19 DECEMBER 2018 / TRANSFER LEARNING

10 Exciting Ideas of 2018 in NLP

This post gathers 10 ideas that I found exciting and impactful this year—and that we'll likely see more of in the future.

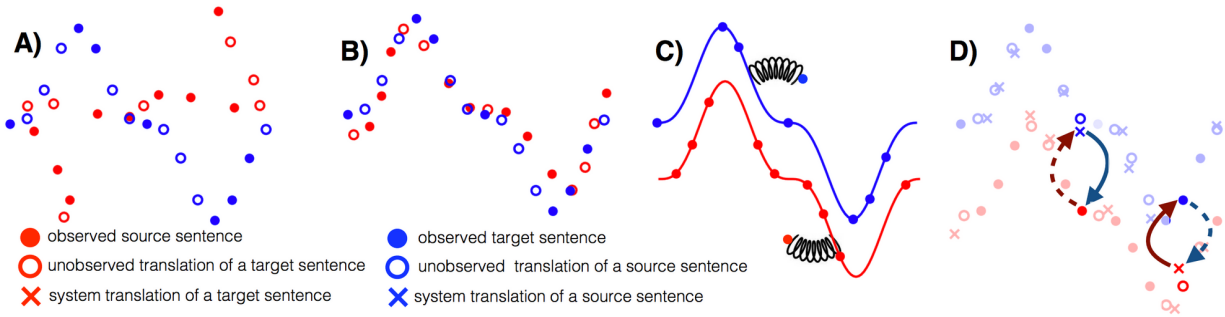
For each idea, I will highlight 1-2 papers that execute them well. I tried to keep the list succinct, so apologies if I did not cover all relevant work. The list is necessarily subjective and covers ideas mainly related to transfer learning and generalization. Most of these (with some exceptions) are not trends (but I suspect that some might become more 'trendy' in 2019). Finally, I would love to read about your highlights in the comments or see highlights posts about other areas.

1) Unsupervised MT

There were two unsupervised MT papers at ICLR 2018. They were *surprising* in that they worked at all, but results were still low compared

with two papers from the same two groups that significantly improve upon their previous methods. My highlight:

- **Phrase-Based & Neural Unsupervised Machine Translation** (EMNLP 2018): The paper does a nice job in distilling the three key requirements for unsupervised MT: a good initialization, language modelling, and modelling the inverse task (via back-translation). All three are also beneficial in other unsupervised scenarios, as we will see below. Modelling the inverse task enforces cyclical consistency, which has been employed in different approaches—most prominently in CycleGAN. The paper performs extensive experiments and evaluates even on two low-resource language pairs, English-Urdu and English-Romanian. We will hopefully see more work on low-resource languages in the future.



Toy illustration of the three principles of unsupervised MT. A) Two monolingual datasets. B) Initialization. C) Language modelling. D) Back-translation (Lample et al., 2018).

2) Pretrained language models

Using pretrained language models is probably the most significant NLP trend this year, so I won't spend much time on it here. There have been a slew of memorable approaches: ELMo, ULMFiT, OpenAI Transformer, and BERT. My highlight:

- **Deep contextualized word representations** (NAACL-HLT 2018): The paper that introduced ELMo has been much lauded.

careful analysis section that teases out the impact of various factors and analyses the information captured in the representations. The word sense disambiguation (WSD) analysis by itself (below on the left) is well executed. Both demonstrate that a LM on its own provides WSD and POS tagging performance close to the state-of-the-art.

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

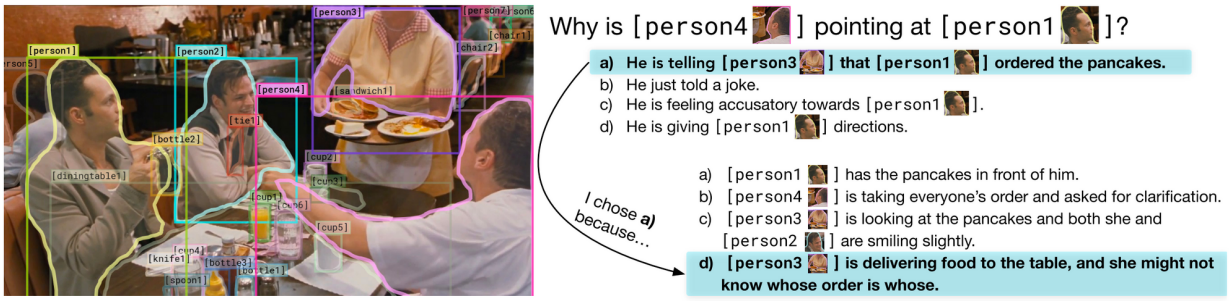
Word sense disambiguation (left) and POS tagging (right) results of first and second layer bidirectional language model compared to baselines (Peters et al., 2018).

3) Common sense inference datasets

Incorporating common sense into our models is one of the most important directions moving forward. However, creating good datasets is not easy and even popular ones show large biases. This year, there have been some well-executed datasets that seek to teach models some common sense such as Event2Mind and SWAG, both from the University of Washington. SWAG was solved unexpectedly quickly. My highlight:

- **Visual Commonsense Reasoning** (arXiv 2018): This is the first visual QA dataset that includes a rationale (an explantation) with each answer. In addition, questions require complex reasoning. The creators go to great lengths to address possible bias by ensuring that every answer's prior probability of being correct is 25% (every answer appears 4 times in the entire dataset, 3 times as an incorrect answer and 1 time as the correct answer); this requires solving a constrained optimization problem using models

possible bias will become a common component when creating datasets. Finally, just look at the gorgeous presentation of the data 👇.

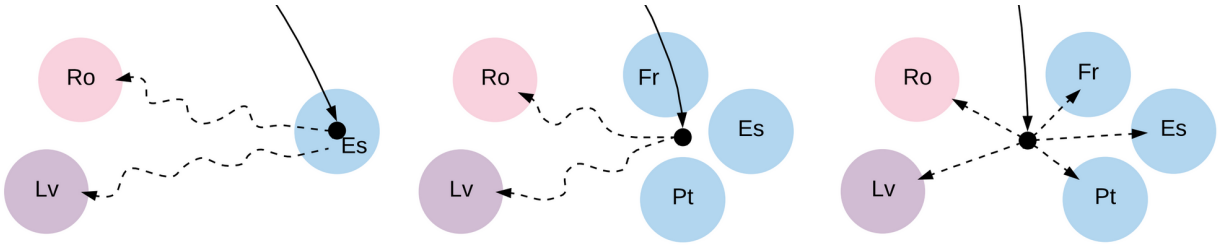


VCR: Given an image, a list of regions, and a question, a model must answer the question and provide a rationale explaining why its answer is right (Zellers et al., 2018).

4) Meta-learning

Meta-learning has seen much use in few-shot learning, reinforcement learning, and robotics—the most prominent example: model-agnostic meta-learning (MAML)—but successful applications in NLP have been rare. Meta-learning is most useful for problems with a limited number of training examples. My highlight:

- **Meta-Learning for Low-Resource Neural Machine Translation** (EMNLP 2018): The authors use MAML to learn a good initialization for translation, treating each language pair as a separate meta-task. Adapting to low-resource languages is probably the most useful setting for meta-learning in NLP. In particular, combining multilingual transfer learning (such as multilingual BERT), unsupervised learning, and meta-learning is a promising direction.



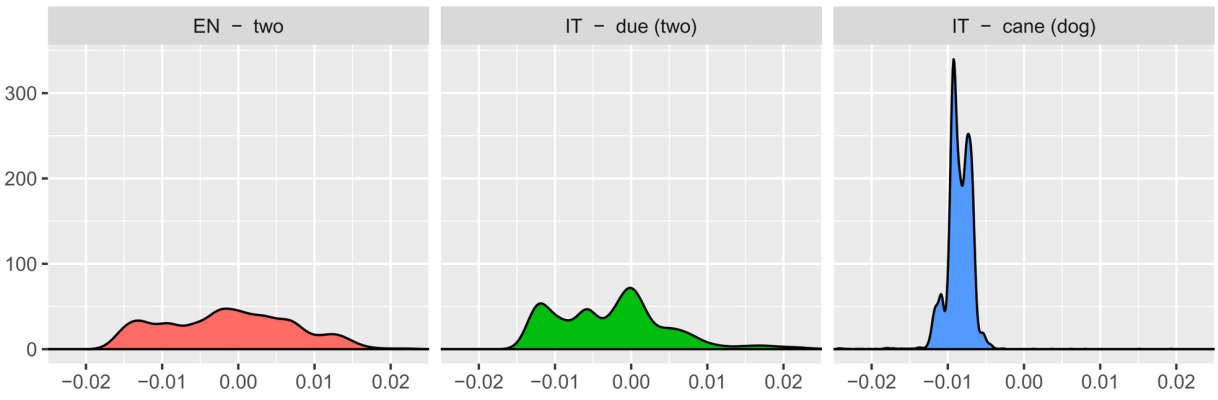
(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning

The difference between transfer learning multilingual transfer learning, and meta-learning. Solid lines: learning of the initialization. Dashed lines: Path of fine-tuning (Gu et al., 2018).

5) Robust unsupervised methods

This year, we and others have observed that unsupervised cross-lingual word embedding methods break down when languages are dissimilar. This is a common phenomenon in transfer learning where a discrepancy between source and target settings (e.g. domains in domain adaptation, tasks in continual learning and multi-task learning) leads to deterioration or failure of the model. Making models more robust to such changes is thus important. My highlight:

- **A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings** (ACL 2018): Instead of meta-learning an initialization, this paper uses their understanding of the problem to craft a better initialization. In particular, they pair words in both languages that have a similar distribution of words they are similar to. This is a great example of using domain expertise and insights from an analysis to make a model more robust.

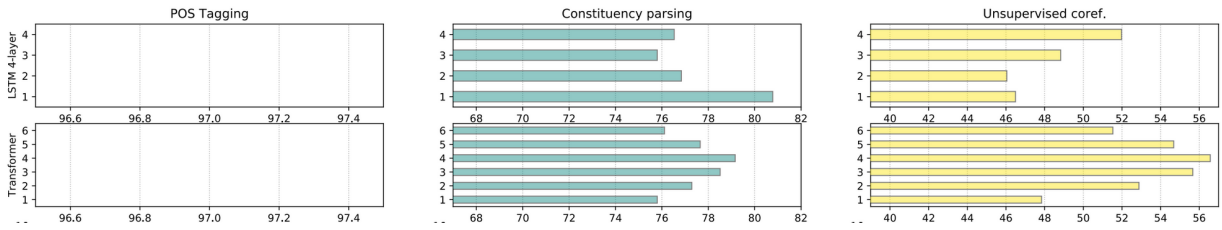


more similar distributions than non-related words ('two' and 'cane'—meaning 'dog'; Artexte et al., 2018).

6) Understanding representations

There have been a lot of efforts in better understanding representations. In particular, 'diagnostic classifiers' (tasks that aim to measure if learned representations can predict certain attributes) have become quite common. My highlight:

- **Dissecting Contextual Word Embeddings: Architecture and Representation** (EMNLP 2018): This paper does a great job of better understanding pretrained language model representations. They extensively study learned word and span representations on carefully designed unsupervised and supervised tasks. The resulting finding: Pretrained representations learn tasks related to low-level morphological and syntactic tasks at lower layers and longer range semantics at higher layers. To me this really shows that pretrained language models indeed capture similar properties as computer vision models pretrained on ImageNet.



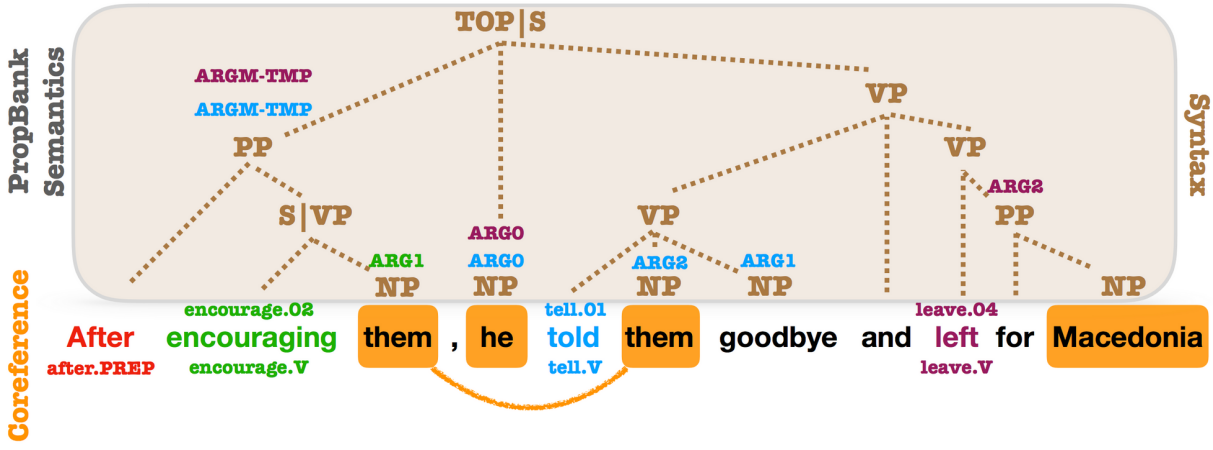
Per-layer performance of BiLSTM and Transformer pretrained representations on (from left to right) POS tagging, constituency parsing, and unsupervised coreference resolution (Peters et al., 2018).

7) Clever auxiliary tasks

In many settings, we have seen an increasing usage of multi-task learning with carefully chosen auxiliary tasks. For a good auxiliary task, data must

uses next-sentence prediction (that has been used in Skip-thoughts and more recently in Quick-thoughts) to great effect. My highlights:

- **Syntactic Scaffolds for Semantic Structures** (EMNLP 2018): This paper proposes an auxiliary task that pretrains span representations by predicting for each span the corresponding syntactic constituent type. Despite being conceptually simple, the auxiliary task leads to large improvements on span-level prediction tasks such as semantic role labelling and coreference resolution. This papers shows that specialised representations learned at the level required by the target task (here: spans) are immensely beneficial.
- **pair2vec: Compositional Word-Pair Embeddings for Cross-Sentence Inference** (arXiv 2018): In a similar vein, this paper pretrains *word pair representations* by maximizing the pointwise mutual information of pairs of words with their context. This encourages the model to learn more meaningful representations of word pairs than with more general objectives, such as language modelling. The pretrained representations are effective in tasks such as SQuAD and MultiNLI that require cross-sentence inference. We can expect to see more pretraining tasks that capture properties particularly suited to certain downstream tasks and are complementary to more general-purpose tasks like language modelling.



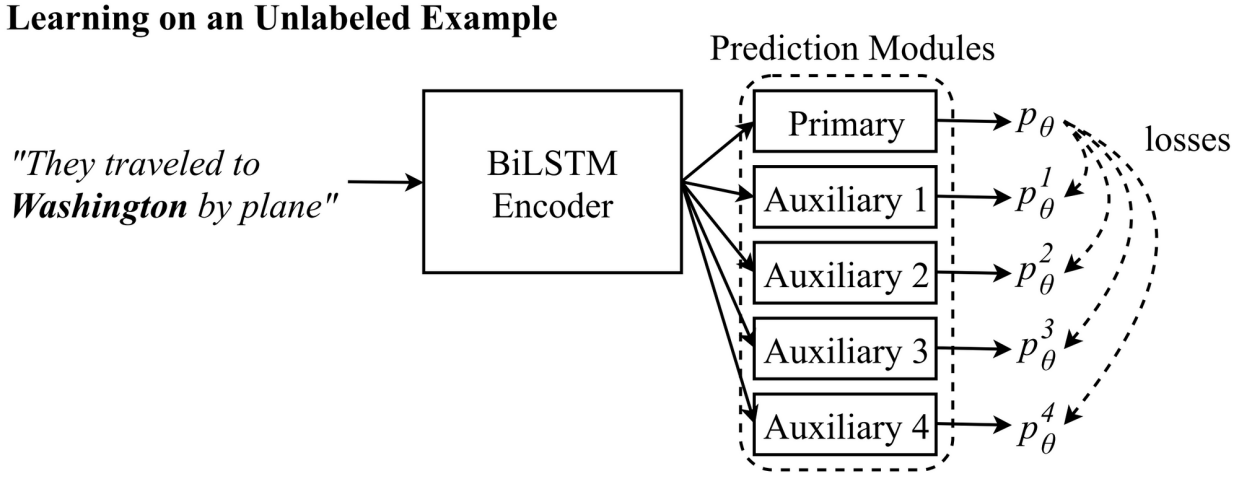
Syntactic, PropBank and coreference annotations from OntoNotes. PropBank SRL arguments and coreference mentions are annotated on top of syntactic constituents. Almost

every argument is related to a syntactic constituent (Swayamdipta et al., 2018).

8) Combining semi-supervised learning with transfer learning

With the recent advances in transfer learning, we should not forget more explicit ways of using target task-specific data. In fact, pretrained representations are complementary with many forms of semi-supervised learning. We have explored self-labelling approaches, a particular category of semi-supervised learning. My highlight:

- **Semi-Supervised Sequence Modeling with Cross-View Training** (EMNLP 2018): This paper shows that a conceptually very simple idea, making sure that the predictions on different views of the input agree with the prediction of the main model, can lead to gains on a diverse set of tasks. The idea is similar to word dropout but allows leveraging unlabelled data to make the model more robust. Compared to other self-ensembling models such as mean teacher, it is specifically designed for particular NLP tasks. With much work on *implicit* semi-supervised learning, we will hopefully see more work that explicitly tries to model the target predictions going forward.



Inputs seen by auxiliary prediction modules: Auxiliary 1: *They traveled to* _____. Auxiliary 2: *They traveled to* **Washington** _____. Auxiliary

9) QA and reasoning with large documents

There have been a lot of developments in question answering (QA), with an array of new QA datasets. Besides conversational QA and performing multi-step reasoning, the most challenging aspect of QA is to synthesize narratives and large bodies of information. My highlight:

- **The NarrativeQA Reading Comprehension Challenge** (TACL 2018): This paper proposes a challenging new QA dataset based on answering questions about entire movie scripts and books. While this task is still out of reach for current methods, models are provided the option of using a summary (rather than the entire book) as context, of selecting the answer (rather than generate it), and of using the output from an IR model. These variants make the task more feasible and enable models to gradually scale up to the full setting. We need more datasets like this that present ambitious problems, but still manage to make them accessible.

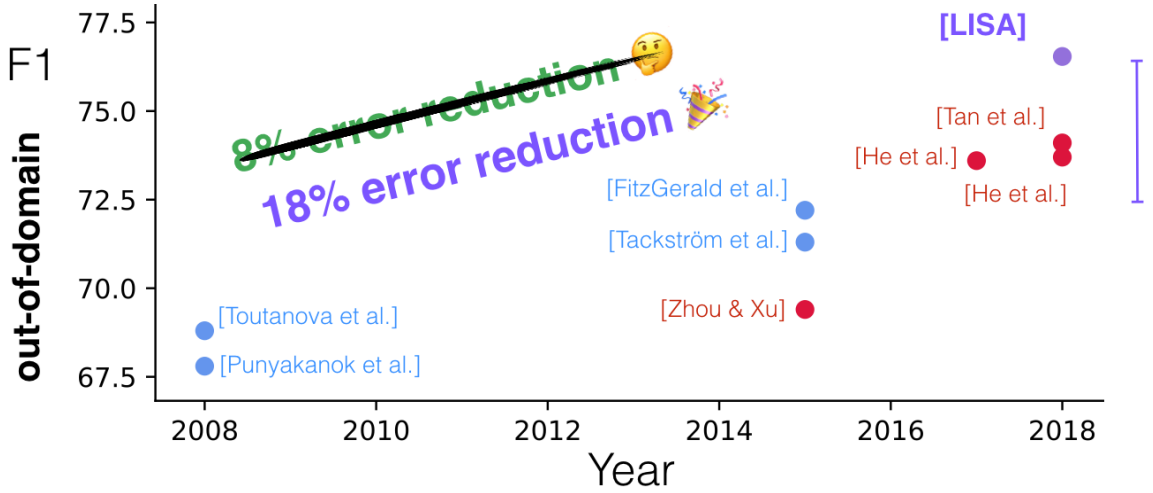
Dataset	Documents	Questions	Answers
MCTest (Richardson et al., 2013)	660 short stories, grade school level	2640 human generated, based on the document	multiple choice
CNN/Daily Mail (Hermann et al., 2015)	93K+220K news articles	387K+997K Cloze-form, based on highlights	entities
Children's Book Test (CBT) (Hill et al., 2016)	687K of 20 sentence passages from 108 children's books	Cloze-form, from the 21st sentence	multiple choice
BookTest (Bajgar et al., 2016)	14.2M, similar to CBT	Cloze-form, similar to CBT	multiple choice
SQuAD (Rajpurkar et al., 2016)	23K paragraphs from 536 Wikipedia articles	108K human generated, based on the paragraphs	spans
NewsQA (Trischler et al., 2016)	13K news articles from the CNN dataset	120K human generated, based on headline, highlights	spans
MS MARCO (Nguyen et al., 2016)	1M passages from 200K+ documents retrieved using the queries	100K search queries	human generated, based on the passages
SearchQA (Dunn et al., 2017)	6.9m passages retrieved from a search engine using the queries	140k human generated Jeopardy! questions	human generated Jeopardy! answers
NarrativeQA (this paper)	1,572 stories (books, movie scripts) & human generated summaries	46,765 human generated, based on summaries	human generated, based on summaries

Comparison of QA datasets (Kočiský et al., 2018).

10) Inductive bias

Inductive biases such as convolutions in a CNN, regularization, dropout, and other mechanisms are core parts of neural network models that act as a regularizer and make models more sample-efficient. However, coming up with a broadly useful inductive bias and incorporating it into a model is challenging. My highlights:

- **Sequence classification with human attention** (CoNLL 2018): This paper proposes to use human attention from eye-tracking corpora to regularize attention in RNNs. Given that many current models such as Transformers use attention, finding ways to train it more efficiently is an important direction. It is also great to see another example that human language learning can help improve our computational models.
- **Linguistically-Informed Self-Attention for Semantic Role Labeling** (EMNLP 2018): This paper has a lot to like: a Transformer trained jointly on both syntactic and semantic tasks; the ability to inject high-quality parses at test time; and out-of-domain evaluation. It also regularizes the Transformer's multi-head attention to be more sensitive to syntax by training one attention head to attend to the syntactic parents of each token. We will likely see more examples of Transformer attention heads used as auxiliary predictors focusing on particular aspects of the input.



10 years of PropBank semantic role labeling. Comparison of Linguistically-Informed Self-

Sebastian Ruder
Read more posts by this author.

Read More

— Sebastian Ruder —
transfer learning

EMNLP 2018 Highlights: Inductive bias, cross-lingual learning, and more

A Review of the Neural History of Natural Language Processing

ACL 2018 Highlights: Understanding Representations and Evaluation in More Challenging Settings

See all 10 posts →

EVENTS

EMNLP 2018 Highlights: Inductive bias, cross-lingual learning, and more

This post discusses highlights of EMNLP 2018. It focuses on talks and papers dealing with inductive bias, cross-lingual learning, word embeddings, latent variable models, language models, and datasets.

11 MIN READ