled examples, only a knowledge base and a collection of unannotated texts from the corresponding domain. In order to achieve this, we frame the task as a multi-instance learning problem and rely on surface matching to create initial noisy

labels. As the learning signal is weak and our surrogate labels are noisy, we introduce a noise detection component in our model: it lets the model detect and disregard examples which are likely to be noisy. Our method, jointly learning to detect noise and link entities, greatly outperforms the surface matching baseline. For a subset of entity categories, it even approaches the performance of supervised learning.

Learning How to Active Learn by Dreaming

- https://www.aclweb.org/anthology/P19-1401/

Heuristic-based active learning (AL) methods are limited when the data distribution of the underlying learning problems vary. Recent data-driven AL policy learning methods are also restricted to learn from closely related domains. We introduce a new sample-efficient method that learns the AL policy directly on the target domain of interest by using wake and dream cycles. Our approach interleaves between querying the annotation of the selected datapoints to update the underlying student learner and improving AL policy using simulation where the current student learner acts as an imperfect annotator. We evaluate our method on cross-domain and cross-lingual text classification and named entity recognition tasks. Experimental results show that our dream-based AL policy training strategy is more effective than applying the pretrained policy without further fine-tuning and better than the existing strong baseline methods that use heuristics or reinforcement learning.

Few-Shot Representation Learning for Out-Of-Vocabulary Words

- https://www.aclweb.org/anthology/P19-1402/

Existing approaches for learning word embedding often assume there are sufficient occurrences for each word in the corpus, such that the representation of words can be accurately estimated from their contexts. However, in real-world scenarios, out-of-vocabulary (a.k.a. OOV) words that do not appear in training corpus emerge frequently. How to learn accurate representations of these words to augment a pre-trained embedding by only a few observations is a challenging research problem. In this paper, we formulate the learning of OOV embedding as a few-shot regression problem by fitting a representation function to predict an oracle embedding vector (defined as embedding trained with abundant observations) based on limited contexts. Specifically, we propose a novel hierarchical attention network-based embedding framework to serve as the neural regression function, in which the context information of a word is encoded and aggregated from K observations. Furthermore, we propose to use Model-Agnostic Meta-Learning (MAML) for adapting the learned model to the new corpus fast and

robustly. Experiments show that the proposed approach significantly outperforms existing methods in constructing an accurate embedding for OOV words and improves downstream tasks when the embedding is utilized.

Neural Temporality Adaptation for Document Classification: Diachronic Word Embeddings and Domain Adaptation Models

- https://www.aclweb.org/anthology/P19-1403/

Language usage can change across periods of time, but document classifiers models are usually trained and tested on corpora spanning multiple years without considering temporal variations. This paper describes two complementary ways to adapt classifiers to shifts across time. First, we show that diachronic word embeddings, which were originally developed to study language change, can also improve document classification, and we show a simple method for constructing this type of embedding. Second, we propose a time-driven neural classification model inspired by methods for domain adaptation. Experiments on six corpora show how these methods can make classifiers more robust over time.

Learning Transferable Feature Representations Using Neural Networks

- https://www.aclweb.org/anthology/P19-1404/

Learning representations such that the source and target distributions appear as similar as possible has benefited transfer learning tasks across several applications. Generally it requires labeled data from the source and only unlabeled data from the target to learn such representations. While these representations act like a bridge to transfer knowledge learned in the source to the target; they may lead to negative transfer when the source specific characteristics detract their ability to represent the target data. We present a novel neural network architecture to simultaneously learn a two-part representation which is based on the principle of segregating source specific representation from the common representation. The first part captures the source specific characteristics while the second part captures the truly common representation. Our architecture optimizes an objective function which acts adversarial for the source specific part if it contributes towards the cross-domain learning. We empirically show that two parts of the representation, in different arrangements, outperforms existing learning algorithms on the source learning as well as cross-domain tasks on multiple datasets.

Bayes Test of Precision, Recall, and F1 Measure for Comparison of Two Natural Language Processing Models

- https://www.aclweb.org/anthology/P19-1405/

Direct comparison on point estimation of the precision (P), recall (R), and F1 measure of two natural language processing (NLP) models on a common test corpus is unreasonable and results in less replicable conclusions due to a lack of a statistical test. However, the existing t-tests in cross-validation (CV) for model comparison are inappropriate because the distributions of P, R, F1 are skewed and an interval estimation of P, R, and F1 based on a t-test may exceed [0,1]. In this study, we propose to use a block-regularized 3×2 CV (3×2 BCV) in model comparison because it could regularize the difference in certain frequency distributions over linguistic units between training and validation sets and yield stable estimators of P, R, and F1. On the basis of the 3×2 BCV, we calibrate the posterior distributions of P, R, and F1 and derive an accurate interval estimation of P, R, and F1. Furthermore, we formulate the comparison into a hypothesis testing problem and propose a novel Bayes test. The test could directly compute the probabilities of the hypotheses on the basis of the posterior distributions and provide more informative decisions than the existing significance t-tests. Three experiments with regard to NLP chunking tasks are conducted, and the results illustrate the validity of the Bayes test.

TIGS: An Inference Algorithm for Text Infilling with Gradient Search

- https://www.aclweb.org/anthology/P19-1406/

Text infilling aims at filling in the missing part of a sentence or paragraph, which has been applied to a variety of real-world natural language generation scenarios. Given a well-trained sequential generative model, it is challenging for its unidirectional decoder to generate missing symbols conditioned on the past and future information around the missing part. In this paper, we propose an iterative inference algorithm based on gradient search, which could be the first inference algorithm that can be broadly applied to any neural sequence generative models for text infilling tasks. Extensive experimental comparisons show the effectiveness and efficiency of the proposed method on three different text infilling tasks with various mask ratios and different mask strategies, comparing with five state-of-the-art methods.

Keeping Notes: Conditional Natural Language Generation with a Scratchpad Encoder

- https://www.aclweb.org/anthology/P19-1407/

We introduce the Scratchpad Mechanism, a novel addition to the sequence-to-sequence (seq2seq) neural network architecture and demonstrate its effectiveness in improving the overall fluency of seq2seq models for natural language generation tasks. By enabling the decoder at each time step to write to all of the encoder output layers, Scratchpad can employ the encoder as a "scratchpad" memory to keep track of what has been generated so far and thereby guide future generation. We evaluate Scratchpad in the context of three well-studied natural language generation tasks — Machine Translation, Question Generation, and Text Summarization — and obtain state-of-the-art or comparable performance on standard datasets for each task. Qualitative assessments in the form of human judgements (question generation), attention visualization (MT), and sample output (summarization) provide further evidence of the ability of Scratchpad to generate fluent and expressive output.

Using Automatically Extracted Minimum Spans to Disentangle Coreference Evaluation from Boundary Detection

- https://www.aclweb.org/anthology/P19-1408/

The common practice in coreference resolution is to identify and evaluate the maximum span of mentions. The use of maximum spans tangles coreference evaluation with the challenges of mention boundary detection like prepositional phrase attachment. To address this problem, minimum spans are manually annotated in smaller corpora. However, this additional annotation is costly and therefore, this solution does not scale to large corpora. In this paper, we propose the MINA algorithm for automatically extracting minimum spans to benefit from minimum span evaluation in all corpora. We show that the extracted minimum spans by MINA are consistent with those that are manually annotated by experts. Our experiments show that using minimum spans is in particular important in cross-dataset coreference evaluation, in which detected mention boundaries are noisier due to domain shift. We have integrated MINA into https://github.com/ns-moosavi/coval for reporting standard coreference scores based on both maximum and automatically detected minimum spans.

Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution

- https://www.aclweb.org/anthology/P19-1409/

Recognizing coreferring events and entities across multiple texts is crucial for many NLP applications. Despite the task's importance, research focus was given mostly to within-document entity coreference, with rather little attention to the other variants. We propose a neural architecture for cross-document coreference

resolution. Inspired by Lee et al. (2012), we jointly model entity and event coreference. We represent an event (entity) mention using its lexical span, surrounding context, and relation to entity (event) mentions via predicate-arguments structures. Our model outperforms the previous state-of-the-art event coreference model on ECB+, while providing the first entity coreference results on this corpus. Our analysis confirms that all our representation elements, including the mention span itself, its context, and the relation to other mentions contribute to the model's success.

A Unified Linear-Time Framework for Sentence-Level Discourse Parsing

- https://www.aclweb.org/anthology/P19-1410/

We propose an efficient neural framework for sentence-level discourse analysis in accordance with Rhetorical Structure Theory (RST). Our framework comprises a discourse segmenter to identify the elementary discourse units (EDU) in a text, and a discourse parser that constructs a discourse tree in a top-down fashion. Both the segmenter and the parser are based on Pointer Networks and operate in linear time. Our segmenter yields an F1 score of 95.4%, and our parser achieves an F1 score of 81.7% on the aggregated labeled (relation) metric, surpassing previous approaches by a good margin and approaching human agreement on both tasks (98.3 and 83.0 F1).

Employing the Correspondence of Relations and Connectives to Identify Implicit Discourse Relations via Label Embeddings

- https://www.aclweb.org/anthology/P19-1411/

It has been shown that implicit connectives can be exploited to improve the performance of the models for implicit discourse relation recognition (IDRR). An important property of the implicit connectives is that they can be accurately mapped into the discourse relations conveying their functions. In this work, we explore this property in a multi-task learning framework for IDRR in which the relations and the connectives are simultaneously predicted, and the mapping is leveraged to transfer knowledge between the two prediction tasks via the embeddings of relations and connectives. We propose several techniques to enable such knowledge transfer that yield the state-of-the-art performance for IDRR on several settings of the benchmark dataset (i.e., the Penn Discourse Treebank dataset).

Do You Know That Florence Is Packed with Visitors? Evaluating State-of-the-art Models of Speaker Commitment

- https://www.aclweb.org/anthology/P19-1412/

When a speaker, Mary, asks "Do you know that Florence is packed with visitors?", we take her to believe that Florence is packed with visitors, but not if she asks "Do you think that Florence is packed with visitors?". Inferring speaker commitment (aka event factuality) is crucial for information extraction and question answering. Here, we explore the hypothesis that linguistic deficits drive the error patterns of existing speaker commitment models by analyzing the linguistic correlates of model error on a challenging naturalistic dataset. We evaluate two state-of-the-art speaker commitment models on the CommitmentBank, an English dataset of naturally occurring discourses. The CommitmentBank is annotated with speaker commitment towards the content of the complement ("Florence is packed with visitors" in our example) of clause-embedding verbs ("know", "think") under four entailment-canceling environments (negation, modal, question, conditional). A breakdown of items by linguistic features reveals asymmetrical error patterns: while the models achieve good performance on some classes (e.g., negation), they fail to generalize to the diverse linguistic constructions (e.g., conditionals) in natural language, highlighting directions for improvement.

Multi-Relational Script Learning for Discourse Relations

- https://www.aclweb.org/anthology/P19-1413/

Modeling script knowledge can be useful for a wide range of NLP tasks. Current statistical script learning approaches embed the events, such that their relationships are indicated by their similarity in the embedding. While intuitive, these approaches fall short of representing nuanced relations, needed for downstream tasks. In this paper, we suggest to view learning event embedding as a multi-relational problem, which allows us to capture different aspects of event pairs. We model a rich set of event relations, such as Cause and Contrast, derived from the Penn Discourse Tree Bank. We evaluate our model on three types of tasks, the popular Mutli-Choice Narrative Cloze and its variants, several multi-relational prediction tasks, and a related downstream task—implicit discourse sense classification.

Open-Domain Why-Question Answering with Adversarial Learning to Encode Answer Texts

- https://www.aclweb.org/anthology/P19-1414/

In this paper, we propose a method for why-question answering (why-QA) that uses an adversarial learning framework. Existing why-QA methods retrieve "answer passages" that usually consist of several sentences. These multi-sentence passages contain not only the reason sought by a why-question and its connection to the why-question, but also redundant and/or unrelated parts. We use our proposed "Adversarial networks for Generating compact-answer Representation" (AGR) to generate from a passage a vector representation of the non-redundant reason sought by a why-question and exploit the representation for judging whether the passage actually answers the why-question. Through a series of experiments using Japanese why-QA datasets, we show that these representations improve the performance of our why-QA neural model as well as that of a BERT-based why-QA model. We show that they also improve a state-of-the-art distantly supervised open-domain QA (DS-QA) method on publicly available English datasets, even though the target task is not a why-QA.

Learning to Ask Unanswerable Questions for Machine Reading Comprehension

- https://www.aclweb.org/anthology/P19-1415/

Machine reading comprehension with unanswerable questions is a challenging task. In this work, we propose a data augmentation technique by automatically generating relevant unanswerable questions according to an answerable question paired with its corresponding paragraph that contains the answer. We introduce a pair-to-sequence model for unanswerable question generation, which effectively captures the interactions between the question and the paragraph. We also present a way to construct training data for our question generation models by leveraging the existing reading comprehension dataset. Experimental results show that the pair-to-sequence model performs consistently better compared with the sequence-to-sequence baseline. We further use the automatically generated unanswerable questions as a means of data augmentation on the SQuAD 2.0 dataset, yielding 1.9 absolute F1 improvement with BERT-base model and 1.7 absolute F1 improvement with BERT-large model.

Compositional Questions Do Not Necessitate Multi-hop Reasoning

- https://www.aclweb.org/anthology/P19-1416/

Multi-hop reading comprehension (RC) questions are challenging because they require reading and reasoning over multiple paragraphs. We argue that it can be difficult to construct large multi-hop RC datasets. For example, even highly compositional questions can be answered with a single hop if they target specific entity types, or the facts needed to answer them are redundant. Our analysis

is centered on HotpotQA, where we show that single-hop reasoning can solve much more of the dataset than previously thought. We introduce a single-hop BERT-based RC model that achieves 67 F1—comparable to state-of-the-art multi-hop models. We also design an evaluation setting where humans are not shown all of the necessary paragraphs for the intended multi-hop reasoning but can still answer over 80% of questions. Together with detailed error analysis, these results suggest there should be an increasing focus on the role of evidence in multi-hop reasoning and possibly even a shift towards information retrieval style evaluations with large and diverse evidence collections.

Improving Question Answering over Incomplete KBs with Knowledge-Aware Reader

- https://www.aclweb.org/anthology/P19-1417/

We propose a new end-to-end question answering model, which learns to aggregate answer evidence from an incomplete knowledge base (KB) and a set of retrieved text snippets.Under the assumptions that structured data is easier to query and the acquired knowledge can help the understanding of unstructured text, our model first accumulates knowledge ofKB entities from a question-related KB sub-graph; then reformulates the question in the latent space and reads the text with the accumulated entity knowledge at hand. The evidence from KB and text are finally aggregated to predict answers. On the widely-used KBQA benchmark WebQSP, our model achieves consistent improvements across settings with different extents of KB incompleteness.

AdaNSP: Uncertainty-driven Adaptive Decoding in Neural Semantic Parsing

- https://www.aclweb.org/anthology/P19-1418/

Neural semantic parsers utilize the encoder-decoder framework to learn an end-to-end model for semantic parsing that transduces a natural language sentence to the formal semantic representation. To keep the model aware of the underlying grammar in target sequences, many constrained decoders were devised in a multi-stage paradigm, which decode to the sketches or syntax trees first, and then decode to target semantic tokens. We instead to propose an adaptive decoding method to avoid such intermediate representations. The decoder is guided by model uncertainty and automatically uses deeper computations when necessary. Thus it can predict tokens adaptively. Our model outperforms the state-of-the-art neural models and does not need any expertise like predefined grammar or sketches in the meantime.

The Language of Legal and Illegal Activity on the Darknet

- https://www.aclweb.org/anthology/P19-1419/

The non-indexed parts of the Internet (the Darknet) have become a haven for both legal and illegal anonymous activity. Given the magnitude of these networks, scalably monitoring their activity necessarily relies on automated tools, and notably on NLP tools. However, little is known about what characteristics texts communicated through the Darknet have, and how well do off-the-shelf NLP tools do on this domain. This paper tackles this gap and performs an in-depth investigation of the characteristics of legal and illegal text in the Darknet, comparing it to a clear net website with similar content as a control condition. Taking drugs-related websites as a test case, we find that texts for selling legal and illegal drugs have several linguistic characteristics that distinguish them from one another, as well as from the control condition, among them the distribution of POS tags, and the coverage of their named entities in Wikipedia.

Eliciting Knowledge from Experts: Automatic Transcript Parsing for Cognitive Task Analysis

- https://www.aclweb.org/anthology/P19-1420/

Cognitive task analysis (CTA) is a type of analysis in applied psychology aimed at eliciting and representing the knowledge and thought processes of domain experts. In CTA, often heavy human labor is involved to parse the interview transcript into structured knowledge (e.g., flowchart for different actions). To reduce human efforts and scale the process, automated CTA transcript parsing is desirable. However, this task has unique challenges as (1) it requires the understanding of long-range context information in conversational text; and (2) the amount of labeled data is limited and indirect—i.e., context-aware, noisy, and low-resource. In this paper, we propose a weakly-supervised information extraction framework for automated CTA transcript parsing. We partition the parsing process into a sequence labeling task and a text span-pair relation extraction task, with distant supervision from human-curated protocol files. To model long-range context information for extracting sentence relations, neighbor sentences are involved as a part of input. Different types of models for capturing context dependency are then applied. We manually annotate real-world CTA transcripts to facilitate the evaluation of the parsing tasks.

Course Concept Expansion in MOOCs with External Knowledge and Interactive Game

- https://www.aclweb.org/anthology/P19-1421/

As Massive Open Online Courses (MOOCs) become increasingly popular, it is promising to automatically provide extracurricular knowledge for MOOC users. Suffering from semantic drifts and lack of knowledge guidance, existing methods can not effectively expand course concepts in complex MOOC environments. In this paper, we first build a novel boundary during searching for new concepts via external knowledge base and then utilize heterogeneous features to verify the high-quality results. In addition, to involve human efforts in our model, we design an interactive optimization mechanism based on a game. Our experiments on the four datasets from Coursera and XuetangX show that the proposed method achieves significant improvements(+0.19 by MAP) over existing methods.

Towards Near-imperceptible Steganographic Text

- https://www.aclweb.org/anthology/P19-1422/

We show that the imperceptibility of several existing linguistic steganographic systems (Fang et al., 2017; Yang et al., 2018) relies on implicit assumptions on statistical behaviors of fluent text. We formally analyze them and empirically evaluate these assumptions. Furthermore, based on these observations, we propose an encoding algorithm called patient-Huffman with improved near-imperceptible guarantees.

Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network

- https://www.aclweb.org/anthology/P19-1423/

Inter-sentence relation extraction deals with a number of complex semantic relationships in documents, which require local, non-local, syntactic and semantic dependencies. Existing methods do not fully exploit such dependencies. We present a novel inter-sentence relation extraction model that builds a labelled edge graph convolutional neural network model on a document-level graph. The graph is constructed using various inter- and intra-sentence dependencies to capture local and non-local dependency information. In order to predict the relation of an entity pair, we utilise multi-instance learning with bi-affine pairwise scoring. Experimental results show that our model achieves comparable performance to the state-of-the-art neural models on two biochemistry datasets. Our analysis shows that all the types in the graph are effective for inter-sentence relation extraction.

Neural Legal Judgment Prediction in English

- https://www.aclweb.org/anthology/P19-1424/

Legal judgment prediction is the task of automatically predicting the outcome of a court case, given a text describing the case's facts. Previous work on using neural models for this task has focused on Chinese; only feature-based models (e.g., using bags of words and topics) have been considered in English. We release a new English legal judgment prediction dataset, containing cases from the European Court of Human Rights. We evaluate a broad variety of neural models on the new dataset, establishing strong baselines that surpass previous feature-based models in three tasks: (1) binary violation classification; (2) multi-label classification; (3) case importance prediction. We also explore if models are biased towards demographic information via data anonymization. As a side-product, we propose a hierarchical version of BERT, which bypasses BERT's length limitation.

Robust Neural Machine Translation with Doubly Adversarial Inputs

- https://www.aclweb.org/anthology/P19-1425/

Neural machine translation (NMT) often suffers from the vulnerability to noisy perturbations in the input. We propose an approach to improving the robustness of NMT models, which consists of two parts: (1) attack the translation model with adversarial source examples; (2) defend the translation model with adversarial target inputs to improve its robustness against the adversarial source inputs. For the generation of adversarial inputs, we propose a gradient-based method to craft adversarial examples informed by the translation loss over the clean inputs. Experimental results on Chinese-English and English-German translation tasks demonstrate that our approach achieves significant improvements (2.8 and 1.6 BLEU points) over Transformer on standard clean benchmarks as well as exhibiting higher robustness on noisy data.

Bridging the Gap between Training and Inference for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1426/

Neural Machine Translation (NMT) generates target words sequentially in the way of predicting the next word conditioned on the context words. At training time, it predicts with the ground truth words as context while at inference it has to generate the entire sequence from scratch. This discrepancy of the fed context leads to error accumulation among the way. Furthermore, word-level training requires strict matching between the generated sequence and the ground truth sequence which leads to overcorrection over different but reasonable translations. In this paper, we address these issues by sampling context words not only from the ground truth sequence but also from the predicted sequence by the model during training, where the predicted sequence is selected

with a sentence-level optimum. Experiment results on Chinese->English and WMT'14 English->German translation tasks demonstrate that our approach can achieve significant improvements on multiple datasets.

Beyond BLEU:Training Neural Machine Translation with Semantic Similarity

- https://www.aclweb.org/anthology/P19-1427/

While most neural machine translation (NMT)systems are still trained using maximum likelihood estimation, recent work has demonstrated that optimizing systems to directly improve evaluation metrics such as BLEU can significantly improve final translation accuracy. However, training with BLEU has some limitations: it doesn't assign partial credit, it has a limited range of output values, and it can penalize semantically correct hypotheses if they differ lexically from the reference. In this paper, we introduce an alternative reward function for optimizing NMT systems that is based on recent work in semantic similarity. We evaluate on four disparate languages trans-lated to English, and find that training with our proposed metric results in better translations as evaluated by BLEU, semantic similarity, and human evaluation, and also that the optimization procedure converges faster. Analysis suggests that this is because the proposed metric is more conducive to optimization, assigning partial credit and providing more diversity in scores than BLEU

AutoML Strategy Based on Grammatical Evolution: A Case Study about Knowledge Discovery from Text

- https://www.aclweb.org/anthology/P19-1428/

The process of extracting knowledge from natural language text poses a complex problem that requires both a combination of machine learning techniques and proper feature selection. Recent advances in Automatic Machine Learning (AutoML) provide effective tools to explore large sets of algorithms, hyperparameters and features to find out the most suitable combination of them. This paper proposes a novel AutoML strategy based on probabilistic grammatical evolution, which is evaluated on the health domain by facing the knowledge discovery challenge in Spanish text documents. Our approach achieves state-of-the-art results and provides interesting insights into the best combination of parameters and algorithms to use when dealing with this challenge. Source code is provided for the research community.

Distilling Discrimination and Generalization Knowledge for Event Detection via Delta-Representation Learning

- https://www.aclweb.org/anthology/P19-1429/

Event detection systems rely on discrimination knowledge to distinguish ambiguous trigger words and generalization knowledge to detect unseen/sparse trigger words. Current neural event detection approaches focus on trigger-centric representations, which work well on distilling discrimination knowledge, but poorly on learning generalization knowledge. To address this problem, this paper proposes a Delta-learning approach to distill discrimination and generalization knowledge by effectively decoupling, incrementally learning and adaptively fusing event representation. Experiments show that our method significantly outperforms previous approaches on unseen/sparse trigger words, and achieves state-of-the-art performance on both ACE2005 and KBP2017 datasets.

Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge

- https://www.aclweb.org/anthology/P19-1430/

Chinese relation extraction is conducted using neural networks with either character-based or word-based inputs, and most existing methods typically suffer from segmentation errors and ambiguity of polysemy. To address the issues, we propose a multi-grained lattice framework (MG lattice) for Chinese relation extraction to take advantage of multi-grained language information and external linguistic knowledge. In this framework, (1) we incorporate word-level information into character sequence inputs so that segmentation errors can be avoided. (2) We also model multiple senses of polysemous words with the help of external linguistic knowledge, so as to alleviate polysemy ambiguity. Experiments on three real-world datasets in distinct domains show consistent and significant superiority and robustness of our model, as compared with other baselines. We will release the source code of this paper in the future.

A2N: Attending to Neighbors for Knowledge Graph Inference

- https://www.aclweb.org/anthology/P19-1431/

State-of-the-art models for knowledge graph completion aim at learning a fixed embedding representation of entities in a multi-relational graph which can generalize to infer unseen entity relationships at test time. This can be sub-optimal as it requires memorizing and generalizing to all possible entity relationships using these fixed representations. We thus propose a novel attention-based method to learn query-dependent representation of entities which adaptively combines

the relevant graph neighborhood of an entity leading to more accurate KG completion. The proposed method is evaluated on two benchmark datasets for knowledge graph completion, and experimental results show that the proposed model performs competitively or better than existing state-of-the-art, including recent methods for explicit multi-hop reasoning. Qualitative probing offers insight into how the model can reason about facts involving multiple hops in the knowledge graph, through the use of neighborhood attention.

Graph based Neural Networks for Event Factuality Prediction using Syntactic and Semantic Structures

- https://www.aclweb.org/anthology/P19-1432/

Event factuality prediction (EFP) is the task of assessing the degree to which an event mentioned in a sentence has happened. For this task, both syntactic and semantic information are crucial to identify the important context words. The previous work for EFP has only combined these information in a simple way that cannot fully exploit their coordination. In this work, we introduce a novel graph-based neural network for EFP that can integrate the semantic and syntactic information more effectively. Our experiments demonstrate the advantage of the proposed model for EFP.

Embedding Time Expressions for Deep Temporal Ordering Models

- https://www.aclweb.org/anthology/P19-1433/

Data-driven models have demonstrated state-of-the-art performance in inferring the temporal ordering of events in text. However, these models often overlook explicit temporal signals, such as dates and time windows. Rule-based methods can be used to identify the temporal links between these time expressions (timexes), but they fail to capture timexes' interactions with events and are hard to integrate with the distributed representations of neural net models. In this paper, we introduce a framework to infuse temporal awareness into such models by learning a pre-trained model to embed timexes. We generate synthetic data consisting of pairs of timexes, then train a character LSTM to learn embeddings and classify the timexes' temporal relation. We evaluate the utility of these embeddings in the context of a strong neural model for event temporal ordering, and show a small increase in performance on the MATRES dataset and more substantial gains on an automatically collected dataset with more frequent event-timex interactions.

Episodic Memory Reader: Learning What to Remember for Question Answering from Streaming Data

- https://www.aclweb.org/anthology/P19-1434/

We consider a novel question answering (QA) task where the machine needs to read from large streaming data (long documents or videos) without knowing when the questions will be given, which is difficult to solve with existing QA methods due to their lack of scalability. To tackle this problem, we propose a novel end-to-end deep network model for reading comprehension, which we refer to as Episodic Memory Reader (EMR) that sequentially reads the input contexts into an external memory, while replacing memories that are less important for answering unseen questions. Specifically, we train an RL agent to replace a memory entry when the memory is full, in order to maximize its QA accuracy at a future timepoint, while encoding the external memory using either the GRU or the Transformer architecture to learn representations that considers relative importance between the memory entries. We validate our model on a synthetic dataset (bAbI) as well as real-world large-scale textual QA (TriviaQA) and video QA (TVQA) datasets, on which it achieves significant improvements over rule based memory scheduling policies or an RL based baseline that independently learns the query-specific importance of each memory.

Selection Bias Explorations and Debias Methods for Natural Language Sentence Matching Datasets

- https://www.aclweb.org/anthology/P19-1435/

Natural Language Sentence Matching (NLSM) has gained substantial attention from both academics and the industry, and rich public datasets contribute a lot to this process. However, biased datasets can also hurt the generalization performance of trained models and give untrustworthy evaluation results. For many NLSM datasets, the providers select some pairs of sentences into the datasets, and this sampling procedure can easily bring unintended pattern, i.e., selection bias. One example is the QuoraQP dataset, where some content-independent naive features are unreasonably predictive. Such features are the reflection of the selection bias and termed as the "leakage features." In this paper, we investigate the problem of selection bias on six NLSM datasets and find that four out of them are significantly biased. We further propose a training and evaluation framework to alleviate the bias. Experimental results on QuoraQP suggest that the proposed framework can improve the generalization ability of trained models, and give more trustworthy evaluation results for real-world adoptions.

Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index

- https://www.aclweb.org/anthology/P19-1436/

Existing open-domain question answering (QA) models are not suitable for real-time usage because they need to process several long documents on-demand for every input query, which is computationally prohibitive. In this paper, we introduce query-agnostic indexable representations of document phrases that can drastically speed up open-domain QA. In particular, our dense-sparse phrase encoding effectively captures syntactic, semantic, and lexical information of the phrases and eliminates the pipeline filtering of context documents. Leveraging strategies for optimizing training and inference time, our model can be trained and deployed even in a single 4-GPU server. Moreover, by representing phrases as pointers to their start and end tokens, our model indexes phrases in the entire English Wikipedia (up to 60 billion phrases) using under 2TB. Our experiments on SQuAD-Open show that our model is on par with or more accurate than previous models with 6000x reduced computational cost, which translates into at least 68x faster end-to-end inference benchmark on CPUs. Code and demo are available at nlp.cs.washington.edu/denspi

Language Modeling with Shared Grammar

- https://www.aclweb.org/anthology/P19-1437/

Sequential recurrent neural networks have achieved superior performance on language modeling, but overlook the structure information in natural language. Recent works on structure-aware models have shown promising results on language modeling. However, how to incorporate structure knowledge on corpus without syntactic annotations remains an open problem. In this work, we propose neural variational language model (NVLM), which enables the sharing of grammar knowledge among different corpora. Experimental results demonstrate the effectiveness of our framework on two popular benchmark datasets. With the help of shared grammar, our language model converges significantly faster to a lower perplexity on new training corpus.

Zero-Shot Semantic Parsing for Instructions

- https://www.aclweb.org/anthology/P19-1438/

We consider a zero-shot semantic parsing task: parsing instructions into compositional logical forms, in domains that were not seen during training. We present a new dataset with 1,390 examples from 7 application domains (e.g. a calendar or a file manager), each example consisting of a triplet: (a) the application's initial state, (b) an instruction, to be carried out in the context of that state,

and (c) the state of the application after carrying out the instruction. We introduce a new training algorithm that aims to train a semantic parser on examples from a set of source domains, so that it can effectively parse instructions from an unknown target domain. We integrate our algorithm into the floating parser of Pasupat and Liang (2015), and further augment the parser with features and a logical form candidate filtering logic, to support zero-shot adaptation. Our experiments with various zero-shot adaptation setups demonstrate substantial performance gains over a non-adapted parser.

Can You Tell Me How to Get Past Sesame Street? Sentence-Level Pretraining Beyond Language Modeling

- https://www.aclweb.org/anthology/P19-1439/

Natural language understanding has recently seen a surge of progress with the use of sentence encoders like ELMo (Peters et al., 2018a) and BERT (Devlin et al., 2019) which are pretrained on variants of language modeling. We conduct the first large-scale systematic study of candidate pretraining tasks, comparing 19 different tasks both as alternatives and complements to language modeling. Our primary results support the use language modeling, especially when combined with pretraining on additional labeled-data tasks. However, our results are mixed across pretraining tasks and show some concerning trends: In ELMo's pretrain-then-freeze paradigm, random baselines are worryingly strong and results vary strikingly across target tasks. In addition, fine-tuning BERT on an intermediate task often negatively impacts downstream transfer. In a more positive trend, we see modest gains from multitask training, suggesting the development of more sophisticated multitask and transfer learning techniques as an avenue for further research.

Complex Question Decomposition for Semantic Parsing

- https://www.aclweb.org/anthology/P19-1440/

In this work, we focus on complex question semantic parsing and propose a novel Hierarchical Semantic Parsing (HSP) method, which utilizes the decompositionality of complex questions for semantic parsing. Our model is designed within a three-stage parsing architecture based on the idea of decomposition-integration. In the first stage, we propose a question decomposer which decomposes a complex question into a sequence of sub-questions. In the second stage, we design an information extractor to derive the type and predicate information of these questions. In the last stage, we integrate the generated information from previous stages and generate a logical form for the complex question. We conduct experiments on COMPLEXWEBQUESTIONS which is a large scale complex question semantic parsing dataset, results show that our model achieves significant improvement compared to state-of-the-art methods.

Multi-Task Deep Neural Networks for Natural Language Understanding

- https://www.aclweb.org/anthology/P19-1441/

In this paper, we present a Multi-Task Deep Neural Network (MT-DNN) for learning representations across multiple natural language understanding (NLU) tasks. MT-DNN not only leverages large amounts of cross-task data, but also benefits from a regularization effect that leads to more general representations to help adapt to new tasks and domains. MT-DNN extends the model proposed in Liu et al. (2015) by incorporating a pre-trained bidirectional transformer language model, known as BERT (Devlin et al., 2018). MT-DNN obtains new state-of-the-art results on ten NLU tasks, including SNLI, SciTail, and eight out of nine GLUE tasks, pushing the GLUE benchmark to 82.7% (2.2% absolute improvement) as of February 25, 2019 on the latest GLUE test set. We also demonstrate using the SNLI and SciTail datasets that the representations learned by MT-DNN allow domain adaptation with substantially fewer in-domain labels than the pre-trained BERT representations. Our code and pre-trained models will be made publicly available.

DisSent: Learning Sentence Representations from Explicit Discourse Relations

- https://www.aclweb.org/anthology/P19-1442/

Learning effective representations of sentences is one of the core missions of natural language understanding. Existing models either train on a vast amount of text, or require costly, manually curated sentence relation datasets. We show that with dependency parsing and rule-based rubrics, we can curate a high quality sentence relation task by leveraging explicit discourse relations. We show that our curated dataset provides an excellent signal for learning vector representations of sentence meaning, representing relations that can only be determined when the meanings of two sentences are combined. We demonstrate that the automatically curated corpus allows a bidirectional LSTM sentence encoder to yield high quality sentence embeddings and can serve as a supervised fine-tuning dataset for larger models such as BERT. Our fixed sentence embeddings achieve high performance on a variety of transfer tasks, including SentEval, and we achieve state-of-the-art results on Penn Discourse Treebank's implicit relation prediction task.

SParC: Cross-Domain Semantic Parsing in Context

- https://www.aclweb.org/anthology/P19-1443/

We present SParC, a dataset for cross-domainSemanticParsing inContext that consists of 4,298 coherent question sequences (12k+ individual questions annotated with SQL queries). It is obtained from controlled user interactions with 200 complex databases over 138 domains. We provide an in-depth analysis of SParC and show that it introduces new challenges compared to existing datasets. SParC demonstrates complex contextual dependencies, (2) has greater semantic diversity, and (3) requires generalization to unseen domains due to its cross-domain nature and the unseen databases at test time. We experiment with two state-of-the-art text-to-SQL models adapted to the context-dependent, cross-domain setup. The best model obtains an exact match accuracy of 20.2% over all questions and less than10% over all interaction sequences, indicating that the cross-domain setting and the con-textual phenomena of the dataset present significant challenges for future research. The dataset, baselines, and leaderboard are released at https://yale-lily.github.io/sparc.

Towards Complex Text-to-SQL in Cross-Domain Database with Intermediate Representation

- https://www.aclweb.org/anthology/P19-1444/

We present a neural approach called IRNet for complex and cross-domain Text-to-SQL. IRNet aims to address two challenges: 1) the mismatch between intents expressed in natural language (NL) and the implementation details in SQL; 2) the challenge in predicting columns caused by the large number of out-of-domain words. Instead of end-to-end synthesizing a SQL query, IRNet decomposes the synthesis process into three phases. In the first phase, IRNet performs a schema linking over a question and a database schema. Then, IRNet adopts a grammar-based neural model to synthesize a SemQL query which is an intermediate representation that we design to bridge NL and SQL. Finally, IRNet deterministically infers a SQL query from the synthesized SemQL query with domain knowledge. On the challenging Text-to-SQL benchmark Spider, IRNet achieves 46.7% accuracy, obtaining 19.5% absolute improvement over previous state-of-the-art approaches. At the time of writing, IRNet achieves the first position on the Spider leaderboard.

EigenSent: Spectral sentence embeddings using higher-order Dynamic Mode Decomposition

- https://www.aclweb.org/anthology/P19-1445/

Distributed representation of words, or word embeddings, have motivated methods for calculating semantic representations of word sequences such as phrases, sentences and paragraphs. Most of the existing methods to do so either use algorithms to learn such representations, or improve on calculating weighted averages of the word vectors. In this work, we experiment with spectral methods

of signal representation and summarization as mechanisms for constructing such word-sequence embeddings in an unsupervised fashion. In particular, we explore an algorithm rooted in fluid-dynamics, known as higher-order Dynamic Mode Decomposition, which is designed to capture the eigenfrequencies, and hence the fundamental transition dynamics, of periodic and quasi-periodic systems. It is empirically observed that this approach, which we call EigenSent, can summarize transitions in a sequence of words and generate an embedding that can represent well the sequence itself. To the best of the authors' knowledge, this is the first application of a spectral decomposition and signal summarization technique on text, to create sentence embeddings. We test the efficacy of this algorithm in creating sentence embeddings on three public datasets, where it performs appreciably well. Moreover it is also shown that, due to the positive combination of their complementary properties, concatenating the embeddings generated by EigenSent with simple word vector averaging achieves state-of-the-art results.

SemBleu: A Robust Metric for AMR Parsing Evaluation

- https://www.aclweb.org/anthology/P19-1446/

Evaluating AMR parsing accuracy involves comparing pairs of AMR graphs. The major evaluation metric, SMATCH (Cai and Knight, 2013), searches for one-to-one mappings between the nodes of two AMRs with a greedy hill-climbing algorithm, which leads to search errors. We propose SEMBLEU, a robust metric that extends BLEU (Papineni et al., 2002) to AMRs. It does not suffer from search errors and considers non-local correspondences in addition to local ones. SEMBLEU is fully content-driven and punishes situations where a system's output does not preserve most information from the input. Preliminary experiments on both sentence and corpus levels show that SEMBLEU has slightly higher consistency with human judgments than SMATCH. Our code is available at http://github.com/ freesunshine0316/sembleu.

Reranking for Neural Semantic Parsing

- https://www.aclweb.org/anthology/P19-1447/

Semantic parsing considers the task of transducing natural language (NL) utterances into machine executable meaning representations (MRs). While neural network-based semantic parsers have achieved impressive improvements over previous methods, results are still far from perfect, and cursory manual inspection can easily identify obvious problems such as lack of adequacy or coherence of the generated MRs. This paper presents a simple approach to quickly iterate and improve the performance of an existing neural semantic parser by reranking an n-best list of predicted MRs, using features that are designed to fix observed problems with baseline models. We implement our reranker in a competitive

neural semantic parser and test on four semantic parsing (GEO, ATIS) and Python code generation (Django, CoNaLa) tasks, improving the strong baseline parser by up to 5.7% absolute in BLEU (CoNaLa) and 2.9% in accuracy (Django), outperforming the best published neural parser results on all four datasets.

Representing Schema Structure with Graph Neural Networks for Text-to-SQL Parsing

- https://www.aclweb.org/anthology/P19-1448/

Research on parsing language to SQL has largely ignored the structure of the database (DB) schema, either because the DB was very simple, or because it was observed at both training and test time. In spider, a recently-released text-to-SQL dataset, new and complex DBs are given at test time, and so the structure of the DB schema can inform the predicted SQL query. In this paper, we present an encoder-decoder semantic parser, where the structure of the DB schema is encoded with a graph neural network, and this representation is later used at both encoding and decoding time. Evaluation shows that encoding the schema structure improves our parser accuracy from 33.8% to 39.4%, dramatically above the current state of the art, which is at 19.7%.

Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark

- https://www.aclweb.org/anthology/P19-1449/

The GLUE benchmark (Wang et al., 2019b) is a suite of language understanding tasks which has seen dramatic progress in the past year, with average performance moving from 70.0 at launch to 83.9, state of the art at the time of writing (May 24, 2019). Here, we measure human performance on the benchmark, in order to learn whether significant headroom remains for further progress. We provide a conservative estimate of human performance on the benchmark through crowdsourcing: Our annotators are non-experts who must learn each task from a brief set of instructions and 20 examples. In spite of limited training, these annotators robustly outperform the state of the art on six of the nine GLUE tasks and achieve an average score of 87.1. Given the fast pace of progress however, the headroom we observe is quite limited. To reproduce the data-poor setting that our annotators must learn in, we also train the BERT model (Devlin et al., 2019) in limited-data regimes, and conclude that low-resource sentence classification remains a challenge for modern neural network approaches to text understanding.

Compositional Semantic Parsing across Graphbanks

- https://www.aclweb.org/anthology/P19-1450/

Most semantic parsers that map sentences to graph-based meaning representations are hand-designed for specific graphbanks. We present a compositional neural semantic parser which achieves, for the first time, competitive accuracies across a diverse range of graphbanks. Incorporating BERT embeddings and multi-task learning improves the accuracy further, setting new states of the art on DM, PAS, PSD, AMR 2015 and EDS.

Rewarding Smatch: Transition-Based AMR Parsing with Reinforcement Learning

- https://www.aclweb.org/anthology/P19-1451/

Our work involves enriching the Stack-LSTM transition-based AMR parser (Ballesteros and Al-Onaizan, 2017) by augmenting training with Policy Learning and rewarding the Smatch score of sampled graphs. In addition, we also combined several AMR-to-text alignments with an attention mechanism and we supplemented the parser with pre-processed concept identification, named entities and contextualized embeddings. We achieve a highly competitive performance that is comparable to the best published results. We show an in-depth study ablating each of the new components of the parser.

BERT Rediscovers the Classical NLP Pipeline

- https://www.aclweb.org/anthology/P19-1452/

Pre-trained text encoders have rapidly advanced the state of the art on many NLP tasks. We focus on one such model, BERT, and aim to quantify where linguistic information is captured within the network. We find that the model represents the steps of the traditional NLP pipeline in an interpretable and localizable way, and that the regions responsible for each step appear in the expected sequence: POS tagging, parsing, NER, semantic roles, then coreference. Qualitative analysis reveals that the model can and often does adjust this pipeline dynamically, revising lower-level decisions on the basis of disambiguating information from higher-level representations.

Simple and Effective Paraphrastic Similarity from Parallel Translations

- https://www.aclweb.org/anthology/P19-1453/

We present a model and methodology for learning paraphrastic sentence embeddings directly from bitext, removing the time-consuming intermediate step of creating para-phrase corpora. Further, we show that the resulting model can be applied to cross lingual tasks where it both outperforms and is orders of magnitude faster than more complex state-of-the-art baselines.

Second-Order Semantic Dependency Parsing with End-to-End Neural Networks

- https://www.aclweb.org/anthology/P19-1454/

Semantic dependency parsing aims to identify semantic relationships between words in a sentence that form a graph. In this paper, we propose a second-order semantic dependency parser, which takes into consideration not only individual dependency edges but also interactions between pairs of edges. We show that second-order parsing can be approximated using mean field (MF) variational inference or loopy belief propagation (LBP). We can unfold both algorithms as recurrent layers of a neural network and therefore can train the parser in an end-to-end manner. Our experiments show that our approach achieves state-of-the-art performance.

Towards Multimodal Sarcasm Detection (An *Obviously* Perfect Paper)

- https://www.aclweb.org/anthology/P19-1455/

Sarcasm is often expressed through several verbal and non-verbal cues, e.g., a change of tone, overemphasis in a word, a drawn-out syllable, or a straight looking face. Most of the recent work in sarcasm detection has been carried out on textual data. In this paper, we argue that incorporating multimodal cues can improve the automatic classification of sarcasm. As a first step towards enabling the development of multimodal approaches for sarcasm detection, we propose a new sarcasm dataset, Multimodal Sarcasm Detection Dataset (MUStARD), compiled from popular TV shows. MUStARD consists of audiovisual utterances annotated with sarcasm labels. Each utterance is accompanied by its context of historical utterances in the dialogue, which provides additional information on the scenario where the utterance occurs. Our initial results show that the use of multimodal information can reduce the relative error rate of sarcasm detection by up to 12.9% in F-score when compared to the use of individual modalities. The full dataset is publicly available for use at https://github.com/soujanyaporia/MUStARD.

Determining Relative Argument Specificity and Stance for Complex Argumentative Structures

- https://www.aclweb.org/anthology/P19-1456/

Systems for automatic argument generation and debate require the ability to (1) determine the stance of any claims employed in the argument and (2) assess the specificity of each claim relative to the argument context. Existing work on understanding claim specificity and stance, however, has been limited to the study of argumentative structures that are relatively shallow, most often consisting of a single claim that directly supports or opposes the argument thesis. In this paper, we tackle these tasks in the context of complex arguments on a diverse set of topics. In particular, our dataset consists of manually curated argument trees for 741 controversial topics covering 95,312 unique claims; lines of argument are generally of depth 2 to 6. We find that as the distance between a pair of claims increases along the argument path, determining the relative specificity of a pair of claims becomes easier and determining their relative stance becomes harder.

Latent Variable Sentiment Grammar

- https://www.aclweb.org/anthology/P19-1457/

Neural models have been investigated for sentiment classification over constituent trees. They learn phrase composition automatically by encoding tree structures but do not explicitly model sentiment composition, which requires to encode sentiment class labels. To this end, we investigate two formalisms with deep sentiment representations that capture sentiment subtype expressions by latent variables and Gaussian mixture vectors, respectively. Experiments on Stanford Sentiment Treebank (SST) show the effectiveness of sentiment grammar over vanilla neural encoders. Using ELMo embeddings, our method gives the best results on this benchmark.

An Investigation of Transfer Learning-Based Sentiment Analysis in Japanese

- https://www.aclweb.org/anthology/P19-1458/

Text classification approaches have usually required task-specific model architectures and huge labeled datasets. Recently, thanks to the rise of text-based transfer learning techniques, it is possible to pre-train a language model in an unsupervised manner and leverage them to perform effective on downstream tasks. In this work we focus on Japanese and show the potential use of transfer learning techniques in text classification. Specifically, we perform binary and multi-class sentiment classification on the Rakuten product review and Yahoo movie review

datasets. We show that transfer learning-based approaches perform better than task-specific models trained on 3 times as much data. Furthermore, these approaches perform just as well for language modeling pre-trained on only 1/30 of the data. We release our pre-trained models and code as open source.

Probing Neural Network Comprehension of Natural Language Arguments

- https://www.aclweb.org/anthology/P19-1459/

We are surprised to find that BERT's peak performance of 77% on the Argument Reasoning Comprehension Task reaches just three points below the average untrained human baseline. However, we show that this result is entirely accounted for by exploitation of spurious statistical cues in the dataset. We analyze the nature of these cues and demonstrate that a range of models all exploit them. This analysis informs the construction of an adversarial dataset on which all models achieve random accuracy. Our adversarial dataset provides a more robust assessment of argument comprehension and should be adopted as the standard in future work.

Recognising Agreement and Disagreement between Stances with Reason Comparing Networks

- https://www.aclweb.org/anthology/P19-1460/

We identify agreement and disagreement between utterances that express stances towards a topic of discussion. Existing methods focus mainly on conversational settings, where dialogic features are used for (dis)agreement inference. We extend this scope and seek to detect stance (dis)agreement in a broader setting, where independent stance-bearing utterances, which prevail in many stance corpora and real-world scenarios, are compared. To cope with such non-dialogic utterances, we find that the reasons uttered to back up a specific stance can help predict stance (dis)agreements. We propose a reason comparing network (RCN) to leverage reason information for stance comparison. Empirical results on a well-known stance corpus show that our method can discover useful reason information, enabling it to outperform several baselines in stance (dis)agreement detection.

Toward Comprehensive Understanding of a Sentiment Based on Human Motives

- https://www.aclweb.org/anthology/P19-1461/

In sentiment detection, the natural language processing community has focused on determining holders, facets, and valences, but has paid little attention to the reasons for sentiment decisions. Our work considers human motives as the driver for human sentiments and addresses the problem of motive detection as the first step. Following a study in psychology, we define six basic motives that cover a wide range of topics appearing in review texts, annotate 1,600 texts in restaurant and laptop domains with the motives, and report the performance of baseline methods on this new dataset. We also show that cross-domain transfer learning boosts detection performance, which indicates that these universal motives exist across different domains.

Context-aware Embedding for Targeted Aspect-based Sentiment Analysis

- https://www.aclweb.org/anthology/P19-1462/

Attention-based neural models were employed to detect the different aspects and sentiment polarities of the same target in targeted aspect-based sentiment analysis (TABSA). However, existing methods do not specifically pre-train reasonable embeddings for targets and aspects in TABSA. This may result in targets or aspects having the same vector representations in different contexts and losing the context-dependent information. To address this problem, we propose a novel method to refine the embeddings of targets and aspects. Such pivotal embedding refinement utilizes a sparse coefficient vector to adjust the embeddings of target and aspect from the context. Hence the embeddings of targets and aspects can be refined from the highly correlative words instead of using context-independent or randomly initialized vectors. Experiment results on two benchmark datasets show that our approach yields the state-of-the-art performance in TABSA task.

Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates

- https://www.aclweb.org/anthology/P19-1463/

Political debates offer a rare opportunity for citizens to compare the candidates' positions on the most controversial topics of the campaign. Thus they represent a natural application scenario for Argument Mining. As existing research lacks solid empirical investigation of the typology of argument components in political debates, we fill this gap by proposing an Argument Mining approach to political debates. We address this task in an empirical manner by annotating 39 political debates from the last 50 years of US presidential campaigns, creating a new corpus of 29k argument components, labeled as premises and claims. We then propose two tasks: (1) identifying the argumentative components in

such debates, and (2) classifying them as premises and claims. We show that feature-rich SVM learners and Neural Network architectures outperform standard baselines in Argument Mining over such complex data. We release the new corpus USElecDeb60To16 and the accompanying software under free licenses to the research community.

An Empirical Study of Span Representations in Argumentation Structure Parsing

- https://www.aclweb.org/anthology/P19-1464/

For several natural language processing (NLP) tasks, span representation design is attracting considerable attention as a promising new technique; a common basis for an effective design has been established. With such basis, exploring task-dependent extensions for argumentation structure parsing (ASP) becomes an interesting research direction. This study investigates (i) span representation originally developed for other NLP tasks and (ii) a simple task-dependent extension for ASP. Our extensive experiments and analysis show that these representations yield high performance for ASP and provide some challenging types of instances to be parsed.

Simple and Effective Text Matching with Richer Alignment Features

- https://www.aclweb.org/anthology/P19-1465/

In this paper, we present a fast and strong neural approach for general purpose text matching applications. We explore what is sufficient to build a fast and well-performed text matching model and propose to keep three key features available for inter-sequence alignment: original point-wise features, previous aligned features, and contextual features while simplifying all the remaining components. We conduct experiments on four well-studied benchmark datasets across tasks of natural language inference, paraphrase identification and answer selection. The performance of our model is on par with the state-of-the-art on all datasets with much fewer parameters and the inference speed is at least 6 times faster compared with similarly performed ones.

Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs

- https://www.aclweb.org/anthology/P19-1466/

The recent proliferation of knowledge graphs (KGs) coupled with incomplete or partial information, in the form of missing relations (links) between entities, has

fueled a lot of research on knowledge base completion (also known as relation prediction). Several recent works suggest that convolutional neural network (CNN) based models generate richer and more expressive feature embeddings and hence also perform well on relation prediction. However, we observe that these KG embeddings treat triples independently and thus fail to cover the complex and hidden information that is inherently implicit in the local neighborhood surrounding a triple. To this effect, our paper proposes a novel attention-based feature embedding that captures both entity and relation features in any given entity's neighborhood. Additionally, we also encapsulate relation clusters and multi-hop relations in our model. Our empirical study offers insights into the efficacy of our attention-based model and we show marked performance gains in comparison to state-of-the-art methods on all datasets.

Neural Network Alignment for Sentential Paraphrases

- https://www.aclweb.org/anthology/P19-1467/

We present a monolingual alignment system for long, sentence- or clause-level alignments, and demonstrate that systems designed for word- or short phrase-based alignment are ill-suited for these longer alignments. Our system is capable of aligning semantically similar spans of arbitrary length. We achieve significantly higher recall on aligning phrases of four or more words and outperform state-of-the- art aligners on the long alignments in the MSR RTE corpus.

Duality of Link Prediction and Entailment Graph Induction

- https://www.aclweb.org/anthology/P19-1468/

Link prediction and entailment graph induction are often treated as different problems. In this paper, we show that these two problems are actually complementary. We train a link prediction model on a knowledge graph of assertions extracted from raw text. We propose an entailment score that exploits the new facts discovered by the link prediction model, and then form entailment graphs between relations. We further use the learned entailments to predict improved link prediction scores. Our results show that the two tasks can benefit from each other. The new entailment score outperforms prior state-of-the-art results on a standard entialment dataset and the new link prediction scores show improvements over the raw link prediction scores.

A Cross-Sentence Latent Variable Model for Semi-Supervised Text Sequence Matching

- https://www.aclweb.org/anthology/P19-1469/

We present a latent variable model for predicting the relationship between a pair of text sequences. Unlike previous auto-encoding–based approaches that consider each sequence separately, our proposed framework utilizes both sequences within a single model by generating a sequence that has a given relationship with a source sequence. We further extend the cross-sentence generating framework to facilitate semi-supervised training. We also define novel semantic constraints that lead the decoder network to generate semantically plausible and diverse sequences. We demonstrate the effectiveness of the proposed model from quantitative and qualitative experiments, while achieving state-of-the-art results on semi-supervised natural language inference and paraphrase identification.

COMET: Commonsense Transformers for Automatic Knowledge Graph Construction

- https://www.aclweb.org/anthology/P19-1470/

We present the first comprehensive study on automatic knowledge base construction for two prevalent commonsense knowledge graphs: ATOMIC (Sap et al., 2019) and ConceptNet (Speer et al., 2017). Contrary to many conventional KBs that store knowledge with canonical templates, commonsense KBs only store loosely structured open-text descriptions of knowledge. We posit that an important step toward automatic commonsense completion is the development of generative models of commonsense knowledge, and propose COMmonsEnse Transformers (COMET) that learn to generate rich and diverse commonsense descriptions in natural language. Despite the challenges of commonsense modeling, our investigation reveals promising results when implicit knowledge from deep pre-trained language models is transferred to generate explicit knowledge in commonsense knowledge graphs. Empirical results demonstrate that COMET is able to generate novel knowledge that humans rate as high quality, with up to 77.5% (ATOMIC) and 91.7% (ConceptNet) precision at top 1, which approaches human performance for these resources. Our findings suggest that using generative commonsense models for automatic commonsense KB completion could soon be a plausible alternative to extractive methods.

Detecting Subevents using Discourse and Narrative Features

- https://www.aclweb.org/anthology/P19-1471/

Recognizing the internal structure of events is a challenging language processing task of great importance for text understanding. We present a supervised model for automatically identifying when one event is a subevent of another. Building on prior work, we introduce several novel features, in particular discourse and narrative features, that significantly improve upon prior state-of-the-art performance. Error analysis further demonstrates the utility of these features. We

evaluate our model on the only two annotated corpora with event hierarchies: HiEve and the Intelligence Community corpus. No prior system has been evaluated on both corpora. Our model outperforms previous systems on both corpora, achieving 0.74 BLANC F1 on the Intelligence Community corpus and 0.70 F1 on the HiEve corpus, respectively a 15 and 5 percentage point improvement over previous models.

HellaSwag: Can a Machine Really Finish Your Sentence?

- https://www.aclweb.org/anthology/P19-1472/

Recent work by Zellers et al. (2018) introduced a new task of commonsense natural language inference: given an event description such as "A woman sits at a piano," a machine must select the most likely followup: "She sets her fingers on the keys." With the introduction of BERT, near human-level performance was reached. Does this mean that machines can perform human level commonsense inference? In this paper, we show that commonsense inference still proves difficult for even state-of-the-art models, by presenting HellaSwag, a new challenge dataset. Though its questions are trivial for humans (>95% accuracy), state-of-the-art models struggle (<48%). We achieve this via Adversarial Filtering (AF), a data collection paradigm wherein a series of discriminators iteratively select an adversarial set of machine-generated wrong answers. AF proves to be surprisingly robust. The key insight is to scale up the length and complexity of the dataset examples towards a critical 'Goldilocks' zone wherein generated text is ridiculous to humans, yet often misclassified by state-of-the-art models. Our construction of HellaSwag, and its resulting difficulty, sheds light on the inner workings of deep pretrained models. More broadly, it suggests a new path forward for NLP research, in which benchmarks co-evolve with the evolving state-of-the-art in an adversarial way, so as to present ever-harder challenges.

Unified Semantic Parsing with Weak Supervision

- https://www.aclweb.org/anthology/P19-1473/

Semantic parsing over multiple knowledge bases enables a parser to exploit structural similarities of programs across the multiple domains. However, the fundamental challenge lies in obtaining high-quality annotations of (utterance, program) pairs across various domains needed for training such models. To overcome this, we propose a novel framework to build a unified multi-domain enabled semantic parser trained only with weak supervision (denotations). Weakly supervised training is particularly arduous as the program search space grows exponentially in a multi-domain setting. To solve this, we incorporate a multi-policy distillation mechanism in which we first train domain-specific semantic parsers (teachers) using weak supervision in the absence of the ground truth programs, followed by training a single unified parser (student) from the domain

specific policies obtained from these teachers. The resultant semantic parser is not only compact but also generalizes better, and generates more accurate programs. It further does not require the user to provide a domain label while querying. On the standard Overnight dataset (containing multiple domains), we demonstrate that the proposed model improves performance by 20% in terms of denotation accuracy in comparison to baseline techniques.

Every Child Should Have Parents: A Taxonomy Refinement Algorithm Based on Hyperbolic Term Embeddings

- https://www.aclweb.org/anthology/P19-1474/

We introduce the use of Poincaré embeddings to improve existing state-of-the-art approaches to domain-specific taxonomy induction from text as a signal for both relocating wrong hyponym terms within a (pre-induced) taxonomy as well as for attaching disconnected terms in a taxonomy. This method substantially improves previous state-of-the-art results on the SemEval-2016 Task 13 on taxonomy extraction. We demonstrate the superiority of Poincaré embeddings over distributional semantic representations, supporting the hypothesis that they can better capture hierarchical lexical-semantic relationships than embeddings in the Euclidean space.

Learning to Rank for Plausible Plausibility

- https://www.aclweb.org/anthology/P19-1475/

Researchers illustrate improvements in contextual encoding strategies via resultant performance on a battery of shared Natural Language Understanding (NLU) tasks. Many of these tasks are of a categorical prediction variety: given a conditioning context (e.g., an NLI premise), provide a label based on an associated prompt (e.g., an NLI hypothesis). The categorical nature of these tasks has led to common use of a cross entropy log-loss objective during training. We suggest this loss is intuitively wrong when applied to plausibility tasks, where the prompt by design is neither categorically entailed nor contradictory given the context. Log-loss naturally drives models to assign scores near 0.0 or 1.0, in contrast to our proposed use of a margin-based loss. Following a discussion of our intuition, we describe a confirmation study based on an extreme, synthetically curated task derived from MultiNLI. We find that a margin-based loss leads to a more plausible model of plausibility. Finally, we illustrate improvements on the Choice Of Plausible Alternative (COPA) task through this change in loss.

Generalized Tuning of Distributional Word Vectors for Monolingual and Cross-Lingual Lexical Entailment

- https://www.aclweb.org/anthology/P19-1476/

Lexical entailment (LE; also known as hyponymy-hypernymy or is-a relation) is a core asymmetric lexical relation that supports tasks like taxonomy induction and text generation. In this work, we propose a simple and effective method for fine-tuning distributional word vectors for LE. Our Generalized Lexical ENtailment model (GLEN) is decoupled from the word embedding model and applicable to any distributional vector space. Yet – unlike existing retrofitting models – it captures a general specialization function allowing for LE-tuning of the entire distributional space and not only the vectors of words seen in lexical constraints. Coupled with a multilingual embedding space, GLEN seamlessly enables cross-lingual LE detection. We demonstrate the effectiveness of GLEN in graded LE and report large improvements (over 20% in accuracy) over state-of-the-art in cross-lingual LE detection.

Attention Is (not) All You Need for Commonsense Reasoning

- https://www.aclweb.org/anthology/P19-1477/

The recently introduced BERT model exhibits strong performance on several language understanding benchmarks. In this paper, we describe a simple re-implementation of BERT for commonsense reasoning. We show that the attentions produced by BERT can be directly utilized for tasks such as the Pronoun Disambiguation Problem and Winograd Schema Challenge. Our proposed attention-guided commonsense reasoning method is conceptually simple yet empirically powerful. Experimental analysis on multiple datasets demonstrates that our proposed system performs remarkably well on all cases while outperforming the previously reported state of the art by a margin. While results suggest that BERT seems to implicitly learn to establish complex relationships between entities, solving commonsense reasoning tasks might require more than unsupervised models learned from huge text corpora.

A Surprisingly Robust Trick for the Winograd Schema Challenge

- https://www.aclweb.org/anthology/P19-1478/

The Winograd Schema Challenge (WSC) dataset WSC273 and its inference counterpart WNLI are popular benchmarks for natural language understanding and commonsense reasoning. In this paper, we show that the performance of

three language models on WSC273 consistently and robustly improves when fine-tuned on a similar pronoun disambiguation problem dataset (denoted WSCR). We additionally generate a large unsupervised WSC-like dataset. By fine-tuning the BERT language model both on the introduced and on the WSCR dataset, we achieve overall accuracies of 72.5% and 74.7% on WSC273 and WNLI, improving the previous state-of-the-art solutions by 8.8% and 9.6%, respectively. Furthermore, our fine-tuned models are also consistently more accurate on the "complex" subsets of WSC273, introduced by Trichelair et al. (2018).

Coherent Comments Generation for Chinese Articles with a Graph-to-Sequence Model

- https://www.aclweb.org/anthology/P19-1479/

Automatic article commenting is helpful in encouraging user engagement on online news platforms. However, the news documents are usually too long for models under traditional encoder-decoder frameworks, which often results in general and irrelevant comments. In this paper, we propose to generate comments with a graph-to-sequence model that models the input news as a topic interaction graph. By organizing the article into graph structure, our model can better understand the internal structure of the article and the connection between topics, which makes it better able to generate coherent and informative comments. We collect and release a large scale news-comment corpus from a popular Chinese online news platform Tencent Kuaibao. Extensive experiment results show that our model can generate much more coherent and informative comments compared with several strong baseline models.

Interconnected Question Generation with Coreference Alignment and Conversation Flow Modeling

- https://www.aclweb.org/anthology/P19-1480/

We study the problem of generating interconnected questions in question-answering style conversations. Compared with previous works which generate questions based on a single sentence (or paragraph), this setting is different in two major aspects: (1) Questions are highly conversational. Almost half of them refer back to conversation history using coreferences. (2) In a coherent conversation, questions have smooth transitions between turns. We propose an end-to-end neural model with coreference alignment and conversation flow modeling. The coreference alignment modeling explicitly aligns coreferent mentions in conversation history with corresponding pronominal references in generated questions, which makes generated questions interconnected to conversation history. The conversation flow modeling builds a coherent conversation by starting questioning on the first few sentences in a text

passage and smoothly shifting the focus to later parts. Extensive experiments show that our system outperforms several baselines and can generate highly conversational questions. The code implementation is released at https://github.com/Evan-Gao/conversaional-QG.

Cross-Lingual Training for Automatic Question Generation

- https://www.aclweb.org/anthology/P19-1481/

Automatic question generation (QG) is a challenging problem in natural language understanding. QG systems are typically built assuming access to a large number of training instances where each instance is a question and its corresponding answer. For a new language, such training instances are hard to obtain making the QG problem even more challenging. Using this as our motivation, we study the reuse of an available large QG dataset in a secondary language (e.g. English) to learn a QG model for a primary language (e.g. Hindi) of interest. For the primary language, we assume access to a large amount of monolingual text but only a small QG dataset. We propose a cross-lingual QG model which uses the following training regime: (i) Unsupervised pretraining of language models in both primary and secondary languages and (ii) joint supervised training for QG in both languages. We demonstrate the efficacy of our proposed approach using two different primary languages, Hindi and Chinese. Our proposed framework clearly outperforms a number of baseline models, including a fully-supervised transformer-based model trained on the QG datasets in the primary language. We also create and release a new question answering dataset for Hindi consisting of 6555 sentences.

A Hierarchical Reinforced Sequence Operation Method for Unsupervised Text Style Transfer

- https://www.aclweb.org/anthology/P19-1482/

Unsupervised text style transfer aims to alter text styles while preserving the content, without aligned data for supervision. Existing seq2seq methods face three challenges: 1) the transfer is weakly interpretable, 2) generated outputs struggle in content preservation, and 3) the trade-off between content and style is intractable. To address these challenges, we propose a hierarchical reinforced sequence operation method, named Point-Then-Operate (PTO), which consists of a high-level agent that proposes operation positions and a low-level agent that alters the sentence. We provide comprehensive training objectives to control the fluency, style, and content of the outputs and a mask-based inference algorithm that allows for multi-step revision based on the single-step trained agents. Experimental results on two text style transfer datasets show that our method significantly outperforms recent methods and effectively addresses the aforementioned challenges.

Handling Divergent Reference Texts when Evaluating Table-to-Text Generation

- https://www.aclweb.org/anthology/P19-1483/

Automatically constructed datasets for generating text from semi-structured data (tables), such as WikiBio, often contain reference texts that diverge from the information in the corresponding semi-structured data. We show that metrics which rely solely on the reference texts, such as BLEU and ROUGE, show poor correlation with human judgments when those references diverge. We propose a new metric, PARENT, which aligns n-grams from the reference and generated texts to the semi-structured data before computing their precision and recall. Through a large scale human evaluation study of table-to-text models for WikiBio, we show that PARENT correlates with human judgments better than existing text generation metrics. We also adapt and evaluate the information extraction based evaluation proposed by Wiseman et al (2017), and show that PARENT has comparable correlation to it, while being easier to use. We show that PARENT is also applicable when the reference texts are elicited from humans using the data from the WebNLG challenge.

Unsupervised Question Answering by Cloze Translation

- https://www.aclweb.org/anthology/P19-1484/

Obtaining training data for Question Answering (QA) is time-consuming and resource-intensive, and existing QA datasets are only available for limited domains and languages. In this work, we explore to what extent high quality training data is actually required for Extractive QA, and investigate the possibility of unsupervised Extractive QA. We approach this problem by first learning to generate context, question and answer triples in an unsupervised manner, which we then use to synthesize Extractive QA training data automatically. To generate such triples, we first sample random context paragraphs from a large corpus of documents and then random noun phrases or Named Entity mentions from these paragraphs as answers. Next we convert answers in context to "fill-in-the-blank" cloze questions and finally translate them into natural questions. We propose and compare various unsupervised ways to perform cloze-to-natural question translation, including training an unsupervised NMT model using non-aligned corpora of natural questions and cloze questions as well as a rule-based approach. We find that modern QA models can learn to answer human questions surprisingly well using only synthetic training data. We demonstrate that, without using the SQuAD training data at all, our approach achieves 56.4 F1 on SQuAD v1 (64.5 F1 when the answer is a Named Entity mention), outperforming early supervised models.

MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension

- https://www.aclweb.org/anthology/P19-1485/

A large number of reading comprehension (RC) datasets has been created recently, but little analysis has been done on whether they generalize to one another, and the extent to which existing datasets can be leveraged for improving performance on new ones. In this paper, we conduct such an investigation over ten RC datasets, training on one or more source RC datasets, and evaluating generalization, as well as transfer to a target RC dataset. We analyze the factors that contribute to generalization, and show that training on a source RC dataset and transferring to a target dataset substantially improves performance, even in the presence of powerful contextual representations from BERT (Devlin et al., 2019). We also find that training on multiple source RC datasets leads to robust generalization and transfer, and can reduce the cost of example collection for a new RC dataset. Following our analysis, we propose MultiQA, a BERT-based model, trained on multiple RC datasets, which leads to state-of-the-art performance on five RC datasets. We share our infrastructure for the benefit of the research community.

Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives

- https://www.aclweb.org/anthology/P19-1486/

This paper tackles the problem of reading comprehension over long narratives where documents easily span over thousands of tokens. We propose a curriculum learning (CL) based Pointer-Generator framework for reading/sampling over large documents, enabling diverse training of the neural model based on the notion of alternating contextual difficulty. This can be interpreted as a form of domain randomization and/or generative pretraining during training. To this end, the usage of the Pointer-Generator softens the requirement of having the answer within the context, enabling us to construct diverse training samples for learning. Additionally, we propose a new Introspective Alignment Layer (IAL), which reasons over decomposed alignments using block-based self-attention. We evaluate our proposed method on the NarrativeQA reading comprehension benchmark, achieving state-of-the-art performance, improving existing baselines by 51% relative improvement on BLEU-4 and 17% relative improvement on Rouge-L. Extensive ablations confirm the effectiveness of our proposed IAL and CL components.

Explain Yourself! Leveraging Language Models for Commonsense Reasoning

- https://www.aclweb.org/anthology/P19-1487/

Deep learning models perform poorly on tasks that require commonsense reasoning, which often necessitates some form of world-knowledge or reasoning over information not immediately present in the input. We collect human explanations for commonsense reasoning in the form of natural language sequences and highlighted annotations in a new dataset called Common Sense Explanations (CoS-E). We use CoS-E to train language models to automatically generate explanations that can be used during training and inference in a novel Commonsense Auto-Generated Explanation (CAGE) framework. CAGE improves the state-of-the-art by 10% on the challenging CommonsenseQA task. We further study commonsense reasoning in DNNs using both human and auto-generated explanations including transfer to out-of-domain tasks. Empirical results indicate that we can effectively leverage language models for commonsense reasoning.

Interpretable Question Answering on Knowledge Bases and Text

- https://www.aclweb.org/anthology/P19-1488/

Interpretability of machine learning (ML) models becomes more relevant with their increasing adoption. In this work, we address the interpretability of ML based question answering (QA) models on a combination of knowledge bases (KB) and text documents. We adapt post hoc explanation methods such as LIME and input perturbation (IP) and compare them with the self-explanatory attention mechanism of the model. For this purpose, we propose an automatic evaluation paradigm for explanation methods in the context of QA. We also conduct a study with human annotators to evaluate whether explanations help them identify better QA models. Our results suggest that IP provides better explanations than LIME or attention, according to both automatic and human evaluation. We obtain the same ranking of methods in both experiments, which supports the validity of our automatic evaluation paradigm.

A Resource-Free Evaluation Metric for Cross-Lingual Word Embeddings Based on Graph Modularity

- https://www.aclweb.org/anthology/P19-1489/

Cross-lingual word embeddings encode the meaning of words from different languages into a shared low-dimensional space. An important requirement for many

downstream tasks is that word similarity should be independent of language—i.e., word vectors within one language should not be more similar to each other than to words in another language. We measure this characteristic using modularity, a network measurement that measures the strength of clusters in a graph. Modularity has a moderate to strong correlation with three downstream tasks, even though modularity is based only on the structure of embeddings and does not require any external resources. We show through experiments that modularity can serve as an intrinsic validation metric to improve unsupervised cross-lingual word embeddings, particularly on distant language pairs in low-resource settings.

Multilingual and Cross-Lingual Graded Lexical Entailment

- https://www.aclweb.org/anthology/P19-1490/

Grounded in cognitive linguistics, graded lexical entailment (GR-LE) is concerned with fine-grained assertions regarding the directional hierarchical relationships between concepts on a continuous scale. In this paper, we present the first work on cross-lingual generalisation of GR-LE relation. Starting from HyperLex, the only available GR-LE dataset in English, we construct new monolingual GR-LE datasets for three other languages, and combine those to create a set of six cross-lingual GR-LE datasets termed CL-HYPERLEX. We next present a novel method dubbed CLEAR (Cross-Lingual Lexical Entailment Attract-Repel) for effectively capturing graded (and binary) LE, both monolingually in different languages as well as across languages (i.e., on CL-HYPERLEX). Coupled with a bilingual dictionary, CLEAR leverages taxonomic LE knowledge in a resource-rich language (e.g., English) and propagates it to other languages. Supported by cross-lingual LE transfer, CLEAR sets competitive baseline performance on three new monolingual GR-LE datasets and six cross-lingual GR-LE datasets. In addition, we show that CLEAR outperforms current state-of-the-art on binary cross-lingual LE detection by a wide margin for diverse language pairs.

What Kind of Language Is Hard to Language-Model?

- https://www.aclweb.org/anthology/P19-1491/

How language-agnostic are current state-of-the-art NLP tools? Are there some types of language that are easier to model with current methods? In prior work (Cotterell et al., 2018) we attempted to address this question for language modeling, and observed that recurrent neural network language models do not perform equally well over all the high-resource European languages found in the Europarl corpus. We speculated that inflectional morphology may be the primary culprit for the discrepancy. In this paper, we extend these earlier experiments to cover 69 languages from 13 language families using a multilingual

Bible corpus. Methodologically, we introduce a new paired-sample multiplicative mixed-effects model to obtain language difficulty coefficients from at-least-pairwise parallel corpora. In other words, the model is aware of inter-sentence variation and can handle missing data. Exploiting this model, we show that "translationese" is not any easier to model than natively written language in a fair comparison. Trying to answer the question of what features difficult languages have in common, we try and fail to reproduce our earlier (Cotterell et al., 2018) observation about morphological complexity and instead reveal far simpler statistics of the data that seem to drive complexity in a much larger sample.

Analyzing the Limitations of Cross-lingual Word Embedding Mappings

- https://www.aclweb.org/anthology/P19-1492/

Recent research in cross-lingual word embeddings has almost exclusively focused on offline methods, which independently train word embeddings in different languages and map them to a shared space through linear transformations. While several authors have questioned the underlying isomorphism assumption, which states that word embeddings in different languages have approximately the same structure, it is not clear whether this is an inherent limitation of mapping approaches or a more general issue when learning cross-lingual embeddings. So as to answer this question, we experiment with parallel corpora, which allows us to compare offline mapping to an extension of skip-gram that jointly learns both embedding spaces. We observe that, under these ideal conditions, joint learning yields to more isomorphic embeddings, is less sensitive to hubness, and obtains stronger results in bilingual lexicon induction. We thus conclude that current mapping methods do have strong limitations, calling for further research to jointly learn cross-lingual embeddings with a weaker cross-lingual signal.

How Multilingual is Multilingual BERT?

- https://www.aclweb.org/anthology/P19-1493/

In this paper, we show that Multilingual BERT (M-BERT), released by Devlin et al. (2018) as a single language model pre-trained from monolingual corpora in 104 languages, is surprisingly good at zero-shot cross-lingual model transfer, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. To understand why, we present a large number of probing experiments, showing that transfer is possible even to languages in different scripts, that transfer works best between typologically similar languages, that monolingual corpora can train models for code-switching, and that the model can find translation pairs. From these results, we can

conclude that M-BERT does create multilingual representations, but that these representations exhibit systematic deficiencies affecting certain language pairs.

Bilingual Lexicon Induction through Unsupervised Machine Translation

- https://www.aclweb.org/anthology/P19-1494/

A recent research line has obtained strong results on bilingual lexicon induction by aligning independently trained word embeddings in two languages and using the resulting cross-lingual embeddings to induce word translation pairs through nearest neighbor or related retrieval methods. In this paper, we propose an alternative approach to this problem that builds on the recent work on unsupervised machine translation. This way, instead of directly inducing a bilingual lexicon from cross-lingual embeddings, we use them to build a phrase-table, combine it with a language model, and use the resulting machine translation system to generate a synthetic parallel corpus, from which we extract the bilingual lexicon using statistical word alignment techniques. As such, our method can work with any word embedding and cross-lingual mapping technique, and it does not require any additional resource besides the monolingual corpus used to train the embeddings. When evaluated on the exact same cross-lingual embeddings, our proposed method obtains an average improvement of 6 accuracy points over nearest neighbor and 4 points over CSLS retrieval, establishing a new state-of-the-art in the standard MUSE dataset.

Automatically Identifying Complaints in Social Media

- https://www.aclweb.org/anthology/P19-1495/

Complaining is a basic speech act regularly used in human and computer mediated communication to express a negative mismatch between reality and expectations in a particular situation. Automatically identifying complaints in social media is of utmost importance for organizations or brands to improve the customer experience or in developing dialogue systems for handling and responding to complaints. In this paper, we introduce the first systematic analysis of complaints in computational linguistics. We collect a new annotated data set of written complaints expressed on Twitter. We present an extensive linguistic analysis of complaining as a speech act in social media and train strong feature-based and neural models of complaints across nine domains achieving a predictive performance of up to 79 F1 using distant supervision.

TWEETQA: A Social Media Focused Question Answering Dataset

- https://www.aclweb.org/anthology/P19-1496/

With social media becoming increasingly popular on which lots of news and real-time events are reported, developing automated question answering systems is critical to the effective-ness of many applications that rely on real-time knowledge. While previous datasets have concentrated on question answering (QA) for formal text like news and Wikipedia, we present the first large-scale dataset for QA over social media data. To ensure that the tweets we collected are useful, we only gather tweets used by journalists to write news articles. We then ask human annotators to write questions and answers upon these tweets. Unlike otherQA datasets like SQuAD in which the answers are extractive, we allow the answers to be ive. We show that two recently proposed neural models that perform well on formal texts are limited in their performance when applied to our dataset. In addition, even the fine-tuned BERT model is still lagging behind human performance with a large margin. Our results thus point to the need of improved QA systems targeting social media text.

Asking the Crowd: Question Analysis, Evaluation and Generation for Open Discussion on Online Forums

- https://www.aclweb.org/anthology/P19-1497/

Teaching machines to ask questions is an important yet challenging task. Most prior work focused on generating questions with fixed answers. As contents are highly limited by given answers, these questions are often not worth discussing. In this paper, we take the first step on teaching machines to ask open-answered questions from real-world news for open discussion (openQG). To generate high-qualified questions, effective ways for question evaluation are required. We take the perspective that the more answers a question receives, the better it is for open discussion, and analyze how language use affects the number of answers. Compared with other factors, e.g. topic and post time, linguistic factors keep our evaluation from being domain-specific. We carefully perform variable control on 11.5M questions from online forums to get a dataset, OQRanD, and further perform question analysis. Based on these conclusions, several models are built for question evaluation. For openQG task, we construct OQGenD, the first dataset as far as we know, and propose a model based on conditional generative adversarial networks and our question evaluation model. Experiments show that our model can generate questions with higher quality compared with commonly-used text generation methods.

Tree LSTMs with Convolution Units to Predict Stance and Rumor Veracity in Social Media Conversations

- https://www.aclweb.org/anthology/P19-1498/

Learning from social-media conversations has gained significant attention recently because of its applications in areas like rumor detection. In this research, we propose a new way to represent social-media conversations as binarized constituency trees that allows comparing features in source-posts and their replies effectively. Moreover, we propose to use convolution units in Tree LSTMs that are better at learning patterns in features obtained from the source and reply posts. Our Tree LSTM models employ multi-task (stance + rumor) learning and propagate the useful stance signal up in the tree for rumor classification at the root node. The proposed models achieve state-of-the-art performance, outperforming the current best model by 12% and 15% on F1-macro for rumor-veracity classification and stance classification tasks respectively.

HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization

- https://www.aclweb.org/anthology/P19-1499/

Neural extractive summarization models usually employ a hierarchical encoder for document encoding and they are trained using sentence-level labels, which are created heuristically using rule-based methods. Training the hierarchical encoder with these inaccurate labels is challenging. Inspired by the recent work on pre-training transformer sentence encoders (Devlin et al., 2018), we propose Hibert (as shorthand for HIerachical Bidirectional Encoder Representations from Transformers) for document encoding and a method to pre-train it using unlabeled data. We apply the pre-trained Hibert to our summarization model and it outperforms its randomly initialized counterpart by 1.25 ROUGE on the CNN/Dailymail dataset and by 2.0 ROUGE on a version of New York Times dataset. We also achieve the state-of-the-art performance on these two datasets.

Hierarchical Transformers for Multi-Document Summarization

- https://www.aclweb.org/anthology/P19-1500/

In this paper, we develop a neural summarization model which can effectively process multiple input documents and distill Transformer architecture with the ability to encode documents in a hierarchical manner. We represent cross-document relationships via an attention mechanism which allows to share information as opposed to simply concatenating text spans and processing them as a flat sequence. Our model learns latent dependencies among textual units, but

can also take advantage of explicit graph representations focusing on similarity or discourse relations. Empirical results on the WikiSum dataset demonstrate that the proposed architecture brings substantial improvements over several strong baselines.

ive Text Summarization Based on Deep Learning and Semantic Content Generalization

- https://www.aclweb.org/anthology/P19-1501/

This work proposes a novel framework for enhancing ive text summarization based on the combination of deep learning techniques along with semantic data transformations. Initially, a theoretical model for semantic-based text generalization is introduced and used in conjunction with a deep encoder-decoder architecture in order to produce a summary in generalized form. Subsequently, a methodology is proposed which transforms the aforementioned generalized summary into human-readable form, retaining at the same time important informational aspects of the original text and addressing the problem of out-of-vocabulary or rare words. The overall approach is evaluated on two popular datasets with encouraging results.

Studying Summarization Evaluation Metrics in the Appropriate Scoring Range

- https://www.aclweb.org/anthology/P19-1502/

In summarization, automatic evaluation metrics are usually compared based on their ability to correlate with human judgments. Unfortunately, the few existing human judgment datasets have been created as by-products of the manual evaluations performed during the DUC/TAC shared tasks. However, modern systems are typically better than the best systems submitted at the time of these shared tasks. We show that, surprisingly, evaluation metrics which behave similarly on these datasets (average-scoring range) strongly disagree in the higher-scoring range in which current systems now operate. It is problematic because metrics disagree yet we can't decide which one to trust. This is a call for collecting human judgments for high-scoring summaries as this would resolve the debate over which metrics to trust. This would also be greatly beneficial to further improve summarization systems and metrics alike.

Simple Unsupervised Summarization by Contextual Matching

- https://www.aclweb.org/anthology/P19-1503/

We propose an unsupervised method for sentence summarization using only language modeling. The approach employs two language models, one that is generic (i.e. pretrained), and the other that is specific to the target domain. We show that by using a product-of-experts criteria these are enough for maintaining continuous contextual matching while maintaining output fluency. Experiments on both ive and extractive sentence summarization data sets show promising results of our method without being exposed to any paired data.

Generating Summaries with Topic Templates and Structured Convolutional Decoders

- https://www.aclweb.org/anthology/P19-1504/

Existing neural generation approaches create multi-sentence text as a single sequence. In this paper we propose a structured convolutional decoder that is guided by the content structure of target summaries. We compare our model with existing sequential decoders on three data sets representing different domains. Automatic and human evaluation demonstrate that our summaries have better content coverage.

Morphological Irregularity Correlates with Frequency

- https://www.aclweb.org/anthology/P19-1505/

We present a study of morphological irregularity. Following recent work, we define an information-theoretic measure of irregularity based on the predictability of forms in a language. Using a neural transduction model, we estimate this quantity for the forms in 28 languages. We first present several validatory and exploratory analyses of irregularity. We then show that our analyses provide evidence for a correlation between irregularity and frequency: higher frequency items are more likely to be irregular and irregular items are more likely be highly frequent. To our knowledge, this result is the first of its breadth and confirms longstanding proposals from the linguistics literature. The correlation is more robust when aggregated at the level of whole paradigms—providing support for models of linguistic structure in which inflected forms are unified by underlying stems or lexemes.

Like a Baby: Visually Situated Neural Language Acquisition

- https://www.aclweb.org/anthology/P19-1506/

We examine the benefits of visual context in training neural language models to perform next-word prediction. A multi-modal neural architecture is introduced

that outperform its equivalent trained on language alone with a 2% decrease in perplexity, even when no visual context is available at test. Fine-tuning the embeddings of a pre-trained state-of-the-art bidirectional language model (BERT) in the language modeling framework yields a 3.5% improvement. The advantage for training with visual context when testing without is robust across different languages (English, German and Spanish) and different models (GRU, LSTM, Delta-RNN, as well as those that use BERT embeddings). Thus, language models perform better when they learn like a baby, i.e, in a multi-modal environment. This finding is compatible with the theory of situated cognition: language is inseparable from its physical context.

Relating Simple Sentence Representations in Deep Neural Networks and the Brain

- https://www.aclweb.org/anthology/P19-1507/

What is the relationship between sentence representations learned by deep recurrent models against those encoded by the brain? Is there any correspondence between hidden layers of these recurrent models and brain regions when processing sentences? Can these deep models be used to synthesize brain data which can then be utilized in other extrinsic tasks? We investigate these questions using sentences with simple syntax and semantics (e.g., The bone was eaten by the dog.). We consider multiple neural network architectures, including recently proposed ELMo and BERT. We use magnetoencephalography (MEG) brain recording data collected from human subjects when they were reading these simple sentences. Overall, we find that BERT's activations correlate the best with MEG brain data. We also find that the deep network representation can be used to generate brain data from new sentences to augment existing brain data. To the best of our knowledge, this is the first work showing that the MEG brain recording when reading a word in a sentence can be used to distinguish earlier words in the sentence. Our exploration is also the first to use deep neural network representations to generate synthetic brain data and to show that it helps in improving subsequent stimuli decoding task accuracy.

Modeling Affirmative and Negated Action Processing in the Brain with Lexical and Compositional Semantic Models

- https://www.aclweb.org/anthology/P19-1508/

Recent work shows that distributional semantic models can be used to decode patterns of brain activity associated with individual words and sentence meanings. However, it is yet unclear to what extent such models can be used to study and decode fMRI patterns associated with specific aspects of semantic composition such as the negation function. In this paper, we apply lexical and compositional semantic models to decode fMRI patterns associated with negated and

affirmative sentences containing hand-action verbs. Our results show reduced decoding (correlation) of sentences where the verb is in the negated context, as compared to the affirmative one, within brain regions implicated in action-semantic processing. This supports behavioral and brain imaging studies, suggesting that negation involves reduced access to aspects of the affirmative mental representation. The results pave the way for testing alternate semantic models of negation against human semantic processing in the brain.

Word-order Biases in Deep-agent Emergent Communication

- https://www.aclweb.org/anthology/P19-1509/

Sequence-processing neural networks led to remarkable progress on many NLP tasks. As a consequence, there has been increasing interest in understanding to what extent they process language as humans do. We aim here to uncover which biases such models display with respect to "natural" word-order constraints. We train models to communicate about paths in a simple gridworld, using miniature languages that reflect or violate various natural language trends, such as the tendency to avoid redundancy or to minimize long-distance dependencies. We study how the controlled characteristics of our miniature languages affect individual learning and their stability across multiple network generations. The results draw a mixed picture. On the one hand, neural networks show a strong tendency to avoid long-distance dependencies. On the other hand, there is no clear preference for the efficient, non-redundant encoding of information that is widely attested in natural language. We thus suggest inoculating a notion of "effort" into neural networks, as a possible way to make their linguistic behavior more human-like.

NNE: A Dataset for Nested Named Entity Recognition in English Newswire

- https://www.aclweb.org/anthology/P19-1510/

Named entity recognition (NER) is widely used in natural language processing applications and downstream tasks. However, most NER tools target flat annotation from popular datasets, eschewing the semantic information available in nested entity mentions. We describe NNE—a fine-grained, nested named entity dataset over the full Wall Street Journal portion of the Penn Treebank (PTB). Our annotation comprises 279,795 mentions of 114 entity types with up to 6 layers of nesting. We hope the public release of this large dataset for English newswire will encourage development of new techniques for nested NER.

Sequence-to-Nuggets: Nested Entity Mention Detection via Anchor-Region Networks

- https://www.aclweb.org/anthology/P19-1511/

Sequential labeling-based NER approaches restrict each word belonging to at most one entity mention, which will face a serious problem when recognizing nested entity mentions. In this paper, we propose to resolve this problem by modeling and leveraging the head-driven phrase structures of entity mentions, i.e., although a mention can nest other mentions, they will not share the same head word. Specifically, we propose Anchor-Region Networks (ARNs), a sequence-to-nuggets architecture for nested mention detection. ARNs first identify anchor words (i.e., possible head words) of all mentions, and then recognize the mention boundaries for each anchor word by exploiting regular phrase structures. Furthermore, we also design Bag Loss, an objective function which can train ARNs in an end-to-end manner without using any anchor word annotation. Experiments show that ARNs achieve the state-of-the-art performance on three standard nested entity mention detection benchmarks.

Improving Textual Network Embedding with Global Attention via Optimal Transport

- https://www.aclweb.org/anthology/P19-1512/

Constituting highly informative network embeddings is an essential tool for network analysis. It encodes network topology, along with other useful side information, into low dimensional node-based feature representations that can be exploited by statistical modeling. This work focuses on learning context-aware network embeddings augmented with text data. We reformulate the network embedding problem, and present two novel strategies to improve over traditional attention mechanisms: (i) a content-aware sparse attention module based on optimal transport; and (ii) a high-level attention parsing module. Our approach yields naturally sparse and self-normalized relational inference. It can capture long-term interactions between sequences, thus addressing the challenges faced by existing textual network embedding schemes. Extensive experiments are conducted to demonstrate our model can consistently outperform alternative state-of-the-art methods.

Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction

- https://www.aclweb.org/anthology/P19-1513/

While the fast-paced inception of novel tasks and new datasets helps foster active research in a community towards interesting directions, keeping track of the

abundance of research activity in different areas on different datasets is likely to become increasingly difficult. The community could greatly benefit from an automatic system able to summarize scientific results, e.g., in the form of a leaderboard. In this paper we build two datasets and develop a framework (TDMS-IE) aimed at automatically extracting task, dataset, metric and score from NLP papers, towards the automatic construction of leaderboards. Experiments show that our model outperforms several baselines by a large margin. Our model is a first step towards automatic leaderboard construction, e.g., in the NLP domain.

Scaling up Open Tagging from Tens to Thousands: Comprehension Empowered Attribute Value Extraction from Product Title

- https://www.aclweb.org/anthology/P19-1514/

Supplementing product information by extracting attribute values from title is a crucial task in e-Commerce domain. Previous studies treat each attribute only as an entity type and build one set of NER tags (e.g., BIO) for each of them, leading to a scalability issue which unfits to the large sized attribute system in real world e-Commerce. In this work, we propose a novel approach to support value extraction scaling up to thousands of attributes without losing performance: (1) We propose to regard attribute as a query and adopt only one global set of BIO tags for any attributes to reduce the burden of attribute tag or model explosion; (2) We explicitly model the semantic representations for attribute and title, and develop an attention mechanism to capture the interactive semantic relations in-between to enforce our framework to be attribute comprehensive. We conduct extensive experiments in real-life datasets. The results show that our model not only outperforms existing state-of-the-art NER tagging models, but also is robust and generates promising results for up to 8,906 attributes.

Incorporating Linguistic Constraints into Keyphrase Generation

- https://www.aclweb.org/anthology/P19-1515/

Keyphrases, that concisely describe the high-level topics discussed in a document, are very useful for a wide range of natural language processing tasks. Though existing keyphrase generation methods have achieved remarkable performance on this task, they generate many overlapping phrases (including subphrases or super-phrases) of keyphrases. In this paper, we propose the parallel Seq2Seq network with the coverage attention to alleviate the overlapping phrase problem. Specifically, we integrate the linguistic constraints of keyphrase into

the basic Seq2Seq network on the source side, and employ the multi-task learning framework on the target side. In addition, in order to prevent from generating overlapping phrases of keyphrases with correct syntax, we introduce the coverage vector to keep track of the attention history and to decide whether the parts of source text have been covered by existing generated keyphrases. Experimental results show that our method can outperform the state-of-the-art CopyRNN on scientific datasets, and is also more effective in news domain.

A Unified Multi-task Adversarial Learning Framework for Pharmacovigilance Mining

- https://www.aclweb.org/anthology/P19-1516/

The mining of adverse drug reaction (ADR) has a crucial role in the pharmacovigilance. The traditional ways of identifying ADR are reliable but time-consuming, non-scalable and offer a very limited amount of ADR relevant information. With the unprecedented growth of information sources in the forms of social media texts (Twitter, Blogs, Reviews etc.), biomedical literature, and Electronic Medical Records (EMR), it has become crucial to extract the most pertinent ADR related information from these free-form texts. In this paper, we propose a neural network inspired multi- task learning framework that can simultaneously extract ADRs from various sources. We adopt a novel adversarial learning-based approach to learn features across multiple ADR information sources. Unlike the other existing techniques, our approach is capable to extracting fine-grained information (such as 'Indications', 'Symptoms', 'Finding', 'Disease', 'Drug') which provide important cues in pharmacovigilance. We evaluate our proposed approach on three publicly available real- world benchmark pharmacovigilance datasets, a Twitter dataset from PSB 2016 Social Me- dia Shared Task, CADEC corpus and Medline ADR corpus. Experiments show that our unified framework achieves state-of-the-art performance on individual tasks associated with the different benchmark datasets. This establishes the fact that our proposed approach is generic, which enables it to achieve high performance on the diverse datasets.

Quantity Tagger: A Latent-Variable Sequence Labeling Approach to Solving Addition-Subtraction Word Problems

- https://www.aclweb.org/anthology/P19-1517/

An arithmetic word problem typically includes a textual description containing several constant quantities. The key to solving the problem is to reveal the underlying mathematical relations (such as addition and subtraction) among quantities, and then generate equations to find solutions. This work presents a novel approach, Quantity Tagger, that automatically discovers such hidden

relations by tagging each quantity with a sign corresponding to one type of mathematical operation. For each quantity, we assume there exists a latent, variable-sized quantity span surrounding the quantity token in the text, which conveys information useful for determining its sign. Empirical results show that our method achieves 5 and 8 points of accuracy gains on two datasets respectively, compared to prior approaches.

A Deep Reinforced Sequence-to-Set Model for Multi-Label Classification

- https://www.aclweb.org/anthology/P19-1518/

Multi-label classification (MLC) aims to predict a set of labels for a given instance. Based on a pre-defined label order, the sequence-to-sequence (Seq2Seq) model trained via maximum likelihood estimation method has been successfully applied to the MLC task and shows powerful ability to capture high-order correlations between labels. However, the output labels are essentially an unordered set rather than an ordered sequence. This inconsistency tends to result in some intractable problems, e.g., sensitivity to the label order. To remedy this, we propose a simple but effective sequence-to-set model. The proposed model is trained via reinforcement learning, where reward feedback is designed to be independent of the label order. In this way, we can reduce the dependence of the model on the label order, as well as capture high-order correlations between labels. Extensive experiments show that our approach can substantially outperform competitive baselines, as well as effectively reduce the sensitivity to the label order.

Joint Slot Filling and Intent Detection via Capsule Neural Networks

- https://www.aclweb.org/anthology/P19-1519/

Being able to recognize words as slots and detect the intent of an utterance has been a keen issue in natural language understanding. The existing works either treat slot filling and intent detection separately in a pipeline manner, or adopt joint models which sequentially label slots while summarizing the utterance-level intent without explicitly preserving the hierarchical relationship among words, slots, and intents. To exploit the semantic hierarchy for effective modeling, we propose a capsule-based neural network model which accomplishes slot filling and intent detection via a dynamic routing-by-agreement schema. A re-routing schema is proposed to further synergize the slot filling performance using the inferred intent representation. Experiments on two real-world datasets show the effectiveness of our model when compared with other alternative model architectures, as well as existing natural language understanding services.

Neural Aspect and Opinion Term Extraction with Mined Rules as Weak Supervision

- https://www.aclweb.org/anthology/P19-1520/

Lack of labeled training data is a major bottleneck for neural network based aspect and opinion term extraction on product reviews. To alleviate this problem, we first propose an algorithm to automatically mine extraction rules from existing training examples based on dependency parsing results. The mined rules are then applied to label a large amount of auxiliary data. Finally, we study training procedures to train a neural model which can learn from both the data automatically labeled by the rules and a small amount of data accurately annotated by human. Experimental results show that although the mined rules themselves do not perform well due to their limited flexibility, the combination of human annotated data and rule labeled auxiliary data can improve the neural model and allow it to achieve performance better than or comparable with the current state-of-the-art.

Cost-sensitive Regularization for Label Confusion-aware Event Detection

- https://www.aclweb.org/anthology/P19-1521/

In supervised event detection, most of the mislabeling occurs between a small number of confusing type pairs, including trigger-NIL pairs and sibling sub-types of the same coarse type. To address this label confusion problem, this paper proposes cost-sensitive regularization, which can force the training procedure to concentrate more on optimizing confusing type pairs. Specifically, we introduce a cost-weighted term into the training loss, which penalizes more on mislabeling between confusing label pairs. Furthermore, we also propose two estimators which can effectively measure such label confusion based on instance-level or population-level statistics. Experiments on TAC-KBP 2017 datasets demonstrate that the proposed method can significantly improve the performances of different models in both English and Chinese event detection.

Exploring Pre-trained Language Models for Event Extraction and Generation

- https://www.aclweb.org/anthology/P19-1522/

Traditional approaches to the task of ACE event extraction usually depend on manually annotated data, which is often laborious to create and limited in size. Therefore, in addition to the difficulty of event extraction itself, insufficient training data hinders the learning process as well. To promote event extraction, we first propose an event extraction model to overcome the roles overlap problem

by separating the argument prediction in terms of roles. Moreover, to address the problem of insufficient training data, we propose a method to automatically generate labeled data by editing prototypes and screen out generated samples by ranking the quality. Experiments on the ACE2005 dataset demonstrate that our extraction model can surpass most existing extraction methods. Besides, incorporating our generation method exhibits further significant improvement. It obtains new state-of-the-art results on the event extraction task, including pushing the F1 score of trigger classification to 81.1%, and the F1 score of argument classification to 58.9%.

Improving Open Information Extraction via Iterative Rank-Aware Learning

- https://www.aclweb.org/anthology/P19-1523/

Open information extraction (IE) is the task of extracting open-domain assertions from natural language sentences. A key step in open IE is confidence modeling, ranking the extractions based on their estimated quality to adjust precision and recall of extracted assertions. We found that the extraction likelihood, a confidence measure used by current supervised open IE systems, is not well calibrated when comparing the quality of assertions extracted from different sentences. We propose an additional binary classification loss to calibrate the likelihood to make it more globally comparable, and an iterative learning process, where extractions generated by the open IE model are incrementally included as training samples to help the model learn from trial and error. Experiments on OIE2016 demonstrate the effectiveness of our method. Code and data are available at https://github.com/jzbjyb/oie_rank.

Towards Improving Neural Named Entity Recognition with Gazetteers

- https://www.aclweb.org/anthology/P19-1524/

Most of the recently proposed neural models for named entity recognition have been purely data-driven, with a strong emphasis on getting rid of the efforts for collecting external resources or designing hand-crafted features. This could increase the chance of overfitting since the models cannot access any supervision signal beyond the small amount of annotated data, limiting their power to generalize beyond the annotated entities. In this work, we show that properly utilizing external gazetteers could benefit segmental neural NER models. We add a simple module on the recently proposed hybrid semi-Markov CRF architecture and observe some promising results.

Towards Improving Neural Named Entity Recognition with Gazetteers

- https://www.aclweb.org/anthology/P19-1524/

Most of the recently proposed neural models for named entity recognition have been purely data-driven, with a strong emphasis on getting rid of the efforts for collecting external resources or designing hand-crafted features. This could increase the chance of overfitting since the models cannot access any supervision signal beyond the small amount of annotated data, limiting their power to generalize beyond the annotated entities. In this work, we show that properly utilizing external gazetteers could benefit segmental neural NER models. We add a simple module on the recently proposed hybrid semi-Markov CRF architecture and observe some promising results.

Span-Level Model for Relation Extraction

- https://www.aclweb.org/anthology/P19-1525/

Relation Extraction is the task of identifying entity mention spans in raw text and then identifying relations between pairs of the entity mentions. Recent approaches for this span-level task have been token-level models which have inherent limitations. They cannot easily define and implement span-level features, cannot model overlapping entity mentions and have cascading errors due to the use of sequential decoding. To address these concerns, we present a model which directly models all possible spans and performs joint entity mention detection and relation extraction. We report a new state-of-the-art performance of 62.83 F1 (prev best was 60.49) on the ACE2005 dataset.

Enhancing Unsupervised Generative Dependency Parser with Contextual Information

- https://www.aclweb.org/anthology/P19-1526/

Most of the unsupervised dependency parsers are based on probabilistic generative models that learn the joint distribution of the given sentence and its parse. Probabilistic generative models usually explicit decompose the desired dependency tree into factorized grammar rules, which lack the global features of the entire sentence. In this paper, we propose a novel probabilistic model called discriminative neural dependency model with valence (D-NDMV) that generates a sentence and its parse from a continuous latent representation, which encodes global contextual information of the generated sentence. We propose two approaches to model the latent representation: the first deterministically summarizes the representation from the sentence and the second probabilistically models the representation conditioned on the sentence. Our approach can be

regarded as a new type of autoencoder model to unsupervised dependency parsing that combines the benefits of both generative and discriminative techniques. In particular, our approach breaks the context-free independence assumption in previous generative approaches and therefore becomes more expressive. Our extensive experimental results on seventeen datasets from various sources show that our approach achieves competitive accuracy compared with both generative and discriminative state-of-the-art unsupervised dependency parsers.

Neural Architectures for Nested NER through Linearization

- https://www.aclweb.org/anthology/P19-1527/

We propose two neural network architectures for nested named entity recognition (NER), a setting in which named entities may overlap and also be labeled with more than one label. We encode the nested labels using a linearized scheme. In our first proposed approach, the nested labels are modeled as multilabels corresponding to the Cartesian product of the nested labels in a standard LSTM-CRF architecture. In the second one, the nested NER is viewed as a sequence-to-sequence problem, in which the input sequence consists of the tokens and output sequence of the labels, using hard attention on the word whose label is being predicted. The proposed methods outperform the nested NER state of the art on four corpora: ACE-2004, ACE-2005, GENIA and Czech CNEC. We also enrich our architectures with the recently published contextual embeddings: ELMo, BERT and Flair, reaching further improvements for the four nested entity corpora. In addition, we report flat NER state-of-the-art results for CoNLL-2002 Dutch and Spanish and for CoNLL-2003 English.

Online Infix Probability Computation for Probabilistic Finite Automata

- https://www.aclweb.org/anthology/P19-1528/

Probabilistic finite automata (PFAs) are com- mon statistical language model in natural lan- guage and speech processing. A typical task for PFAs is to compute the probability of all strings that match a query pattern. An impor- tant special case of this problem is computing the probability of a string appearing as a prefix, suffix, or infix. These problems find use in many natural language processing tasks such word prediction and text error correction. Recently, we gave the first incremental algorithm to efficiently compute the infix probabilities of each prefix of a string (Cognetta et al., 2018). We develop an asymptotic improvement of that algorithm and solve the open problem of computing the infix probabilities of PFAs from streaming data, which is crucial when process- ing queries online and is the ultimate goal of the incremental approach.

How to Best Use Syntax in Semantic Role Labelling

- https://www.aclweb.org/anthology/P19-1529/

There are many different ways in which external information might be used in a NLP task. This paper investigates how external syntactic information can be used most effectively in the Semantic Role Labeling (SRL) task. We evaluate three different ways of encoding syntactic parses and three different ways of injecting them into a state-of-the-art neural ELMo-based SRL sequence labelling model. We show that using a constituency representation as input features improves performance the most, achieving a new state-of-the-art for non-ensemble SRL models on the in-domain CoNLL'05 and CoNLL'12 benchmarks.

PTB Graph Parsing with Tree Approximation

- https://www.aclweb.org/anthology/P19-1530/

The Penn Treebank (PTB) represents syntactic structures as graphs due to nonlocal dependencies. This paper proposes a method that approximates PTB graph-structured representations by trees. By our approximation method, we can reduce nonlocal dependency identification and constituency parsing into single tree-based parsing. An experimental result demonstrates that our approximation method with an off-the-shelf tree-based constituency parser significantly outperforms the previous methods in nonlocal dependency identification.

Sequence Labeling Parsing by Learning across Representations

- https://www.aclweb.org/anthology/P19-1531/

We use parsing as sequence labeling as a common framework to learn across constituency and dependency syntactic ions.To do so, we cast the problem as multitask learning (MTL). First, we show that adding a parsing paradigm as an auxiliary loss consistently improves the performance on the other paradigm. Secondly, we explore an MTL sequence labeling model that parses both representations, at almost no cost in terms of performance and speed. The results across the board show that on average MTL models with auxiliary losses for constituency parsing outperform single-task ones by 1.05 F1 points, and for dependency parsing by 0.62 UAS points.

A Prism Module for Semantic Disentanglement in Name Entity Recognition

- https://www.aclweb.org/anthology/P19-1532/

Natural Language Processing has been perplexed for many years by the problem that multiple semantics are mixed inside a word, even with the help of context. To solve this problem, we propose a prism module to disentangle the semantic aspects of words and reduce noise at the input layer of a model. In the prism module, some words are selectively replaced with task-related semantic aspects, then these denoised word representations can be fed into downstream tasks to make them easier. Besides, we also introduce a structure to train this module jointly with the downstream model without additional data. This module can be easily integrated into the downstream model and significantly improve the performance of baselines on named entity recognition (NER) task. The ablation analysis demonstrates the rationality of the method. As a side effect, the proposed method also provides a way to visualize the contribution of each word.

Label-Agnostic Sequence Labeling by Copying Nearest Neighbors

- https://www.aclweb.org/anthology/P19-1533/

Retrieve-and-edit based approaches to structured prediction, where structures associated with retrieved neighbors are edited to form new structures, have recently attracted increased interest. However, much recent work merely conditions on retrieved structures (e.g., in a sequence-to-sequence framework), rather than explicitly manipulating them. We show we can perform accurate sequence labeling by explicitly (and only) copying labels from retrieved neighbors. Moreover, because this copying is label-agnostic, we can achieve impressive performance in zero-shot sequence-labeling tasks. We additionally consider a dynamic programming approach to sequence labeling in the presence of retrieved neighbors, which allows for controlling the number of distinct (copied) segments used to form a prediction, and leads to both more interpretable and accurate predictions.

Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset

- https://www.aclweb.org/anthology/P19-1534/

One challenge for dialogue agents is recognizing feelings in the conversation partner and replying accordingly, a key communicative skill. While it is straightforward for humans to recognize and acknowledge others' feelings in a conversation, this is a significant challenge for AI systems due to the paucity of suitable publicly-available datasets for training and evaluation. This work proposes a new benchmark for empathetic dialogue generation and EmpatheticDialogues, a novel dataset of 25k conversations grounded in emotional situations. Our experiments indicate that dialogue models that use our dataset are perceived to be

more empathetic by human evaluators, compared to models merely trained on large-scale Internet conversation data. We also present empirical comparisons of dialogue model adaptations for empathetic responding, leveraging existing models or datasets without requiring lengthy re-training of the full model.

Know More about Each Other: Evolving Dialogue Strategy via Compound Assessment

- https://www.aclweb.org/anthology/P19-1535/

In this paper, a novel Generation-Evaluation framework is developed for multi-turn conversations with the objective of letting both participants know more about each other. For the sake of rational knowledge utilization and coherent conversation flow, a dialogue strategy which controls knowledge selection is instantiated and continuously adapted via reinforcement learning. Under the deployed strategy, knowledge grounded conversations are conducted with two dialogue agents. The generated dialogues are comprehensively evaluated on aspects like informativeness and coherence, which are aligned with our objective and human instinct. These assessments are integrated as a compound reward to guide the evolution of dialogue strategy via policy gradient. Comprehensive experiments have been carried out on the publicly available dataset, demonstrating that the proposed method outperforms the other state-of-the-art approaches significantly.

Training Neural Response Selection for Task-Oriented Dialogue Systems

- https://www.aclweb.org/anthology/P19-1536/

Despite their popularity in the chatbot literature, retrieval-based models have had modest impact on task-oriented dialogue systems, with the main obstacle to their application being the low-data regime of most task-oriented dialogue tasks. Inspired by the recent success of pretraining in language modelling, we propose an effective method for deploying response selection in task-oriented dialogue. To train response selection models for task-oriented dialogue tasks, we propose a novel method which: 1) pretrains the response selection model on large general-domain conversational corpora; and then 2) fine-tunes the pretrained model for the target dialogue domain, relying only on the small in-domain dataset to capture the nuances of the given dialogue domain. Our evaluation on five diverse application domains, ranging from e-commerce to banking, demonstrates the effectiveness of the proposed training method.

Collaborative Dialogue in Minecraft

- https://www.aclweb.org/anthology/P19-1537/

We wish to develop interactive agents that can communicate with humans to collaboratively solve tasks in grounded scenarios. Since computer games allow us to simulate such tasks without the need for physical robots, we define a Minecraft-based collaborative building task in which one player (A, the Architect) is shown a target structure and needs to instruct the other player (B, the Builder) to build this structure. Both players interact via a chat interface. A can observe B but cannot place blocks. We present the Minecraft Dialogue Corpus, a collection of 509 conversations and game logs. As a first step towards our goal of developing fully interactive agents for this task, we consider the subtask of Architect utterance generation, and show how challenging it is.

Neural Response Generation with Meta-words

- https://www.aclweb.org/anthology/P19-1538/

We present open domain dialogue generation with meta-words. A meta-word is a structured record that describes attributes of a response, and thus allows us to explicitly model the one-to-many relationship within open domain dialogues and perform response generation in an explainable and controllable manner. To incorporate meta-words into generation, we propose a novel goal-tracking memory network that formalizes meta-word expression as a goal in response generation and manages the generation process to achieve the goal with a state memory panel and a state controller. Experimental results from both automatic evaluation and human judgment on two large-scale data sets indicate that our model can significantly outperform state-of-the-art generation models in terms of response relevance, response diversity, and accuracy of meta-word expression.

Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading

- https://www.aclweb.org/anthology/P19-1539/

Although neural conversational models are effective in learning how to produce fluent responses, their primary challenge lies in knowing what to say to make the conversation contentful and non-vacuous. We present a new end-to-end approach to contentful neural conversation that jointly models response generation and on-demand machine reading. The key idea is to provide the conversation model with relevant long-form text on the fly as a source of external knowledge. The model performs QA-style reading comprehension on this text in response to each conversational turn, thereby allowing for more focused integration of external knowledge than has been possible in prior approaches. To support further

research on knowledge-grounded conversation, we introduce a new large-scale conversation dataset grounded in external web pages (2.8M turns, 7.4M sentences of grounding). Both human evaluation and automated metrics show that our approach results in more contentful responses compared to a variety of previous methods, improving both the informativeness and diversity of generated output.

Ordinal and Attribute Aware Response Generation in a Multimodal Dialogue System

- https://www.aclweb.org/anthology/P19-1540/

Multimodal dialogue systems have opened new frontiers in the traditional goal-oriented dialogue systems. The state-of-the-art dialogue systems are primarily based on unimodal sources, predominantly the text, and hence cannot capture the information present in the other sources such as videos, audios, images etc. With the availability of large scale multimodal dialogue dataset (MMD) (Saha et al., 2018) on the fashion domain, the visual appearance of the products is essential for understanding the intention of the user. Without capturing the information from both the text and image, the system will be incapable of generating correct and desirable responses. In this paper, we propose a novel position and attribute aware attention mechanism to learn enhanced image representation conditioned on the user utterance. Our evaluation shows that the proposed model can generate appropriate responses while preserving the position and attribute information. Experimental results also prove that our proposed approach attains superior performance compared to the baseline models, and outperforms the state-of-the-art approaches on text similarity based evaluation metrics.

Memory Consolidation for Contextual Spoken Language Understanding with Dialogue Logistic Inference

- https://www.aclweb.org/anthology/P19-1541/

Dialogue contexts are proven helpful in the spoken language understanding (SLU) system and they are typically encoded with explicit memory representations. However, most of the previous models learn the context memory with only one objective to maximizing the SLU performance, leaving the context memory under-exploited. In this paper, we propose a new dialogue logistic inference (DLI) task to consolidate the context memory jointly with SLU in the multi-task framework. DLI is defined as sorting a shuffled dialogue session into its original logical order and shares the same memory encoder and retrieval mechanism as the SLU model. Our experimental results show that various popular contextual SLU models can benefit from our approach, and improvements are quite impressive, especially in slot filling.

Personalizing Dialogue Agents via Meta-Learning

- https://www.aclweb.org/anthology/P19-1542/

Existing personalized dialogue models use human designed persona descriptions to improve dialogue consistency. Collecting such descriptions from existing dialogues is expensive and requires hand-crafted feature designs. In this paper, we propose to extend Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) to personalized dialogue learning without using any persona descriptions. Our model learns to quickly adapt to new personas by leveraging only a few dialogue samples collected from the same user, which is fundamentally different from conditioning the response on the persona descriptions. Empirical results on Persona-chat dataset (Zhang et al., 2018) indicate that our solution outperforms non-meta-learning baselines using automatic evaluation metrics, and in terms of human-evaluated fluency and consistency.

Reading Turn by Turn: Hierarchical Attention Architecture for Spoken Dialogue Comprehension

- https://www.aclweb.org/anthology/P19-1543/

Comprehending multi-turn spoken conversations is an emerging research area, presenting challenges different from reading comprehension of passages due to the interactive nature of information exchange from at least two speakers. Unlike passages, where sentences are often the default semantic modeling unit, in multi-turn conversations, a turn is a topically coherent unit embodied with immediately relevant context, making it a linguistically intuitive segment for computationally modeling verbal interactions. Therefore, in this work, we propose a hierarchical attention neural network architecture, combining turn-level and word-level attention mechanisms, to improve spoken dialogue comprehension performance. Experiments are conducted on a multi-turn conversation dataset, where nurses inquire and discuss symptom information with patients. We empirically show that the proposed approach outperforms standard attention baselines, achieves more efficient learning outcomes, and is more robust to lengthy and out-of-distribution test samples.

A Novel Bi-directional Interrelated Model for Joint Intent Detection and Slot Filling

- https://www.aclweb.org/anthology/P19-1544/

A spoken language understanding (SLU) system includes two main tasks, slot filling (SF) and intent detection (ID). The joint model for the two tasks is becoming a tendency in SLU. But the bi-directional interrelated connections between the intent and slots are not established in the existing joint models. In

this paper, we propose a novel bi-directional interrelated model for joint intent detection and slot filling. We introduce an SF-ID network to establish direct connections for the two tasks to help them promote each other mutually. Besides, we design an entirely new iteration mechanism inside the SF-ID network to enhance the bi-directional interrelated connections. The experimental results show that the relative improvement in the sentence-level semantic frame accuracy of our model is 3.79% and 5.42% on ATIS and Snips datasets, respectively, compared to the state-of-the-art model.

Dual Supervised Learning for Natural Language Understanding and Generation

- https://www.aclweb.org/anthology/P19-1545/

Natural language understanding (NLU) and natural language generation (NLG) are both critical research topics in the NLP and dialogue fields. Natural language understanding is to extract the core semantic meaning from the given utterances, while natural language generation is opposite, of which the goal is to construct corresponding sentences based on the given semantics. However, such dual relationship has not been investigated in literature. This paper proposes a novel learning framework for natural language understanding and generation on top of dual supervised learning, providing a way to exploit the duality. The preliminary experiments show that the proposed approach boosts the performance for both tasks, demonstrating the effectiveness of the dual relationship.

SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking

- https://www.aclweb.org/anthology/P19-1546/

In goal-oriented dialog systems, belief trackers estimate the probability distribution of slot-values at every dialog turn. Previous neural approaches have modeled domain- and slot-dependent belief trackers, and have difficulty in adding new slot-values, resulting in lack of flexibility of domain ontology configurations. In this paper, we propose a new approach to universal and scalable belief tracker, called slot-utterance matching belief tracker (SUMBT). The model learns the relations between domain-slot-types and slot-values appearing in utterances through attention mechanisms based on contextual semantic vectors. Furthermore, the model predicts slot-value labels in a non-parametric way. From our experiments on two dialog corpora, WOZ 2.0 and MultiWOZ, the proposed model showed performance improvement in comparison with slot-dependent methods and achieved the state-of-the-art joint accuracy.

Robust Zero-Shot Cross-Domain Slot Filling with Example Values

- https://www.aclweb.org/anthology/P19-1547/

Task-oriented dialog systems increasingly rely on deep learning-based slot filling models, usually needing extensive labeled training data for target domains. Often, however, little to no target domain training data may be available, or the training and target domain schemas may be misaligned, as is common for web forms on similar websites. Prior zero-shot slot filling models use slot descriptions to learn concepts, but are not robust to misaligned schemas. We propose utilizing both the slot description and a small number of examples of slot values, which may be easily available, to learn semantic representations of slots which are transferable across domains and robust to misaligned schemas. Our approach outperforms state-of-the-art models on two multi-domain datasets, especially in the low-data setting.

Deep Unknown Intent Detection with Margin Loss

- https://www.aclweb.org/anthology/P19-1548/

Identifying the unknown (novel) user intents that have never appeared in the training set is a challenging task in the dialogue system. In this paper, we present a two-stage method for detecting unknown intents. We use bidirectional long short-term memory (BiLSTM) network with the margin loss as the feature extractor. With margin loss, we can learn discriminative deep features by forcing the network to maximize inter-class variance and to minimize intra-class variance. Then, we feed the feature vectors to the density-based novelty detection algorithm, local outlier factor (LOF), to detect unknown intents. Experiments on two benchmark datasets show that our method can yield consistent improvements compared with the baseline methods.

Modeling Semantic Relationship in Multi-turn Conversations with Hierarchical Latent Variables

- https://www.aclweb.org/anthology/P19-1549/

Multi-turn conversations consist of complex semantic structures, and it is still a challenge to generate coherent and diverse responses given previous utterances. It's practical that a conversation takes place under a background, meanwhile, the query and response are usually most related and they are consistent in topic but also different in content. However, little work focuses on such hierarchical relationship among utterances. To address this problem, we propose a Conversational Semantic Relationship RNN (CSRR) model to construct the dependency explicitly. The model contains latent variables in three hierarchies. The

discourse-level one captures the global background, the pair-level one stands for the common topic information between query and response, and the utterance-level ones try to represent differences in content. Experimental results show that our model significantly improves the quality of responses in terms of fluency, coherence, and diversity compared to baseline methods.

Rationally Reappraising ATIS-based Dialogue Systems

- https://www.aclweb.org/anthology/P19-1550/

The Air Travel Information Service (ATIS) corpus has been the most common benchmark for evaluating Spoken Language Understanding (SLU) tasks for more than three decades since it was released. Recent state-of-the-art neural models have obtained F1-scores near 98% on the task of slot filling. We developed a rule-based grammar for the ATIS domain that achieves a 95.82% F1-score on our evaluation set. In the process, we furthermore discovered numerous shortcomings in the ATIS corpus annotation, which we have fixed. This paper presents a detailed account of these shortcomings, our proposed repairs, our rule-based grammar and the neural slot-filling architectures associated with ATIS. We also rationally reappraise the motivations for choosing a neural architecture in view of this account. Fixing the annotation errors results in a relative error reduction of between 19.4 and 52% across all architectures. We nevertheless argue that neural models must play a different role in ATIS dialogues because of the latter's lack of variety.

Learning Latent Trees with Stochastic Perturbations and Differentiable Dynamic Programming

- https://www.aclweb.org/anthology/P19-1551/

We treat projective dependency trees as latent variables in our probabilistic model and induce them in such a way as to be beneficial for a downstream task, without relying on any direct tree supervision. Our approach relies on Gumbel perturbations and differentiable dynamic programming. Unlike previous approaches to latent tree learning, we stochastically sample global structures and our parser is fully differentiable. We illustrate its effectiveness on sentiment analysis and natural language inference tasks. We also study its properties on a synthetic structure induction task. Ablation studies emphasize the importance of both stochasticity and constraining latent structures to be projective trees.

Neural-based Chinese Idiom Recommendation for Enhancing Elegance in Essay Writing

- https://www.aclweb.org/anthology/P19-1552/

Although the proper use of idioms can enhance the elegance of writing, the active use of various expressions is a challenge because remembering idioms is difficult. In this study, we address the problem of idiom recommendation by leveraging a neural machine translation framework, in which we suppose that idioms are written with one pseudo target language. Two types of real-life datasets are collected to support this study. Experimental results show that the proposed approach achieves promising performance compared with other baseline methods.

Better Exploiting Latent Variables in Text Modeling

- https://www.aclweb.org/anthology/P19-1553/

We show that sampling latent variables multiple times at a gradient step helps in improving a variational autoencoder and propose a simple and effective method to better exploit these latent variables through hidden state averaging. Consistent gains in performance on two different datasets, Penn Treebank and Yahoo, indicate the generalizability of our method.

Misleading Failures of Partial-input Baselines

- https://www.aclweb.org/anthology/P19-1554/

Recent work establishes dataset difficulty and removes annotation artifacts via partial-input baselines (e.g., hypothesis-only model for SNLI or question-only model for VQA). A successful partial-input baseline indicates that the dataset is cheatable. But the converse is not necessarily true: failures of partial-input baselines do not mean the dataset is free of artifacts. We first design artificial datasets to illustrate how the trivial patterns that are only visible in the full input can evade any partial-input baseline. Next, we identify such artifacts in the SNLI dataset—a hypothesis-only model augmented with trivial patterns in the premise can solve 15% of previously-thought "hard" examples. Our work provides a caveat for the use and creation of partial-input baselines for datasets.

Soft Contextual Data Augmentation for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1555/

While data augmentation is an important trick to boost the accuracy of deep learning methods in computer vision tasks, its study in natural language tasks is still very limited. In this paper, we present a novel data augmentation method for neural machine translation.Different from previous augmentation methods that randomly drop, swap or replace words with other words in a sentence, we

softly augment a randomly chosen word in a sentence by its contextual mixture of multiple related words. More accurately, we replace the one-hot representation of a word by a distribution (provided by a language model) over the vocabulary, i.e., replacing the embedding of this word by a weighted combination of multiple semantically similar words. Since the weights of those words depend on the contextual information of the word to be replaced,the newly generated sentences capture much richer information than previous augmentation methods. Experimental results on both small scale and large scale machine translation data sets demonstrate the superiority of our method over strong baselines.

Reversing Gradients in Adversarial Domain Adaptation for Question Deduplication and Textual Entailment Tasks

- https://www.aclweb.org/anthology/P19-1556/

Adversarial domain adaptation has been recently proposed as an effective technique for textual matching tasks, such as question deduplication. Here we investigate the use of gradient reversal on adversarial domain adaptation to explicitly learn both shared and unshared (domain specific) representations between two textual domains. In doing so, gradient reversal learns features that explicitly compensate for domain mismatch, while still distilling domain specific knowledge that can improve target domain accuracy. We evaluate reversing gradients for adversarial adaptation on multiple domains, and demonstrate that it significantly outperforms other methods on question deduplication as well as on recognizing textual entailment (RTE) tasks, achieving up to 7% absolute boost in base model accuracy on some datasets.

Towards Integration of Statistical Hypothesis Tests into Deep Neural Networks

- https://www.aclweb.org/anthology/P19-1557/

We report our ongoing work about a new deep architecture working in tandem with a statistical test procedure for jointly training texts and their label descriptions for multi-label and multi-class classification tasks. A statistical hypothesis testing method is used to extract the most informative words for each given class. These words are used as a class description for more label-aware text classification. Intuition is to help the model to concentrate on more informative words rather than more frequent ones. The model leverages the use of label descriptions in addition to the input text to enhance text classification performance. Our method is entirely data-driven, has no dependency on other sources of information than the training data, and is adaptable to different classification problems by providing appropriate training data without major

hyper-parameter tuning. We trained and tested our system on several publicly available datasets, where we managed to improve the state-of-the-art on one set with a high margin and to obtain competitive results on all other ones.

Depth Growing for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1558/

While very deep neural networks have shown effectiveness for computer vision and text classification applications, how to increase the network depth of the neural machine translation (NMT) models for better translation quality remains a challenging problem. Directly stacking more blocks to the NMT model results in no improvement and even drop in performance. In this work, we propose an effective two-stage approach with three specially designed components to construct deeper NMT models, which result in significant improvements over the strong Transformer baselines on WMT14 English→German and English→French translation tasks.

Generating Fluent Adversarial Examples for Natural Languages

- https://www.aclweb.org/anthology/P19-1559/

Efficiently building an adversarial attacker for natural language processing (NLP) tasks is a real challenge. Firstly, as the sentence space is discrete, it is difficult to make small perturbations along the direction of gradients. Secondly, the fluency of the generated examples cannot be guaranteed. In this paper, we propose MHA, which addresses both problems by performing Metropolis-Hastings sampling, whose proposal is designed with the guidance of gradients. Experiments on IMDB and SNLI show that our proposed MHAoutperforms the baseline model on attacking capability. Adversarial training with MHA also leads to better robustness and performance.

Towards Explainable NLP: A Generative Explanation Framework for Text Classification

- https://www.aclweb.org/anthology/P19-1560/

Building explainable systems is a critical problem in the field of Natural Language Processing (NLP), since most machine learning models provide no explanations for the predictions. Existing approaches for explainable machine learning systems tend to focus on interpreting the outputs or the connections between inputs and outputs. However, the fine-grained information (e.g. textual explanations for the labels) is often ignored, and the systems do not explicitly generate

the human-readable explanations. To solve this problem, we propose a novel generative explanation framework that learns to make classification decisions and generate fine-grained explanations at the same time. More specifically, we introduce the explainable factor and the minimum risk training approach that learn to generate more reasonable explanations. We construct two new datasets that contain summaries, rating scores, and fine-grained reasons. We conduct experiments on both datasets, comparing with several strong neural network baseline systems. Experimental results show that our method surpasses all baselines on both datasets, and is able to generate concise explanations at the same time.

Combating Adversarial Misspellings with Robust Word Recognition

- https://www.aclweb.org/anthology/P19-1561/

To combat adversarial spelling mistakes, we propose placing a word recognition model in front of the downstream classifier. Our word recognition models build upon the RNN semi-character architecture, introducing several new backoff strategies for handling rare and unseen words. Trained to recognize words corrupted by random adds, drops, swaps, and keyboard mistakes, our method achieves 32% relative (and 3.3% absolute) error reduction over the vanilla semi-character model. Notably, our pipeline confers robustness on the downstream classifier, outperforming both adversarial training and off-the-shelf spell checkers. Against a BERT model fine-tuned for sentiment analysis, a single adversarially-chosen character attack lowers accuracy from 90.3% to 45.8%. Our defense restores accuracy to 75%. Surprisingly, better word recognition does not always entail greater robustness. Our analysis reveals that robustness also depends upon a quantity that we denote the sensitivity.

An Empirical Investigation of Structured Output Modeling for Graph-based Neural Dependency Parsing

- https://www.aclweb.org/anthology/P19-1562/

In this paper, we investigate the aspect of structured output modeling for the state-of-the-art graph-based neural dependency parser (Dozat and Manning, 2017). With evaluations on 14 treebanks, we empirically show that global output-structured models can generally obtain better performance, especially on the metric of sentence-level Complete Match. However, probably because neural models already learn good global views of the inputs, the improvement brought by structured output modeling is modest.

Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes

- https://www.aclweb.org/anthology/P19-1563/

Automatically analyzing dialogue can help understand and guide behavior in domains such as counseling, where interactions are largely mediated by conversation. In this paper, we study modeling behavioral codes used to asses a psychotherapy treatment style called Motivational Interviewing (MI), which is effective for addressing substance abuse and related problems. Specifically, we address the problem of providing real-time guidance to therapists with a dialogue observer that (1) categorizes therapist and client MI behavioral codes and, (2) forecasts codes for upcoming utterances to help guide the conversation and potentially alert the therapist. For both tasks, we define neural network models that build upon recent successes in dialogue modeling. Our experiments demonstrate that our models can outperform several baselines for both tasks. We also report the results of a careful analysis that reveals the impact of the various network design tradeoffs for modeling therapy dialogue.

Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems

- https://www.aclweb.org/anthology/P19-1564/

Developing Video-Grounded Dialogue Systems (VGDS), where a dialogue is conducted based on visual and audio aspects of a given video, is significantly more challenging than traditional image or text-grounded dialogue systems because (1) feature space of videos span across multiple picture frames, making it difficult to obtain semantic information; and (2) a dialogue agent must perceive and process information from different modalities (audio, video, caption, etc.) to obtain a comprehensive understanding. Most existing work is based on RNNs and sequence-to-sequence architectures, which are not very effective for capturing complex long-term dependencies (like in videos). To overcome this, we propose Multimodal Transformer Networks (MTN) to encode videos and incorporate information from different modalities. We also propose query-aware attention through an auto-encoder to extract query-aware features from non-text modalities. We develop a training procedure to simulate token-level decoding to improve the quality of generated responses during inference. We get state of the art performance on Dialogue System Technology Challenge 7 (DSTC7). Our model also generalizes to another multimodal visual-grounded dialogue task, and obtains promising performance.

Target-Guided Open-Domain Conversation

- https://www.aclweb.org/anthology/P19-1565/

Many real-world open-domain conversation applications have specific goals to achieve during open-ended chats, such as recommendation, psychotherapy, education, etc. We study the problem of imposing conversational goals on open-domain chat agents. In particular, we want a conversational system to chat naturally with human and proactively guide the conversation to a designated target subject. The problem is challenging as no public data is available for learning such a target-guided strategy. We propose a structured approach that introduces coarse-grained keywords to control the intended content of system responses. We then attain smooth conversation transition through turn-level supervised learning, and drive the conversation towards the target with discourse-level constraints. We further derive a keyword-augmented conversation dataset for the study. Quantitative and human evaluations show our system can produce meaningful and effective conversations, significantly improving over other approaches

Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good

- https://www.aclweb.org/anthology/P19-1566/

Developing intelligent persuasive conversational agents to change people's opinions and actions for social good is the frontier in advancing the ethical development of automated dialogue systems. To do so, the first step is to understand the intricate organization of strategic disclosures and appeals employed in human persuasion conversations. We designed an online persuasion task where one participant was asked to persuade the other to donate to a specific charity. We collected a large dataset with 1,017 dialogues and annotated emerging persuasion strategies from a subset. Based on the annotation, we built a baseline classifier with context information and sentence-level features to predict the 10 persuasion strategies used in the corpus. Furthermore, to develop an understanding of personalized persuasion processes, we analyzed the relationships between individuals' demographic and psychological backgrounds including personality, morality, value systems, and their willingness for donation. Then, we analyzed which types of persuasion strategies led to a greater amount of donation depending on the individuals' personal backgrounds. This work lays the ground for developing a personalized persuasive dialogue system.

Improving Neural Conversational Models with Entropy-Based Data Filtering

- https://www.aclweb.org/anthology/P19-1567/

Current neural network-based conversational models lack diversity and generate boring responses to open-ended utterances. Priors such as persona, emotion, or topic provide additional information to dialog models to aid response

generation, but annotating a dataset with priors is expensive and such annotations are rarely available. While previous methods for improving the quality of open-domain response generation focused on either the underlying model or the training objective, we present a method of filtering dialog datasets by removing generic utterances from training data using a simple entropy-based approach that does not require human supervision. We conduct extensive experiments with different variations of our method, and compare dialog models across 17 evaluation metrics to show that training on datasets filtered this way results in better conversational quality as chatbots learn to output more diverse responses.

Zero-shot Word Sense Disambiguation using Sense Definition Embeddings

- https://www.aclweb.org/anthology/P19-1568/

Word Sense Disambiguation (WSD) is a long-standing but open problem in Natural Language Processing (NLP). WSD corpora are typically small in size, owing to an expensive annotation process. Current supervised WSD methods treat senses as discrete labels and also resort to predicting the Most-Frequent-Sense (MFS) for words unseen during training. This leads to poor performance on rare and unseen senses. To overcome this challenge, we propose Extended WSD Incorporating Sense Embeddings (EWISE), a supervised model to perform WSD by predicting over a continuous sense embedding space as opposed to a discrete label space. This allows EWISE to generalize over both seen and unseen senses, thus achieving generalized zero-shot learning. To obtain target sense embeddings, EWISE utilizes sense definitions. EWISE learns a novel sentence encoder for sense definitions by using WordNet relations and also ConvE, a recently proposed knowledge graph embedding method. We also compare EWISE against other sentence encoders pretrained on large corpora to generate definition embeddings. EWISE achieves new state-of-the-art WSD performance.

Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation

- https://www.aclweb.org/anthology/P19-1569/

Contextual embeddings represent a new generation of semantic representations learned from Neural Language Modelling (NLM) that addresses the issue of meaning conflation hampering traditional word embeddings. In this work, we show that contextual embeddings can be used to achieve unprecedented gains in Word Sense Disambiguation (WSD) tasks. Our approach focuses on creating sense-level embeddings with full-coverage of WordNet, and without recourse to explicit knowledge of sense distributions or task-specific modelling. As a

result, a simple Nearest Neighbors (k-NN) method using our representations is able to consistently surpass the performance of previous systems using powerful neural sequencing models. We also analyse the robustness of our approach when ignoring part-of-speech and lemma features, requiring disambiguation against the full sense inventory, and revealing shortcomings to be improved. Finally, we explore applications of our sense embeddings for concept-level analyses of contextual embeddings and their respective NLMs.

Word2Sense: Sparse Interpretable Word Embeddings

- https://www.aclweb.org/anthology/P19-1570/

We present an unsupervised method to generate Word2Sense word embeddings that are interpretable — each dimension of the embedding space corresponds to a fine-grained sense, and the non-negative value of the embedding along the j-th dimension represents the relevance of the j-th sense to the word. The underlying LDA-based generative model can be extended to refine the representation of a polysemous word in a short context, allowing us to use the embedings in contextual tasks. On computational NLP tasks, Word2Sense embeddings compare well with other word embeddings generated by unsupervised methods. Across tasks such as word similarity, entailment, sense induction, and contextual interpretation, Word2Sense is competitive with the state-of-the-art method for that task. Word2Sense embeddings are at least as sparse and fast to compute as prior art.

Modeling Semantic Compositionality with Sememe Knowledge

- https://www.aclweb.org/anthology/P19-1571/

Semantic compositionality (SC) refers to the phenomenon that the meaning of a complex linguistic unit can be composed of the meanings of its constituents. Most related works focus on using complicated compositionality functions to model SC while few works consider external knowledge in models. In this paper, we verify the effectiveness of sememes, the minimum semantic units of human languages, in modeling SC by a confirmatory experiment. Furthermore, we make the first attempt to incorporate sememe knowledge into SC models, and employ the sememe-incorporated models in learning representations of multiword expressions, a typical task of SC. In experiments, we implement our models by incorporating knowledge from a famous sememe knowledge base HowNet and perform both intrinsic and extrinsic evaluations. Experimental results show that our models achieve significant performance boost as compared to the baseline methods without considering sememe knowledge. We further conduct quantitative analysis and case studies to demonstrate the effectiveness of applying

sememe knowledge in modeling SC.All the code and data of this paper can be obtained on https://github.com/thunlp/Sememe-SC.

Predicting Humorousness and Metaphor Novelty with Gaussian Process Preference Learning

- https://www.aclweb.org/anthology/P19-1572/

The inability to quantify key aspects of creative language is a frequent obstacle to natural language understanding. To address this, we introduce novel tasks for evaluating the creativeness of language—namely, scoring and ranking text by humorousness and metaphor novelty. To sidestep the difficulty of assigning discrete labels or numeric scores, we learn from pairwise comparisons between texts. We introduce a Bayesian approach for predicting humorousness and metaphor novelty using Gaussian process preference learning (GPPL), which achieves a Spearman's of 0.56 against gold using word embeddings and linguistic features. Our experiments show that given sparse, crowdsourced annotation data, ranking using GPPL outperforms best–worst scaling. We release a new dataset for evaluating humour containing 28,210 pairwise comparisons of 4,030 texts, and make our software freely available.

Empirical Linguistic Study of Sentence Embeddings

- https://www.aclweb.org/anthology/P19-1573/

The purpose of the research is to answer the question whether linguistic information is retained in vector representations of sentences. We introduce a method of analysing the content of sentence embeddings based on universal probing tasks, along with the classification datasets for two contrasting languages. We perform a series of probing and downstream experiments with different types of sentence embeddings, followed by a thorough analysis of the experimental results. Aside from dependency parser-based embeddings, linguistic information is retained best in the recently proposed LASER sentence embeddings.

Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings

- https://www.aclweb.org/anthology/P19-1574/

Word embeddings typically represent different meanings of a word in a single conflated vector. Empirical analysis of embeddings of ambiguous words is currently limited by the small size of manually annotated resources and by the fact that word senses are treated as unrelated individual concepts. We present a large dataset based on manual Wikipedia annotations and word senses, where

word senses from different words are related by semantic classes. This is the basis for novel diagnostic tests for an embedding's content: we probe word embeddings for semantic classes and analyze the embedding space by classifying embeddings into semantic classes. Our main findings are: (i) Information about a sense is generally represented well in a single-vector embedding – if the sense is frequent. (ii) A classifier can accurately predict whether a word is single-sense or multi-sense, based only on its embedding. (iii) Although rare senses are not well represented in single-vector embeddings, this does not have negative impact on an NLP application whose performance depends on frequent senses.

Deep Neural Model Inspection and Comparison via Functional Neuron Pathways

- https://www.aclweb.org/anthology/P19-1575/

We introduce a general method for the interpretation and comparison of neural models. The method is used to factor a complex neural model into its functional components, which are comprised of sets of co-firing neurons that cut across layers of the network architecture, and which we call neural pathways. The function of these pathways can be understood by identifying correlated task level and linguistic heuristics in such a way that this knowledge acts as a lens for approximating what the network has learned to apply to its intended task. As a case study for investigating the utility of these pathways, we present an examination of pathways identified in models trained for two standard tasks, namely Named Entity Recognition and Recognizing Textual Entailment.

Collocation Classification with Unsupervised Relation Vectors

- https://www.aclweb.org/anthology/P19-1576/

Lexical relation classification is the task of predicting whether a certain relation holds between a given pair of words. In this paper, we explore to which extent the current distributional landscape based on word embeddings provides a suitable basis for classification of collocations, i.e., pairs of words between which idiosyncratic lexical relations hold. First, we introduce a novel dataset with collocations categorized according to lexical functions. Second, we conduct experiments on a subset of this benchmark, comparing it in particular to the well known DiffVec dataset. In these experiments, in addition to simple word vector arithmetic operations, we also investigate the role of unsupervised relation vectors as a complementary input. While these relation vectors indeed help, we also show that lexical function classification poses a greater challenge than the syntactic and semantic relations that are typically used for benchmarks in the literature.

Corpus-based Check-up for Thesaurus

- https://www.aclweb.org/anthology/P19-1577/

In this paper we discuss the usefulness of applying a checking procedure to existing thesauri. The procedure is based on the analysis of discrepancies of corpus-based and thesaurus-based word similarities. We applied the procedure to more than 30 thousand words of the Russian wordnet and found some serious errors in word sense description, including inaccurate relationships and missing senses of ambiguous words.

Confusionset-guided Pointer Networks for Chinese Spelling Check

- https://www.aclweb.org/anthology/P19-1578/

This paper proposes Confusionset-guided Pointer Networks for Chinese Spell Check (CSC) task. More concretely, our approach utilizes the off-the-shelf confusionset for guiding the character generation. To this end, our novel Seq2Seq model jointly learns to copy a correct character from an input sentence through a pointer network, or generate a character from the confusionset rather than the entire vocabulary. We conduct experiments on three human-annotated datasets, and results demonstrate that our proposed generative model outperforms all competitor models by a large margin of up to 20% F1 score, achieving state-of-the-art performance on three datasets.

Generalized Data Augmentation for Low-Resource Translation

- https://www.aclweb.org/anthology/P19-1579/

Low-resource language pairs with a paucity of parallel data pose challenges for machine translation in terms of both adequacy and fluency. Data augmentation utilizing a large amount of monolingual data is regarded as an effective way to alleviate the problem. In this paper, we propose a general framework of data augmentation for low-resource machine translation not only using target-side monolingual data, but also by pivoting through a related high-resource language. Specifically, we experiment with a two-step pivoting method to convert high-resource data to the low-resource language, making best use of available resources to better approximate the true distribution of the low-resource language. First, we inject low-resource words into high-resource sentences through an induced bilingual dictionary. Second, we further edit the high-resource data injected with low-resource words using a modified unsupervised machine translation framework. Extensive experiments on four low-resource datasets show that under extreme low-resource settings, our data augmentation techniques improve

translation quality by up to 1.5 to 8 BLEU points compared to supervised back-translation baselines.

Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned

- https://www.aclweb.org/anthology/P19-1580/

Multi-head self-attention is a key component of the Transformer, a state-of-the-art architecture for neural machine translation. In this work we evaluate the contribution made by individual attention heads to the overall performance of the model and analyze the roles played by them in the encoder. We find that the most important and confident heads play consistent and often linguistically-interpretable roles. When pruning heads using a method based on stochastic gates and a differentiable relaxation of the L0 penalty, we observe that specialized heads are last to be pruned. Our novel pruning method removes the vast majority of heads without seriously affecting performance. For example, on the English-Russian WMT dataset, pruning 38 out of 48 encoder heads results in a drop of only 0.15 BLEU.

Better OOV Translation with Bilingual Terminology Mining

- https://www.aclweb.org/anthology/P19-1581/

Unseen words, also called out-of-vocabulary words (OOVs), are difficult for machine translation. In neural machine translation, byte-pair encoding can be used to represent OOVs, but they are still often incorrectly translated. We improve the translation of OOVs in NMT using easy-to-obtain monolingual data. We look for OOVs in the text to be translated and translate them using simple-to-construct bilingual word embeddings (BWEs). In our MT experiments we take the 5-best candidates, which is motivated by intrinsic mining experiments. Using all five of the proposed target language words as queries we mine target-language sentences. We then back-translate, forcing the back-translation of each of the five proposed target-language OOV-translation-candidates to be the original source-language OOV. We show that by using this synthetic data to fine-tune our system the translation of OOVs can be dramatically improved. In our experiments we use a system trained on Europarl and mine sentences containing medical terms from monolingual data.

Simultaneous Translation with Flexible Policy via Restricted Imitation Learning

- https://www.aclweb.org/anthology/P19-1582/

Simultaneous translation is widely useful but remains one of the most difficult tasks in NLP. Previous work either uses fixed-latency policies, or train a complicated two-staged model using reinforcement learning. We propose a much simpler single model that adds a "delay" token to the target vocabulary, and design a restricted dynamic oracle to greatly simplify training. Experiments on Chinese <-> English simultaneous translation show that our work leads to flexible policies that achieve better BLEU scores and lower latencies compared to both fixed and RL-learned policies.

Target Conditioned Sampling: Optimizing Data Selection for Multilingual Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1583/

To improve low-resource Neural Machine Translation (NMT) with multilingual corpus, training on the most related high-resource language only is generally more effective than us- ing all data available (Neubig and Hu, 2018). However, it remains a question whether a smart data selection strategy can further improve low-resource NMT with data from other auxiliary languages. In this paper, we seek to construct a sampling distribution over all multilingual data, so that it minimizes the training loss of the low-resource language. Based on this formulation, we propose and efficient algorithm, (TCS), which first samples a target sentence, and then conditionally samples its source sentence. Experiments show TCS brings significant gains of up to 2 BLEU improvements on three of four languages we test, with minimal training overhead.

Adversarial Learning of Privacy-Preserving Text Representations for De-Identification of Medical Records

- https://www.aclweb.org/anthology/P19-1584/

De-identification is the task of detecting protected health information (PHI) in medical text. It is a critical step in sanitizing electronic health records (EHR) to be shared for research. Automatic de-identification classifiers can significantly speed up the sanitization process. However, obtaining a large and diverse dataset to train such a classifier that works well across many types of medical text poses a challenge as privacy laws prohibit the sharing of raw medical records. We introduce a method to create privacy-preserving shareable representations of medical text (i.e. they contain no PHI) that does not require expensive manual pseudonymization. These representations can be shared between organizations to create unified datasets for training de-identification models. Our representation allows training a simple LSTM-CRF de-identification model to an F1 score of 97.4%, which is comparable to a strong baseline that exposes private information in its representation. A robust, widely available de-identification

classifier based on our representation could potentially enable studies for which de-identification would otherwise be too costly.

Merge and Label: A Novel Neural Network Architecture for Nested NER

- https://www.aclweb.org/anthology/P19-1585/

Named entity recognition (NER) is one of the best studied tasks in natural language processing. However, most approaches are not capable of handling nested structures which are common in many applications. In this paper we introduce a novel neural network architecture that first merges tokens and/or entities into entities forming nested structures, and then labels each of them independently. Unlike previous work, our merge and label approach predicts real-valued instead of discrete segmentation structures, which allow it to combine word and nested entity embeddings while maintaining differentiability. We evaluate our approach using the ACE 2005 Corpus, where it achieves state-of-the-art F1 of 74.6, further improved with contextual embeddings (BERT) to 82.4, an overall improvement of close to 8 F1 points over previous approaches trained on the same data. Additionally we compare it against BiLSTM-CRFs, the dominant approach for flat NER structures, demonstrating that its ability to predict nested structures does not impact performance in simpler cases.

Low-resource Deep Entity Resolution with Transfer and Active Learning

- https://www.aclweb.org/anthology/P19-1586/

Entity resolution (ER) is the task of identifying different representations of the same real-world entities across databases. It is a key step for knowledge base creation and text mining. Recent adaptation of deep learning methods for ER mitigates the need for dataset-specific feature engineering by constructing distributed representations of entity records. While these methods achieve state-of-the-art performance over benchmark data, they require large amounts of labeled data, which are typically unavailable in realistic ER applications. In this paper, we develop a deep learning-based method that targets low-resource settings for ER through a novel combination of transfer learning and active learning. We design an architecture that allows us to learn a transferable model from a high-resource setting to a low-resource one. To further adapt to the target dataset, we incorporate active learning that carefully selects a few informative examples to fine-tune the transferred model. Empirical evaluation demonstrates that our method achieves comparable, if not better, performance compared to state-of-the-art learning-based methods while using an order of magnitude fewer labels.

A Semi-Markov Structured Support Vector Machine Model for High-Precision Named Entity Recognition

- https://www.aclweb.org/anthology/P19-1587/

Named entity recognition (NER) is the backbone of many NLP solutions. F1 score, the harmonic mean of precision and recall, is often used to select/evaluate the best models. However, when precision needs to be prioritized over recall, a state-of-the-art model might not be the best choice. There is little in literature that directly addresses training-time modifications to achieve higher precision information extraction. In this paper, we propose a neural semi-Markov structured support vector machine model that controls the precision-recall trade-off by assigning weights to different types of errors in the loss-augmented inference during training. The semi-Markov property provides more accurate phrase-level predictions, thereby improving performance. We empirically demonstrate the advantage of our model when high precision is required by comparing against strong baselines based on CRF. In our experiments with the CoNLL 2003 dataset, our model achieves a better precision-recall trade-off at various precision levels.

Using Human Attention to Extract Keyphrase from Microblog Post

- https://www.aclweb.org/anthology/P19-1588/

This paper studies automatic keyphrase extraction on social media. Previous works have achieved promising results on it, but they neglect human reading behavior during keyphrase annotating. The human attention is a crucial element of human reading behavior. It reveals the relevance of words to the main topics of the target text. Thus, this paper aims to integrate human attention into keyphrase extraction models. First, human attention is represented by the reading duration estimated from eye-tracking corpus. Then, we merge human attention with neural network models by an attention mechanism. In addition, we also integrate human attention into unsupervised models. To the best of our knowledge, we are the first to utilize human attention on keyphrase extraction tasks. The experimental results show that our models have significant improvements on two Twitter datasets.

Model-Agnostic Meta-Learning for Relation Classification with Limited Supervision

- https://www.aclweb.org/anthology/P19-1589/

In this paper we frame the task of supervised relation classification as an instance of meta-learning. We propose a model-agnostic meta-learning protocol

for training relation classifiers to achieve enhanced predictive performance in limited supervision settings. During training, we aim to not only learn good parameters for classifying relations with sufficient supervision, but also learn model parameters that can be fine-tuned to enhance predictive performance for relations with limited supervision. In experiments conducted on two relation classification datasets, we demonstrate that the proposed meta-learning approach improves the predictive performance of two state-of-the-art supervised relation classification models.

Variational Pretraining for Semi-supervised Text Classification

- https://www.aclweb.org/anthology/P19-1590/

We introduce VAMPIRE, a lightweight pretraining framework for effective text classification when data and computing resources are limited. We pretrain a unigram document model as a variational autoencoder on in-domain, unlabeled data and use its internal states as features in a downstream classifier. Empirically, we show the relative strength of VAMPIRE against computationally expensive contextual embeddings and other popular semi-supervised baselines under low resource settings. We also find that fine-tuning to in-domain data is crucial to achieving decent performance from contextual embeddings when working with limited supervision. We accompany this paper with code to pretrain and use VAMPIRE embeddings in downstream tasks.

Task Refinement Learning for Improved Accuracy and Stability of Unsupervised Domain Adaptation

- https://www.aclweb.org/anthology/P19-1591/

Pivot Based Language Modeling (PBLM) (Ziser and Reichart, 2018a), combining LSTMs with pivot-based methods, has yielded significant progress in unsupervised domain adaptation. However, this approach is still challenged by the large pivot detection problem that should be solved, and by the inherent instability of LSTMs. In this paper we propose a Task Refinement Learning (TRL) approach, in order to solve these problems. Our algorithms iteratively train the PBLM model, gradually increasing the information exposed about each pivot. TRL-PBLM achieves stateof- the-art accuracy in six domain adaptation setups for sentiment classification. Moreover, it is much more stable than plain PBLM across model configurations, making the model much better fitted for practical use.

Optimal Transport-based Alignment of Learned Character Representations for String Similarity

- https://www.aclweb.org/anthology/P19-1592/

String similarity models are vital for record linkage, entity resolution, and search. In this work, we present STANCE–a learned model for computing the similarity of two strings. Our approach encodes the characters of each string, aligns the encodings using Sinkhorn Iteration (alignment is posed as an instance of optimal transport) and scores the alignment with a convolutional neural network. We evaluate STANCE's ability to detect whether two strings can refer to the same entity–a task we term alias detection. We construct five new alias detection datasets (and make them publicly available). We show that STANCE (or one of its variants) outperforms both state-of-the-art and classic, parameter-free similarity models on four of the five datasets. We also demonstrate STANCE's ability to improve downstream tasks by applying it to an instance of cross-document coreference and show that it leads to a 2.8 point improvement in B^3 F1 over the previous state-of-the-art approach.

The Referential Reader: A Recurrent Entity Network for Anaphora Resolution

- https://www.aclweb.org/anthology/P19-1593/

We present a new architecture for storing and accessing entity mentions during online text processing. While reading the text, entity references are identified, and may be stored by either updating or overwriting a cell in a fixed-length memory. The update operation implies coreference with the other mentions that are stored in the same cell; the overwrite operation causes these mentions to be forgotten. By encoding the memory operations as differentiable gates, it is possible to train the model end-to-end, using both a supervised anaphora resolution objective as well as a supplementary language modeling objective. Evaluation on a dataset of pronoun-name anaphora demonstrates strong performance with purely incremental text processing.

Interpolated Spectral NGram Language Models

- https://www.aclweb.org/anthology/P19-1594/

Spectral models for learning weighted non-deterministic automata have nice theoretical and algorithmic properties. Despite this, it has been challenging to obtain competitive results in language modeling tasks, for two main reasons. First, in order to capture long-range dependencies of the data, the method must use statistics from long substrings, which results in very large matrices that are difficult to decompose. The second is that the loss function behind spectral learning,

based on moment matching, differs from the probabilistic metrics used to evaluate language models. In this work we employ a technique for scaling up spectral learning, and use interpolated predictions that are optimized to maximize perplexity. Our experiments in character-based language modeling show that our method matches the performance of state-of-the-art ngram models, while being very fast to train.

BAM! Born-Again Multi-Task Networks for Natural Language Understanding

- https://www.aclweb.org/anthology/P19-1595/

It can be challenging to train multi-task neural networks that outperform or even match their single-task counterparts. To help address this, we propose using knowledge distillation where single-task models teach a multi-task model. We enhance this training with teacher annealing, a novel method that gradually transitions the model from distillation to supervised learning, helping the multi-task model surpass its single-task teachers. We evaluate our approach by multi-task fine-tuning BERT on the GLUE benchmark. Our method consistently improves over standard single-task and multi-task training.

Curate and Generate: A Corpus and Method for Joint Control of Semantics and Style in Neural NLG

- https://www.aclweb.org/anthology/P19-1596/

Neural natural language generation (NNLG) from structured meaning representations has become increasingly popular in recent years. While we have seen progress with generating syntactically correct utterances that preserve semantics, various shortcomings of NNLG systems are clear: new tasks require new training data which is not available or straightforward to acquire, and model outputs are simple and may be dull and repetitive. This paper addresses these two critical challenges in NNLG by: (1) scalably (and at no cost) creating training datasets of parallel meaning representations and reference texts with rich style markup by using data from freely available and naturally descriptive user reviews, and (2) systematically exploring how the style markup enables joint control of semantic and stylistic aspects of neural model output. We present YelpNLG, a corpus of 300,000 rich, parallel meaning representations and highly stylistically varied reference texts spanning different restaurant attributes, and describe a novel methodology that can be scalably reused to generate NLG datasets for other domains. The experiments show that the models control important aspects, including lexical choice of adjectives, output length, and sentiment, allowing the models to successfully hit multiple style targets without sacrificing semantics.

Automated Chess Commentator Powered by Neural Chess Engine

- https://www.aclweb.org/anthology/P19-1597/

In this paper, we explore a new approach for automated chess commentary generation, which aims to generate chess commentary texts in different categories (e.g., description, comparison, planning, etc.). We introduce a neural chess engine into text generation models to help with encoding boards, predicting moves, and analyzing situations. By jointly training the neural chess engine and the generation models for different categories, the models become more effective. We conduct experiments on 5 categories in a benchmark Chess Commentary dataset and achieve inspiring results in both automatic and human evaluations.

Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling

- https://www.aclweb.org/anthology/P19-1598/

Modeling human language requires the ability to not only generate fluent text but also encode factual knowledge. However, traditional language models are only capable of remembering facts seen at training time, and often have difficulty recalling them. To address this, we introduce the knowledge graph language model (KGLM), a neural language model with mechanisms for selecting and copying facts from a knowledge graph that are relevant to the context. These mechanisms enable the model to render information it has never seen before, as well as generate out-of-vocabulary tokens. We also introduce the Linked WikiText-2 dataset, a corpus of annotated text aligned to the Wikidata knowledge graph whose contents (roughly) match the popular WikiText-2 benchmark. In experiments, we demonstrate that the KGLM achieves significantly better performance than a strong baseline language model. We additionally compare different language model's ability to complete sentences requiring factual knowledge, showing that the KGLM outperforms even very large language models in generating facts.

Controllable Paraphrase Generation with a Syntactic Exemplar

- https://www.aclweb.org/anthology/P19-1599/

Prior work on controllable text generation usually assumes that the controlled attribute can take on one of a small set of values known a priori. In this work, we propose a novel task, where the syntax of a generated sentence is controlled

rather by a sentential exemplar. To evaluate quantitatively with standard metrics, we create a novel dataset with human annotations. We also develop a variational model with a neural module specifically designed for capturing syntactic knowledge and several multitask training objectives to promote disentangled representation learning. Empirically, the proposed model is observed to achieve improvements over baselines and learn to capture desirable characteristics.

Towards Comprehensive Description Generation from Factual Attribute-value Tables

- https://www.aclweb.org/anthology/P19-1600/

The comprehensive descriptions for factual attribute-value tables, which should be accurate, informative and loyal, can be very helpful for end users to understand the structured data in this form. However previous neural generators might suffer from key attributes missing, less informative and groundless information problems, which impede the generation of high-quality comprehensive descriptions for tables. To relieve these problems, we first propose force attention (FA) method to encourage the generator to pay more attention to the uncovered attributes to avoid potential key attributes missing. Furthermore, we propose reinforcement learning for information richness to generate more informative as well as more loyal descriptions for tables. In our experiments, we utilize the widely used WIKIBIO dataset as a benchmark. Besides, we create WB-filter based on WIKIBIO to test our model in the simulated user-oriented scenarios, in which the generated descriptions should accord with particular user interests. Experimental results show that our model outperforms the state-of-the-art baselines on both automatic and human evaluation.

Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation

- https://www.aclweb.org/anthology/P19-1601/

Disentangling the content and style in the latent space is prevalent in unpaired text style transfer. However, two major issues exist in most of the current neural models. 1) It is difficult to completely strip the style information from the semantics for a sentence. 2) The recurrent neural network (RNN) based encoder and decoder, mediated by the latent representation, cannot well deal with the issue of the long-term dependency, resulting in poor preservation of non-stylistic semantic content. In this paper, we propose the Style Transformer, which makes no assumption about the latent representation of source sentence and equips the power of attention mechanism in Transformer to achieve better style transfer and better content preservation.

Generating Sentences from Disentangled Syntactic and Semantic Spaces

- https://www.aclweb.org/anthology/P19-1602/

Variational auto-encoders (VAEs) are widely used in natural language generation due to the regularization of the latent space. However, generating sentences from the continuous latent space does not explicitly model the syntactic information. In this paper, we propose to generate sentences from disentangled syntactic and semantic spaces. Our proposed method explicitly models syntactic information in the VAE's latent space by using the linearized tree sequence, leading to better performance of language generation. Additionally, the advantage of sampling in the disentangled syntactic and semantic latent spaces enables us to perform novel applications, such as the unsupervised paraphrase generation and syntax transfer generation. Experimental results show that our proposed model achieves similar or better performance in various tasks, compared with state-of-the-art related work.

Learning to Control the Fine-grained Sentiment for Story Ending Generation

- https://www.aclweb.org/anthology/P19-1603/

Automatic story ending generation is an interesting and challenging task in natural language generation. Previous studies are mainly limited to generate coherent, reasonable and diversified story endings, and few works focus on controlling the sentiment of story endings. This paper focuses on generating a story ending which meets the given fine-grained sentiment intensity. There are two major challenges to this task. First is the lack of story corpus which has fine-grained sentiment labels. Second is the difficulty of explicitly controlling sentiment intensity when generating endings. Therefore, we propose a generic and novel framework which consists of a sentiment analyzer and a sentimental generator, respectively addressing the two challenges. The sentiment analyzer adopts a series of methods to acquire sentiment intensities of the story dataset. The sentimental generator introduces the sentiment intensity into decoder via a Gaussian Kernel Layer to control the sentiment of the output. To the best of our knowledge, this is the first endeavor to control the fine-grained sentiment for story ending generation without manually annotating sentiment labels. Experiments show that our proposed framework can generate story endings which are not only more coherent and fluent but also able to meet the given sentiment intensity better.

Self-Attention Architectures for Answer-Agnostic Neural Question Generation

- https://www.aclweb.org/anthology/P19-1604/

Neural architectures based on self-attention, such as Transformers, recently attracted interest from the research community, and obtained significant improvements over the state of the art in several tasks. We explore how Transformers can be adapted to the task of Neural Question Generation without constraining the model to focus on a specific answer passage. We study the effect of several strategies to deal with out-of-vocabulary words such as copy mechanisms, placeholders, and contextual word embeddings. We report improvements obtained over the state-of-the-art on the SQuAD dataset according to automated metrics (BLEU, ROUGE), as well as qualitative human assessments of the system outputs.

Unsupervised Paraphrasing without Translation

- https://www.aclweb.org/anthology/P19-1605/

Paraphrasing is an important task demonstrating the ability to semantic content from its surface form. Recent literature on automatic paraphrasing is dominated by methods leveraging machine translation as an intermediate step. This contrasts with humans, who can paraphrase without necessarily being bilingual. This work proposes to learn paraphrasing models only from a monolingual corpus. To that end, we propose a residual variant of vector-quantized variational auto-encoder. Our experiments consider paraphrase identification, and paraphrasing for training set augmentation, comparing to supervised and unsupervised translation-based approaches. Monolingual paraphrasing is shown to outperform unsupervised translation in all contexts. The comparison with supervised MT is more mixed: monolingual paraphrasing is interesting for identification and augmentation but supervised MT is superior for generation.

Storyboarding of Recipes: Grounded Contextual Generation

- https://www.aclweb.org/anthology/P19-1606/

Information need of humans is essentially multimodal in nature, enabling maximum exploitation of situated context. We introduce a dataset for sequential procedural (how-to) text generation from images in cooking domain. The dataset consists of 16,441 cooking recipes with 160,479 photos associated with different steps. We setup a baseline motivated by the best performing model in terms of human evaluation for the Visual Story Telling (ViST) task. In addition, we introduce two models to incorporate high level structure learnt by a Finite

State Machine (FSM) in neural sequential generation process by: (1) Scaffolding Structure in Decoder (SSiD) (2) Scaffolding Structure in Loss (SSiL). Our best performing model (SSiL) achieves a METEOR score of 0.31, which is an improvement of 0.6 over the baseline model. We also conducted human evaluation of the generated grounded recipes, which reveal that 61% found that our proposed (SSiL) model is better than the baseline model in terms of overall recipes. We also discuss analysis of the output highlighting key important NLP issues for prospective directions.

Negative Lexically Constrained Decoding for Paraphrase Generation

- https://www.aclweb.org/anthology/P19-1607/

Paraphrase generation can be regarded as monolingual translation. Unlike bilingual machine translation, paraphrase generation rewrites only a limited portion of an input sentence. Hence, previous methods based on machine translation often perform conservatively to fail to make necessary rewrites. To solve this problem, we propose a neural model for paraphrase generation that first identifies words in the source sentence that should be paraphrased. Then, these words are paraphrased by the negative lexically constrained decoding that avoids outputting these words as they are. Experiments on text simplification and formality transfer show that our model improves the quality of paraphrasing by making necessary rewrites to an input sentence.

Large-Scale Transfer Learning for Natural Language Generation

- https://www.aclweb.org/anthology/P19-1608/

Large-scale pretrained language models define state of the art in natural language processing, achieving outstanding performance on a variety of tasks. We study how these architectures can be applied and adapted for natural language generation, comparing a number of architectural and training schemes. We focus in particular on open-domain dialog as a typical high entropy generation task, presenting and comparing different architectures for adapting pretrained models with state of the art results.

Automatic Grammatical Error Correction for Sequence-to-sequence Text Generation: An Empirical Study

- https://www.aclweb.org/anthology/P19-1609/

Sequence-to-sequence (seq2seq) models have achieved tremendous success in text generation tasks. However, there is no guarantee that they can always generate sentences without grammatical errors. In this paper, we present a preliminary empirical study on whether and how much automatic grammatical error correction can help improve seq2seq text generation. We conduct experiments across various seq2seq text generation tasks including machine translation, formality style transfer, sentence compression and simplification. Experiments show the state-of-the-art grammatical error correction system can improve the grammaticality of generated text and can bring task-oriented improvements in the tasks where target sentences are in a formal style.

Improving the Robustness of Question Answering Systems to Question Paraphrasing

- https://www.aclweb.org/anthology/P19-1610/

Despite the advancement of question answering (QA) systems and rapid improvements on held-out test sets, their generalizability is a topic of concern. We explore the robustness of QA models to question paraphrasing by creating two test sets consisting of paraphrased SQuAD questions. Paraphrased questions from the first test set are very similar to the original questions designed to test QA models' over-sensitivity, while questions from the second test set are paraphrased using context words near an incorrect answer candidate in an attempt to confuse QA models. We show that both paraphrased test sets lead to significant decrease in performance on multiple state-of-the-art QA models. Using a neural paraphrasing model trained to generate multiple paraphrased questions for a given source question and a set of paraphrase suggestions, we propose a data augmentation approach that requires no human intervention to re-train the models for improved robustness to question paraphrasing.

RankQA: Neural Question Answering with Answer Re-Ranking

- https://www.aclweb.org/anthology/P19-1611/

The conventional paradigm in neural question answering (QA) for narrative content is limited to a two-stage process: first, relevant text passages are retrieved and, subsequently, a neural network for machine comprehension extracts the likeliest answer. However, both stages are largely isolated in the status quo and, hence, information from the two phases is never properly fused. In contrast, this work proposes RankQA: RankQA extends the conventional two-stage process in neural QA with a third stage that performs an additional answer re-ranking. The re-ranking leverages different features that are directly extracted from the QA pipeline, i.e., a combination of retrieval and comprehension features. While

our intentionally simple design allows for an efficient, data-sparse estimation, it nevertheless outperforms more complex QA systems by a significant margin: in fact, RankQA achieves state-of-the-art performance on 3 out of 4 benchmark datasets. Furthermore, its performance is especially superior in settings where the size of the corpus is dynamic. Here the answer re-ranking provides an effective remedy against the underlying noise-information trade-off due to a variable corpus size. As a consequence, RankQA represents a novel, powerful, and thus challenging baseline for future research in content-based QA.

Latent Retrieval for Weakly Supervised Open Domain Question Answering

- https://www.aclweb.org/anthology/P19-1612/

Recent work on open domain question answering (QA) assumes strong supervision of the supporting evidence and/or assumes a blackbox information retrieval (IR) system to retrieve evidence candidates. We argue that both are suboptimal, since gold evidence is not always available, and QA is fundamentally different from IR. We show for the first time that it is possible to jointly learn the retriever and reader from question-answer string pairs and without any IR system. In this setting, evidence retrieval from all of Wikipedia is treated as a latent variable. Since this is impractical to learn from scratch, we pre-train the retriever with an Inverse Cloze Task. We evaluate on open versions of five QA datasets. On datasets where the questioner already knows the answer, a traditional IR system such as BM25 is sufficient. On datasets where a user is genuinely seeking an answer, we show that learned retrieval is crucial, outperforming BM25 by up to 19 points in exact match.

Multi-hop Reading Comprehension through Question Decomposition and Rescoring

- https://www.aclweb.org/anthology/P19-1613/

Multi-hop Reading Comprehension (RC) requires reasoning and aggregation across several paragraphs. We propose a system for multi-hop RC that decomposes a compositional question into simpler sub-questions that can be answered by off-the-shelf single-hop RC models. Since annotations for such decomposition are expensive, we recast subquestion generation as a span prediction problem and show that our method, trained using only 400 labeled examples, generates sub-questions that are as effective as human-authored sub-questions. We also introduce a new global rescoring approach that considers each decomposition (i.e. the sub-questions and their answers) to select the best final answer, greatly improving overall performance. Our experiments on HotpotQA show that this approach achieves the state-of-the-art results, while providing explainable evidence for its decision making in the form of sub-questions.

Combining Knowledge Hunting and Neural Language Models to Solve the Winograd Schema Challenge

- https://www.aclweb.org/anthology/P19-1614/

Winograd Schema Challenge (WSC) is a pronoun resolution task which seems to require reasoning with commonsense knowledge. The needed knowledge is not present in the given text. Automatic extraction of the needed knowledge is a bottleneck in solving the challenge. The existing state-of-the-art approach uses the knowledge embedded in their pre-trained language model. However, the language models only embed part of the knowledge, the ones related to frequently co-existing concepts. This limits the performance of such models on the WSC problems. In this work, we build-up on the language model based methods and augment them with a commonsense knowledge hunting (using automatic extraction from text) module and an explicit reasoning module. Our end-to-end system built in such a manner improves on the accuracy of two of the available language model based approaches by 5.53% and 7.7% respectively. Overall our system achieves the state-of-the-art accuracy of 71.06% on the WSC dataset, an improvement of 7.36% over the previous best.

Careful Selection of Knowledge to Solve Open Book Question Answering

- https://www.aclweb.org/anthology/P19-1615/

Open book question answering is a type of natural language based QA (NLQA) where questions are expected to be answered with respect to a given set of open book facts, and common knowledge about a topic. Recently a challenge involving such QA, OpenBookQA, has been proposed. Unlike most other NLQA that focus on linguistic understanding, OpenBookQA requires deeper reasoning involving linguistic understanding as well as reasoning with common knowledge. In this paper we address QA with respect to the OpenBookQA dataset and combine state of the art language models with abductive information retrieval (IR), information gain based re-ranking, passage selection and weighted scoring to achieve 72.0% accuracy, an 11.6% improvement over the current state of the art.

Learning Representation Mapping for Relation Detection in Knowledge Base Question Answering

- https://www.aclweb.org/anthology/P19-1616/

Relation detection is a core step in many natural language process applications including knowledge base question answering. Previous efforts show that single-fact questions could be answered with high accuracy. However, one critical

problem is that current approaches only get high accuracy for questions whose relations have been seen in the training data. But for unseen relations, the performance will drop rapidly. The main reason for this problem is that the representations for unseen relations are missing. In this paper, we propose a simple mapping method, named representation adapter, to learn the representation mapping for both seen and unseen relations based on previously learned relation embedding. We employ the adversarial objective and the reconstruction objective to improve the mapping performance. We re-organize the popular SimpleQuestion dataset to reveal and evaluate the problem of detecting unseen relations. Experiments show that our method can greatly improve the performance of unseen relations while the performance for those seen part is kept comparable to the state-of-the-art.

Dynamically Fused Graph Network for Multi-hop Reasoning

- https://www.aclweb.org/anthology/P19-1617/

Text-based question answering (TBQA) has been studied extensively in recent years. Most existing approaches focus on finding the answer to a question within a single paragraph. However, many difficult questions require multiple supporting evidence from scattered text among two or more documents. In this paper, we propose Dynamically Fused Graph Network (DFGN), a novel method to answer those questions requiring multiple scattered evidence and reasoning over them. Inspired by human's step-by-step reasoning behavior, DFGN includes a dynamic fusion layer that starts from the entities mentioned in the given query, explores along the entity graph dynamically built from the text, and gradually finds relevant supporting entities from the given documents. We evaluate DFGN on HotpotQA, a public TBQA dataset requiring multi-hop reasoning. DFGN achieves competitive results on the public board. Furthermore, our analysis shows DFGN produces interpretable reasoning chains.

NLProlog: Reasoning with Weak Unification for Question Answering in Natural Language

- https://www.aclweb.org/anthology/P19-1618/

Rule-based models are attractive for various tasks because they inherently lead to interpretable and explainable decisions and can easily incorporate prior knowledge. However, such systems are difficult to apply to problems involving natural language, due to its large linguistic variability. In contrast, neural models can cope very well with ambiguity by learning distributed representations of words and their composition from data, but lead to models that are difficult to interpret. In this paper, we describe a model combining neural networks with logic

programming in a novel manner for solving multi-hop reasoning tasks over natural language. Specifically, we propose to use an Prolog prover which we extend to utilize a similarity function over pretrained sentence encoders. We fine-tune the representations for the similarity function via backpropagation. This leads to a system that can apply rule-based reasoning to natural language, and induce domain-specific natural language rules from training data. We evaluate the proposed system on two different question answering tasks, showing that it outperforms two baselines – BiDAF (Seo et al., 2016a) and FastQA(Weissenborn et al., 2017) on a subset of the WikiHop corpus and achieves competitive results on the MedHop data set (Welbl et al., 2017).

Modeling Intra-Relation in Math Word Problems with Different Functional Multi-Head Attentions

- https://www.aclweb.org/anthology/P19-1619/

Several deep learning models have been proposed for solving math word problems (MWPs) automatically. Although these models have the ability to capture features without manual efforts, their approaches to capturing features are not specifically designed for MWPs. To utilize the merits of deep learning models with simultaneous consideration of MWPs' specific features, we propose a group attention mechanism to extract global features, quantity-related features, quantity-pair features and question-related features in MWPs respectively. The experimental results show that the proposed approach performs significantly better than previous state-of-the-art methods, and boost performance from 66.9% to 69.5% on Math23K with training-test split, from 65.8% to 66.9% on Math23K with 5-fold cross-validation and from 69.2% to 76.1% on MAWPS.

Synthetic QA Corpora Generation with Roundtrip Consistency

- https://www.aclweb.org/anthology/P19-1620/

We introduce a novel method of generating synthetic question answering corpora by combining models of question generation and answer extraction, and by filtering the results to ensure roundtrip consistency. By pretraining on the resulting corpora we obtain significant improvements on SQuAD2 and NQ, establishing a new state-of-the-art on the latter. Our synthetic data generation models, for both question generation and answer extraction, can be fully reproduced by finetuning a publicly available BERT model on the extractive subsets of SQuAD2 and NQ. We also describe a more powerful variant that does full sequence-to-sequence pretraining for question generation, obtaining exact match and F1 at less than 0.1% and 0.4% from human performance on SQuAD2.

Are Red Roses Red? Evaluating Consistency of Question-Answering Models

- https://www.aclweb.org/anthology/P19-1621/

Although current evaluation of question-answering systems treats predictions in isolation, we need to consider the relationship between predictions to measure true understanding. A model should be penalized for answering "no" to "Is the rose red?" if it answers "red" to "What color is the rose?". We propose a method to automatically extract such implications for instances from two QA datasets, VQA and SQuAD, which we then use to evaluate the consistency of models. Human evaluation shows these generated implications are well formed and valid. Consistency evaluation provides crucial insights into gaps in existing models, while retraining with implication-augmented data improves consistency on both synthetic and human-generated implications.

MC^2: Multi-perspective Convolutional Cube for Conversational Machine Reading Comprehension

- https://www.aclweb.org/anthology/P19-1622/

Conversational machine reading comprehension (CMRC) extends traditional single-turn machine reading comprehension (MRC) by multi-turn interactions, which requires machines to consider the history of conversation. Most of models simply combine previous questions for conversation understanding and only employ recurrent neural networks (RNN) for reasoning. To comprehend context profoundly and efficiently from different perspectives, we propose a novel neural network model, Multi-perspective Convolutional Cube (MC^2). We regard each conversation as a cube. 1D and 2D convolutions are integrated with RNN in our model. To avoid models previewing the next turn of conversation, we also extend causal convolution partially to 2D. Experiments on the Conversational Question Answering (CoQA) dataset show that our model achieves state-of-the-art results.

Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach

- https://www.aclweb.org/anthology/P19-1623/

While neural machine translation (NMT) has achieved remarkable success, NMT systems are prone to make word omission errors. In this work, we propose a contrastive learning approach to reducing word omission errors in NMT. The basic idea is to enable the NMT model to assign a higher probability to a ground-truth translation and a lower probability to an erroneous translation, which is automatically constructed from the ground-truth translation by omitting words.

We design different types of negative examples depending on the number of omitted words, word frequency, and part of speech. Experiments on Chinese-to-English, German-to-English, and Russian-to-English translation tasks show that our approach is effective in reducing word omission errors and achieves better translation performance than three baseline methods.

Exploiting Sentential Context for Neural Machine Translation

- https://www.aclweb.org/anthology/P19-1624/

In this work, we present novel approaches to exploit sentential context for neural machine translation (NMT). Specifically, we show that a shallow sentential context extracted from the top encoder layer only, can improve translation performance via contextualizing the encoding representations of individual words. Next, we introduce a deep sentential context, which aggregates the sentential context representations from all of the internal layers of the encoder to form a more comprehensive context representation. Experimental results on the WMT14 English-German and English-French benchmarks show that our model consistently improves performance over the strong Transformer model, demonstrating the necessity and effectiveness of exploiting sentential context for NMT.

Wetin dey with these comments? Modeling Sociolinguistic Factors Affecting Code-switching Behavior in Nigerian Online Discussions

- https://www.aclweb.org/anthology/P19-1625/

Multilingual individuals code switch between languages as a part of a complex communication process. However, most computational studies have examined only one or a handful of contextual factors predictive of switching. Here, we examine Naija-English code switching in a rich contextual environment to understand the social and topical factors eliciting a switch. We introduce a new corpus of 330K articles and accompanying 389K comments labeled for code switching behavior. In modeling whether a comment will switch, we show that topic-driven variation, tribal affiliation, emotional valence, and audience design all play complementary roles in behavior.

Accelerating Sparse Matrix Operations in Neural Networks on Graphics Processing Units

- https://www.aclweb.org/anthology/P19-1626/

Graphics Processing Units (GPUs) are commonly used to train and evaluate neural networks efficiently. While previous work in deep learning has focused on accelerating operations on dense matrices/tensors on GPUs, efforts have concentrated on operations involving sparse data structures. Operations using sparse structures are common in natural language models at the input and output layers, because these models operate on sequences over discrete alphabets. We present two new GPU algorithms: one at the input layer, for multiplying a matrix by a few-hot vector (generalizing the more common operation of multiplication by a one-hot vector) and one at the output layer, for a fused softmax and top-N selection (commonly used in beam search). Our methods achieve speedups over state-of-the-art parallel GPU baselines of up to 7x and 50x, respectively. We also illustrate how our methods scale on different GPU architectures.

An Automated Framework for Fast Cognate Detection and Bayesian Phylogenetic Inference in Computational Historical Linguistics

- https://www.aclweb.org/anthology/P19-1627/

We present a fully automated workflow for phylogenetic reconstruction on large datasets, consisting of two novel methods, one for fast detection of cognates and one for fast Bayesian phylogenetic inference. Our results show that the methods take less than a few minutes to process language families that have so far required large amounts of time and computational power. Moreover, the cognates and the trees inferred from the method are quite close, both to gold standard cognate judgments and to expert language family trees. Given its speed and ease of application, our framework is specifically useful for the exploration of very large datasets in historical linguistics.

Sentence Centrality Revisited for Unsupervised Summarization

- https://www.aclweb.org/anthology/P19-1628/

Single document summarization has enjoyed renewed interest in recent years thanks to the popularity of neural network models and the availability of large-scale datasets. In this paper we develop an unsupervised approach arguing that it is unrealistic to expect large-scale and high-quality training data to be available or created for different types of summaries, domains, or languages. We revisit a popular graph-based ranking algorithm and modify how node (aka sentence) centrality is computed in two ways: (a) we employ BERT, a state-of-the-art neural representation learning model to better capture sentential meaning and (b) we build graphs with directed edges arguing that the contribution of any

two nodes to their respective centrality is influenced by their relative position in a document. Experimental results on three news summarization datasets representative of different languages and writing styles show that our approach outperforms strong baselines by a wide margin.

Discourse Representation Parsing for Sentences and Documents

- https://www.aclweb.org/anthology/P19-1629/

We introduce a novel semantic parsing task based on Discourse Representation Theory (DRT; Kamp and Reyle 1993). Our model operates over Discourse Representation Tree Structures which we formally define for sentences and documents. We present a general framework for parsing discourse structures of arbitrary length and granularity. We achieve this with a neural model equipped with a supervised hierarchical attention mechanism and a linguistically-motivated copy strategy. Experimental results on sentence- and document-level benchmarks show that our model outperforms competitive baselines by a wide margin.

Inducing Document Structure for Aspect-based Summarization

- https://www.aclweb.org/anthology/P19-1630/

Automatic summarization is typically treated as a 1-to-1 mapping from document to summary. Documents such as news articles, however, are structured and often cover multiple topics or aspects; and readers may be interested in only some of them. We tackle the task of aspect-based summarization, where, given a document and a target aspect, our models generate a summary centered around the aspect. We induce latent document structure jointly with an ive summarization objective, and train our models in a scalable synthetic setup. In addition to improvements in summarization over topic-agnostic baselines, we demonstrate the benefit of the learnt document structure: we show that our models (a) learn to accurately segment documents by aspect; (b) can leverage the structure to produce both ive and extractive aspect-based summaries; and (c) that structure is particularly advantageous for summarizing long documents. All results transfer from synthetic training documents to natural news articles from CNN/Daily Mail and RCV1.

Incorporating Priors with Feature Attribution on Text Classification

- https://www.aclweb.org/anthology/P19-1631/

Feature attribution methods, proposed recently, help users interpret the predictions of complex models. Our approach integrates feature attributions into the objective function to allow machine learning practitioners to incorporate priors in model building. To demonstrate the effectiveness our technique, we apply it to two tasks: (1) mitigating unintended bias in text classifiers by neutralizing identity terms; (2) improving classifier performance in scarce data setting by forcing model to focus on toxic terms. Our approach adds an L2 distance loss between feature attributions and task-specific prior values to the objective. Our experiments show that i) a classifier trained with our technique reduces undesired model biases without a tradeoff on the original task; ii) incorporating prior helps model performance in scarce data settings.

Matching Article Pairs with Graphical Decomposition and Convolutions

- https://www.aclweb.org/anthology/P19-1632/

Identifying the relationship between two articles, e.g., whether two articles published from different sources describe the same breaking news, is critical to many document understanding tasks. Existing approaches for modeling and matching sentence pairs do not perform well in matching longer documents, which embody more complex interactions between the enclosed entities than a sentence does. To model article pairs, we propose the Concept Interaction Graph to represent an article as a graph of concepts. We then match a pair of articles by comparing the sentences that enclose the same concept vertex through a series of encoding techniques, and aggregate the matching signals through a graph convolutional network. To facilitate the evaluation of long article matching, we have created two datasets, each consisting of about 30K pairs of breaking news articles covering diverse topics in the open domain. Extensive evaluations of the proposed methods on the two datasets demonstrate significant improvements over a wide range of state-of-the-art methods for natural language matching.

Hierarchical Transfer Learning for Multi-label Text Classification

- https://www.aclweb.org/anthology/P19-1633/

Multi-Label Hierarchical Text Classification (MLHTC) is the task of categorizing documents into one or more topics organized in an hierarchical taxonomy. MLHTC can be formulated by combining multiple binary classification problems with an independent classifier for each category. We propose a novel transfer learning based strategy, HTrans, where binary classifiers at lower levels in the hierarchy are initialized using parameters of the parent classifier and fine-tuned on the child category classification task. In HTrans, we use a Gated Recurrent

Unit (GRU)-based deep learning architecture coupled with attention. Compared to binary classifiers trained from scratch, our HTrans approach results in significant improvements of 1% on micro-F1 and 3% on macro-F1 on the RCV1 dataset. Our experiments also show that binary classifiers trained from scratch are significantly better than single multi-label models.

Bias Analysis and Mitigation in the Evaluation of Authorship Verification

- https://www.aclweb.org/anthology/P19-1634/

The PAN series of shared tasks is well known for its continuous and high quality research in the field of digital text forensics. Among others, PAN contributions include original corpora, tailored benchmarks, and standardized experimentation platforms. In this paper we review, theoretically and practically, the authorship verification task and conclude that the underlying experiment design cannot guarantee pushing forward the state of the art—in fact, it allows for top benchmarking with a surprisingly straightforward approach. In this regard, we present a "Basic and Fairly Flawed" (BAFF) authorship verifier that is on a par with the best approaches submitted so far, and that illustrates sources of bias that should be eliminated. We pinpoint these sources in the evaluation chain and present a refined authorship corpus as effective countermeasure.

Numeracy-600K: Learning Numeracy for Detecting Exaggerated Information in Market Comments

- https://www.aclweb.org/anthology/P19-1635/

In this paper, we attempt to answer the question of whether neural network models can learn numeracy, which is the ability to predict the magnitude of a numeral at some specific position in a text description. A large benchmark dataset, called Numeracy-600K, is provided for the novel task. We explore several neural network models including CNN, GRU, BiGRU, CRNN, CNN-capsule, GRU-capsule, and BiGRU-capsule in the experiments. The results show that the BiGRU model gets the best micro-averaged F1 score of 80.16%, and the GRU-capsule model gets the best macro-averaged F1 score of 64.71%. Besides discussing the challenges through comprehensive experiments, we also present an important application scenario, i.e., detecting exaggerated information, for the task.

Large-Scale Multi-Label Text Classification on EU Legislation

- https://www.aclweb.org/anthology/P19-1636/

We consider Large-Scale Multi-Label Text Classification (LMTC) in the legal domain. We release a new dataset of 57k legislative documents from EUR-LEX, annotated with 4.3k EUROVOC labels, which is suitable for LMTC, few- and zero-shot learning. Experimenting with several neural classifiers, we show that BIGRUs with label-wise attention perform better than other current state of the art methods. Domain-specific WORD2VEC and context-sensitive ELMO embeddings further improve performance. We also find that considering only particular zones of the documents is sufficient. This allows us to bypass BERT's maximum text length limit and fine-tune BERT, obtaining the best results in all but zero-shot learning cases.

Why Didn't You Listen to Me? Comparing User Control of Human-in-the-Loop Topic Models

- https://www.aclweb.org/anthology/P19-1637/

To address the lack of comparative evaluation of Human-in-the-Loop Topic Modeling (HLTM) systems, we implement and evaluate three contrasting HLTM modeling approaches using simulation experiments. These approaches extend previously proposed frameworks, including constraints and informed prior-based methods. Users should have a sense of control in HLTM systems, so we propose a control metric to measure whether refinement operations' results match users' expectations. Informed prior-based methods provide better control than constraints, but constraints yield higher quality topics.

Encouraging Paragraph Embeddings to Remember Sentence Identity Improves Classification

- https://www.aclweb.org/anthology/P19-1638/

While paragraph embedding models are remarkably effective for downstream classification tasks, what they learn and encode into a single vector remains opaque. In this paper, we investigate a state-of-the-art paragraph embedding method proposed by Zhang et al. (2017) and discover that it cannot reliably tell whether a given sentence occurs in the input paragraph or not. We formulate a sentence content task to probe for this basic linguistic property and find that even a much simpler bag-of-words method has no trouble solving it. This result motivates us to replace the reconstruction-based objective of Zhang et al. (2017) with our sentence content probe objective in a semi-supervised setting. Despite its simplicity, our objective improves over paragraph reconstruction in terms of (1) downstream classification accuracies on benchmark datasets, (2) faster training, and (3) better generalization ability.

A Multi-Task Architecture on Relevance-based Neural Query Translation

- https://www.aclweb.org/anthology/P19-1639/

We describe a multi-task learning approach to train a Neural Machine Translation (NMT) model with a Relevance-based Auxiliary Task (RAT) for search query translation. The translation process for Cross-lingual Information Retrieval (CLIR) task is usually treated as a black box and it is performed as an independent step. However, an NMT model trained on sentence-level parallel data is not aware of the vocabulary distribution of the retrieval corpus. We address this problem and propose a multi-task learning architecture that achieves 16% improvement over a strong baseline on Italian-English query-document dataset. We show using both quantitative and qualitative analysis that our model generates balanced and precise translations with the regularization effect it achieves from multi-task learning paradigm.

Topic Modeling with Wasserstein Autoencoders

- https://www.aclweb.org/anthology/P19-1640/

We propose a novel neural topic model in the Wasserstein autoencoders (WAE) framework. Unlike existing variational autoencoder based models, we directly enforce Dirichlet prior on the latent document-topic vectors. We exploit the structure of the latent space and apply a suitable kernel in minimizing the Maximum Mean Discrepancy (MMD) to perform distribution matching. We discover that MMD performs much better than the Generative Adversarial Network (GAN) in matching high dimensional Dirichlet distribution. We further discover that incorporating randomness in the encoder output during training leads to significantly more coherent topics. To measure the diversity of the produced topics, we propose a simple topic uniqueness metric. Together with the widely used coherence measure NPMI, we offer a more wholistic evaluation of topic quality. Experiments on several real datasets show that our model produces significantly better topics than existing topic models.

Dense Procedure Captioning in Narrated Instructional Videos

- https://www.aclweb.org/anthology/P19-1641/

Understanding narrated instructional videos is important for both research and real-world web applications. Motivated by video dense captioning, we propose a model to generate procedure captions from narrated instructional videos which are a sequence of step-wise clips with description. Previous works on video

dense captioning learn video segments and generate captions without considering transcripts. We argue that transcripts in narrated instructional videos can enhance video representation by providing fine-grained complimentary and semantic textual information. In this paper, we introduce a framework to (1) extract procedures by a cross-modality module, which fuses video content with the entire transcript; and (2) generate captions by encoding video frames as well as a snippet of transcripts within each extracted procedure. Experiments show that our model can achieve state-of-the-art performance in procedure extraction and captioning, and the ablation studies demonstrate that both the video frames and the transcripts are important for the task.

Latent Variable Model for Multi-modal Translation

- https://www.aclweb.org/anthology/P19-1642/

In this work, we propose to model the interaction between visual and textual features for multi-modal neural machine translation (MMT) through a latent variable model. This latent variable can be seen as a multi-modal stochastic embedding of an image and its description in a foreign language. It is used in a target-language decoder and also to predict image features. Importantly, our model formulation utilises visual and textual inputs during training but does not require that images be available at test time. We show that our latent variable MMT formulation improves considerably over strong baselines, including a multi-task learning approach (Elliott and Kadar, 2017) and a conditional variational auto-encoder approach (Toyama et al., 2016). Finally, we show improvements due to (i) predicting image features in addition to only conditioning on them, (ii) imposing a constraint on the KL term to promote models with non-negligible mutual information between inputs and latent variable, and (iii) by training on additional target-language image descriptions (i.e. synthetic data).

Identifying Visible Actions in Lifestyle Vlogs

- https://www.aclweb.org/anthology/P19-1643/

We consider the task of identifying human actions visible in online videos. We focus on the widely spread genre of lifestyle vlogs, which consist of videos of people performing actions while verbally describing them. Our goal is to identify if actions mentioned in the speech description of a video are visually present. We construct a dataset with crowdsourced manual annotations of visible actions, and introduce a multimodal algorithm that leverages information derived from visual and linguistic clues to automatically infer which actions are visible in a video.

A Corpus for Reasoning about Natural Language Grounded in Photographs

- https://www.aclweb.org/anthology/P19-1644/

We introduce a new dataset for joint reasoning about natural language and images, with a focus on semantic diversity, compositionality, and visual reasoning challenges. The data contains 107,292 examples of English sentences paired with web photographs. The task is to determine whether a natural language caption is true about a pair of photographs. We crowdsource the data using sets of visually rich images and a compare-and-contrast task to elicit linguistically diverse language. Qualitative analysis shows the data requires compositional joint reasoning, including about quantities, comparisons, and relations. Evaluation using state-of-the-art visual reasoning methods shows the data presents a strong challenge.

Learning to Discover, Ground and Use Words with Segmental Neural Language Models

- https://www.aclweb.org/anthology/P19-1645/

We propose a segmental neural language model that combines the generalization power of neural networks with the ability to discover word-like units that are latent in unsegmented character sequences. In contrast to previous segmentation models that treat word segmentation as an isolated task, our model unifies word discovery, learning how words fit together to form sentences, and, by conditioning the model on visual context, how words' meanings ground in representations of nonlinguistic modalities. Experiments show that the unconditional model learns predictive distributions better than character LSTM models, discovers words competitively with nonparametric Bayesian word segmentation models, and that modeling language conditional on visual context improves performance on both.

What Should I Ask? Using Conversationally Informative Rewards for Goal-oriented Visual Dialog.

- https://www.aclweb.org/anthology/P19-1646/

The ability to engage in goal-oriented conversations has allowed humans to gain knowledge, reduce uncertainty, and perform tasks more efficiently. Artificial agents, however, are still far behind humans in having goal-driven conversations. In this work, we focus on the task of goal-oriented visual dialogue, aiming to automatically generate a series of questions about an image with a single objective. This task is challenging since these questions must not only be consistent with a strategy to achieve a goal, but also consider the contextual information

in the image. We propose an end-to-end goal-oriented visual dialogue system, that combines reinforcement learning with regularized information gain. Unlike previous approaches that have been proposed for the task, our work is motivated by the Rational Speech Act framework, which models the process of human inquiry to reach a goal. We test the two versions of our model on the GuessWhat?! dataset, obtaining significant results that outperform the current state-of-the-art models in the task of generating questions to find an undisclosed object in an image.

Symbolic Inductive Bias for Visually Grounded Learning of Spoken Language

- https://www.aclweb.org/anthology/P19-1647/

A widespread approach to processing spoken language is to first automatically transcribe it into text. An alternative is to use an end-to-end approach: recent works have proposed to learn semantic embeddings of spoken language from images with spoken captions, without an intermediate transcription step. We propose to use multitask learning to exploit existing transcribed speech within the end-to-end setting. We describe a three-task architecture which combines the objectives of matching spoken captions with corresponding images, speech with text, and text with images. We show that the addition of the speech/text task leads to substantial performance improvements on image retrieval when compared to training the speech/image task in isolation. We conjecture that this is due to a strong inductive bias transcribed speech provides to the model, and offer supporting evidence for this.

Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog

- https://www.aclweb.org/anthology/P19-1648/

This paper presents a new model for visual dialog, Recurrent Dual Attention Network (ReDAN), using multi-step reasoning to answer a series of questions about an image. In each question-answering turn of a dialog, ReDAN infers the answer progressively through multiple reasoning steps. In each step of the reasoning process, the semantic representation of the question is updated based on the image and the previous dialog history, and the recurrently-refined representation is used for further reasoning in the subsequent step. On the VisDial v1.0 dataset, the proposed ReDAN model achieves a new state-of-the-art of 64.47% NDCG score. Visualization on the reasoning process further demonstrates that ReDAN can locate context-relevant visual and textual clues via iterative refinement, which can lead to the correct answer step-by-step.

Lattice Transformer for Speech Translation

- https://www.aclweb.org/anthology/P19-1649/

Recent advances in sequence modeling have highlighted the strengths of the transformer architecture, especially in achieving state-of-the-art machine translation results. However, depending on the up-stream systems, e.g., speech recognition, or word segmentation, the input to translation system can vary greatly. The goal of this work is to extend the attention mechanism of the transformer to naturally consume the lattice in addition to the traditional sequential input. We first propose a general lattice transformer for speech translation where the input is the output of the automatic speech recognition (ASR) which contains multiple paths and posterior scores. To leverage the extra information from the lattice structure, we develop a novel controllable lattice attention mechanism to obtain latent representations. On the LDC Spanish-English speech translation corpus, our experiments show that lattice transformer generalizes significantly better and outperforms both a transformer baseline and a lattice LSTM. Additionally, we validate our approach on the WMT 2017 Chinese-English translation task with lattice inputs from different BPE segmentations. In this task, we also observe the improvements over strong baselines.

Informative Image Captioning with External Sources of Information

- https://www.aclweb.org/anthology/P19-1650/

An image caption should fluently present the essential information in a given image, including informative, fine-grained entity mentions and the manner in which these entities interact. However, current captioning models are usually trained to generate captions that only contain common object names, thus falling short on an important "informativeness" dimension. We present a mechanism for integrating image information together with fine-grained labels (assumed to be generated by some upstream models) into a caption that describes the image in a fluent and informative manner. We introduce a multimodal, multi-encoder model based on Transformer that ingests both image features and multiple sources of entity labels. We demonstrate that we can learn to control the appearance of these entity labels in the output, resulting in captions that are both fluent and informative.

CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication

- https://www.aclweb.org/anthology/P19-1651/

In this work, we propose a goal-driven collaborative task that combines language, perception, and action. Specifically, we develop a Collaborative image-Drawing game between two agents, called CoDraw. Our game is grounded in a virtual world that contains movable clip art objects. The game involves two players: a Teller and a Drawer. The Teller sees an scene containing multiple clip art pieces in a semantically meaningful configuration, while the Drawer tries to reconstruct the scene on an empty canvas using available clip art pieces. The two players communicate with each other using natural language. We collect the CoDraw dataset of ~10K dialogs consisting of ~138K messages exchanged between human players. We define protocols and metrics to evaluate learned agents in this testbed, highlighting the need for a novel "crosstalk" evaluation condition which pairs agents trained independently on disjoint subsets of the training data. We present models for our task and benchmark them using both fully automated evaluation and by having them play the game live with humans.

Bridging by Word: Image Grounded Vocabulary Construction for Visual Captioning

- https://www.aclweb.org/anthology/P19-1652/

Image Captioning aims at generating a short description for an image. Existing research usually employs the architecture of CNN-RNN that views the generation as a sequential decision-making process and the entire dataset vocabulary is used as decoding space. They suffer from generating high frequent n-gram with irrelevant words. To tackle this problem, we propose to construct an image-grounded vocabulary, based on which, captions are generated with limitation and guidance. In specific, a novel hierarchical structure is proposed to construct the vocabulary incorporating both visual information and relations among words. For generation, we propose a word-aware RNN cell incorporating vocabulary information into the decoding process directly. Reinforce algorithm is employed to train the generator using constraint vocabulary as action space. Experimental results on MS COCO and Flickr30k show the effectiveness of our framework compared to some state-of-the-art models.

Distilling Translations with Visual Awareness

- https://www.aclweb.org/anthology/P19-1653/

Previous work on multimodal machine translation has shown that visual information is only needed in very specific cases, for example in the presence of ambiguous words where the textual context is not sufficient. As a consequence, models tend to learn to ignore this information. We propose a translate-and-refine approach to this problem where images are only used by a second stage decoder. This approach is trained jointly to generate a good first draft translation and to improve over this draft by (i) making better use of the target

language textual context (both left and right-side contexts) and (ii) making use of visual context. This approach leads to the state of the art results. Additionally, we show that it has the ability to recover from erroneous or missing words in the source language.

VIFIDEL: Evaluating the Visual Fidelity of Image Descriptions

- https://www.aclweb.org/anthology/P19-1654/

We address the task of evaluating image description generation systems. We propose a novel image-aware metric for this task: VIFIDEL. It estimates the faithfulness of a generated caption with respect to the content of the actual image, based on the semantic similarity between labels of objects depicted in images and words in the description. The metric is also able to take into account the relative importance of objects mentioned in human reference descriptions during evaluation. Even if these human reference descriptions are not available, VIFIDEL can still reliably evaluate system descriptions. The metric achieves high correlation with human judgments on two well-known datasets and is competitive with metrics that depend on and rely exclusively on human references.

Are You Looking? Grounding to Multiple Modalities in Vision-and-Language Navigation

- https://www.aclweb.org/anthology/P19-1655/

Vision-and-Language Navigation (VLN) requires grounding instructions, such as "turn right and stop at the door", to routes in a visual environment. The actual grounding can connect language to the environment through multiple modalities, e.g. "stop at the door" might ground into visual objects, while "turn right" might rely only on the geometric structure of a route. We investigate where the natural language empirically grounds under two recent state-of-the-art VLN models. Surprisingly, we discover that visual features may actually hurt these models: models which only use route structure, ablating visual features, outperform their visual counterparts in unseen new environments on the benchmark Room-to-Room dataset. To better use all the available modalities, we propose to decompose the grounding procedure into a set of expert models with access to different modalities (including object detections) and ensemble them at prediction time, improving the performance of state-of-the-art models on the VLN task.

Multimodal Transformer for Unaligned Multimodal Language Sequences

- https://www.aclweb.org/anthology/P19-1656/

Human language is often multimodal, which comprehends a mixture of natural language, facial gestures, and acoustic behaviors. However, two major challenges in modeling such multimodal human language time-series data exist: 1) inherent data non-alignment due to variable sampling rates for the sequences from each modality; and 2) long-range dependencies between elements across modalities. In this paper, we introduce the Multimodal Transformer (MulT) to generically address the above issues in an end-to-end manner without explicitly aligning the data. At the heart of our model is the directional pairwise crossmodal attention, which attends to interactions between multimodal sequences across distinct time steps and latently adapt streams from one modality to another. Comprehensive experiments on both aligned and non-aligned multimodal time-series show that our model outperforms state-of-the-art methods by a large margin. In addition, empirical analysis suggests that correlated crossmodal signals are able to be captured by the proposed crossmodal attention mechanism in MulT.

Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports

- https://www.aclweb.org/anthology/P19-1657/

Chest X-Ray (CXR) images are commonly used for clinical screening and diagnosis. Automatically writing reports for these images can considerably lighten the workload of radiologists for summarizing descriptive findings and conclusive impressions. The complex structures between and within sections of the reports pose a great challenge to the automatic report generation. Specifically, the section Impression is a diagnostic summarization over the section Findings; and the appearance of normality dominates each section over that of abnormality. Existing studies rarely explore and consider this fundamental structure information. In this work, we propose a novel framework which exploits the structure information between and within report sections for generating CXR imaging reports. First, we propose a two-stage strategy that explicitly models the relationship between Findings and Impression. Second, we design a novel co-operative multi-agent system that implicitly captures the imbalanced distribution between abnormality and normality. Experiments on two CXR report datasets show that our method achieves state-of-the-art performance in terms of various evaluation metrics. Our results expose that the proposed approach is able to generate high-quality medical reports through integrating the structure information.

Visual Story Post-Editing

- https://www.aclweb.org/anthology/P19-1658/

We introduce the first dataset for human edits of machine-generated visual stories and explore how these collected edits may be used for the visual story post-editing task. The dataset ,VIST-Edit, includes 14,905 human-edited versions of 2,981 machine-generated visual stories. The stories were generated by two state-of-the-art visual storytelling models, each aligned to 5 human-edited versions. We establish baselines for the task, showing how a relatively small set of human edits can be leveraged to boost the performance of large visual storytelling models. We also discuss the weak correlation between automatic evaluation scores and human ratings, motivating the need for new automatic metrics.

Multimodal ive Summarization for How2 Videos

- https://www.aclweb.org/anthology/P19-1659/

In this paper, we study ive summarization for open-domain videos. Unlike the traditional text news summarization, the goal is less to "compress" text information but rather to provide a fluent textual summary of information that has been collected and fused from different source modalities, in our case video and audio transcripts (or text). We show how a multi-source sequence-to-sequence model with hierarchical attention can integrate information from different modalities into a coherent output, compare various models trained with different modalities and present pilot experiments on the How2 corpus of instructional videos. We also propose a new evaluation metric (Content F1) for ive summarization task that measures semantic adequacy rather than fluency of the summaries, which is covered by metrics like ROUGE and BLEU.