Research Statement – Shafiq Joty

The Internet is a great source of human knowledge, but most of the information is hidden in unstructured texts. As a researcher in Natural Language Processing (NLP), my goal is to add structure to this text to uncover relevant information, and to use it in developing useful applications. To this end, my research interests span two areas of NLP: (*a*) developing language analysis tools to understand human language, and (*b*) to build NLP applications to support end users. For *(a)*, I am interested in **parsing** texts with syntactic, semantic and discourse structures (§1). For *(b)*, my interests lie in NLP applications that involve not only language understanding but also **generation** (§3). These include large-scale language modeling (LM), machine translation (MT), text summarization, question answering (QA) and dialogue systems. I focus on **multilingual** processing, where I develop NLP models for not only English but other low-resource languages and dialects (§2). As NLP methods are becoming more and more ubiquitous, directly impacting humanity and commerce, I have also been looking into the **security and robustness** of NLP models to ensure that they do not exhibit algorithmic bias and discriminate on the basis of factors such as gender, race, name, location or speaker (§4).[1]

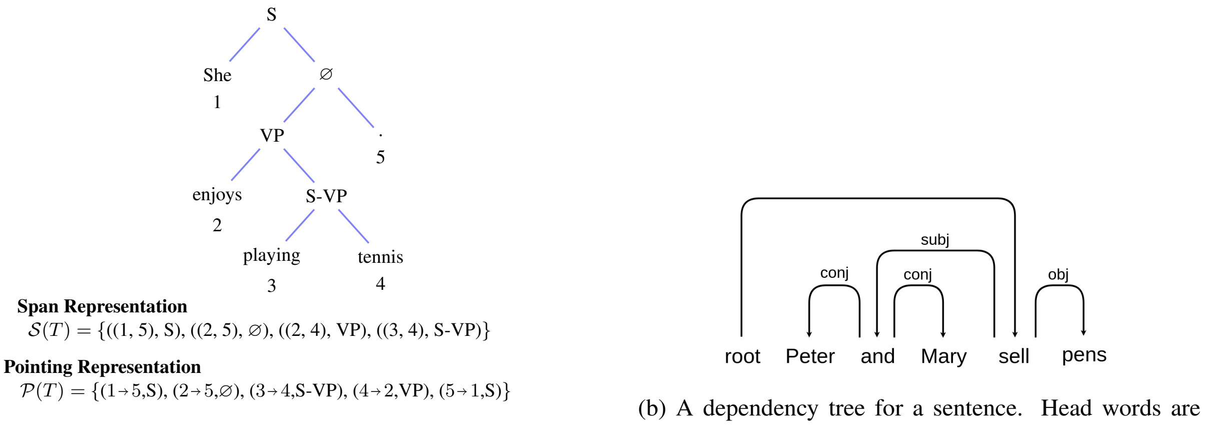
I am also interested in **interdisciplinary** research that goes beyond NLP (§5). I have been collaborating with the (*i*) computer vision group to develop effective **multi-modal** (text and image) representation learning models (§5.1), (*ii*) speech group for effective speech recognition solutions (§5.4), (*iii*) social computing group for crisis computing and fact checking solutions (§5.2), (*iv*) database and data mining group on solutions for more effective database education and recommendation (§5.3), and (*v*) health science group to develop effective health applications (§5.5). I have recently embarked on a joint research project on Covid-19 with Worth Health Organization (WHO) and Lee Kong Chian (LKC) School of Medicine, where my collaborators and I are investigating machine learning models for effective media monitoring using neural search, question answering, multi-document summarization and topic modeling. One methodology emphasized throughout my research is to first identify the inherent semantic structures in a given problem, and then to develop structured machine learning models to exploit such structures effectively. My work has heavily relied on **deep learning** (DL) for better representation of the input text and on **probabilistic graphical models** (PGM) and **reinforcement learning** (RL) for capturing latent dependencies in the output.

1 Language Understanding & Parsing

Natural language is ambiguous. As humans, we can easily disambiguate the meaning of linguistic units (phrases, sentences) as we read or listen. However, for machines it is difficult to understand without explicit representations of syntax, semantics and discourse. In my group, we develop NLP tools to parse natural language in terms of its syntax (constituency, dependency), semantics (*e.g.,* named entities) and discourse structures (*e.g.,* coreference, coherence).

1.1 Syntactic Parsing (Constituency & Dependency)

Constituency and dependency are two different formalisms that represent the grammatical structure of a sentence. *Constituency* (*a.k.a.* phrase-structure) trees organize words and phrases into nested constituents (fig. 1a), whereas words in a *dependency* tree are connected directly with each other by directed links called dependencies (fig. 1b).


S
She
1
∅
VP
.
5
enjoys
2
S-VP
playing
3
tennis
4
Span Representation
$\mathcal{S}(T) = \{((1,5),\text{S}), ((2,5),\varnothing), ((2,4),\text{VP}), ((3,4),\text{S-VP})\}$
Pointing Representation
$\mathcal{P}(T) = \{(1\rightarrow 5,\text{S}), (2\rightarrow 5,\varnothing), (3\rightarrow 4,\text{S-VP}), (4\rightarrow 2,\text{VP}), (5\rightarrow 1,\text{S})\}$
subj
conj
conj
obj
root
Peter
and
Mary
sell
pens

(a) A binarized constituency tree for a sentence. The span and pointing representations of the tree are shown below the tree.

(b) A dependency tree for a sentence. Head words are connected with their modifiers by directed links. Our parser is trained to point to the dependents given a head.

Figure 1: Constituency and dependency tree structures.

[1] Source code (and a few demos) of most of the research projects can be found at https://ntunlpsg.github.io/resources/.

In recent years, neural end-to-end parsing methods have outperformed traditional methods that use grammar, lexicon and hand-crafted features. Transition-based parsers generate trees by performing shift-reduce actions. Though computationally attractive, the local decisions made at each step may propagate errors to subsequent steps due to *exposure bias* [9]. On the other hand, globally optimal models such as Chart methods for constituency and graph-based methods for dependency parsing, are generally slow with at least $\mathcal{O}(n^3)$ time complexity.

In contrast to previous work, our proposed approach [70, 89, 90] casts the parsing tasks into a series of conditional splitting decisions and uses a Pointer network [113] to model the splitting decision as pointing to the split points (fig. 1a) at each decoding step (fig. 2). The conditional probabilities of the splitting decisions are optimized using a cross entropy loss and structural consistency is maintained through a global pointing mechanism. The training process is fully parallelized without requiring expensive structured inference like previous methods. Our model enables efficient top-down decoding with $\mathcal{O}(n)$ running time like transition-based parsers, while also supporting a customized beam search to get the best tree by searching through a reasonable search space of high scoring trees. Moreover, our parser does not rely on any handcrafted features (not even part-of-speech tags), which makes it more efficient and flexible to different domains or languages. In the experiments with the English Treebank, our model achieves state-of-the-art (SoTA) results with/without pre-trained representations. It also performs competitively with SoTA methods on the multilingual parsing tasks in SPMRL 2013/2014. Our model supports faster decoding with a speed of over 1,100 sentences per second (fastest so far).

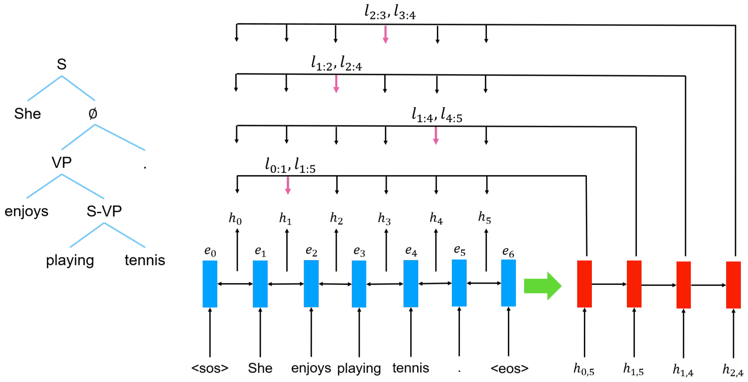


Figure 2: Our syntactic parser along with the decoding process for a given sentence. The input to the decoder at each step is the representation of the span to be split. We predict the splitting point using a biaffine function between the corresponding decoder state and the boundary-based encoder representations. A label classifier is used to assign labels to the left and right spans.

1.2 Discourse Analysis

In syntactic parsing, our goal was to find explicit representation that describes how the *sentences* are formed. However, a well-written text is not merely a sequence of independent sentences, but instead a sequence of related sentences. It addresses a particular topic, often covering multiple subtopics, and is organized in a coherent way that enables the reader to process the information. In discourse analysis, we seek to uncover such underlying structures, which can support many downstream applications. In my PhD [33, 40, 41, 38, 39, 36, 35] which was supported by NSERC CGS-D[2], I proposed novel computational models for discovering the *rhetorical*, *topical* and *conversational* structures of a discourse.[3] I continued working on these topics after my PhD, focusing not only on improving them further but also on using the tools to improve end-user applications. I will describe some of them briefly in the following subsections.

1.2.1 RST Parsing & Its Applications

Different formal theories have been proposed to describe the *coherence* structure of a text. Rhetorical Structure Theory (RST) is perhaps the most influential one, which posits a tree-like discourse structure. For example, consider the discourse tree in Figure 3. The leaves of the tree correspond to contiguous atomic text spans, called elementary discourse units (EDUs). Adjacent EDUs are connected by coherence relations (e.g., *Elaboration*, *Contrast*), forming larger discourse units, which in turn are also subject to this relation linking. Discourse units linked by a relation are further distinguished based on their relative importance in the text: *nuclei* are the core parts of the relation while *satellites* are peripheral ones. For example, in Figure 3, the satellite EDU "— manufacturing strength —" *elaborates* the nucleus EDU "But the thing it's supposed to measure", and two nuclei EDUs "Some people use the purchasers index as a leading indicator" and "some use it as a coincident indicator" *contrast* each other. Conventionally, RST parsing involves two subtasks: (*i*) **discourse segmentation** is the task of breaking the text into a sequence of EDUs, and (*ii*) **discourse parsing** is the task of linking the discourse units (EDUs and larger units) into a labeled tree.

During my PhD [33, 40, 41, 38], I developed CODRA – a COmplete Discriminative framework for Rhetorical Analysis, which comprises a discourse segmenter and a discourse parser. The crucial component is the use of a probabilistic discriminative parsing model, expressed as a Dynamic Conditional Random Field (DCRF), to infer the probability of all possible tree constituents. By representing the structure and the relation of each tree constituent jointly and by explicitly

[2] NSERC CGS-D is awarded to the top-ranked PhD students across Canada.

[3] Conversational (or dialogue) structures are applicable to only conversational discourse (e.g., multi-party conversations).

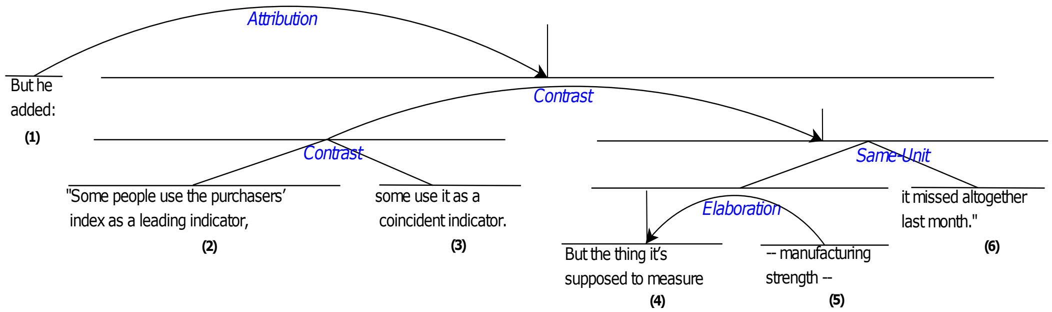

Attribution
But he added:
(1)
Contrast
Contrast
"Some people use the purchasers' index as a leading indicator,
(2)
some use it as a coincident indicator.
(3)
Same-Unit
Elaboration
But the thing it's supposed to measure
(4)
-- manufacturing strength --
(5)
it missed altogether last month."
(6)

Figure 3: A sample discourse tree. Horizontal lines indicate text segments; satellites are connected to their nuclei by curved arrows and two nuclei are connected with straight lines.

capturing the sequential dependencies between tree constituents, the DCRF model does not make any independence assumption among these properties. CODRA uses the inferred probabilities from the parsing models in a probabilistic chart-based (CKY) bottom–up parsing algorithm, which is optimal. Furthermore, a simple modification of this parsing algorithm allows us to generate k-best parse hypotheses, that are later used in a *tree kernel-based reranker* to improve over the initial ranking using additional (global) features of the discourse tree as evidence [49].

CODRA uses the traditional feature-based statistical method which, along with the CKY parsing ($\mathcal{O}(n^3)$), makes it slow for both training and inference, especially for long documents. Also, in CODRA and other existing methods, discourse segmentation is detached from parsing and treated as a prerequisite step. Therefore, the errors in segmentation affect the overall parsing performance. In our work [67, 70, 91], we are the first to propose a neural *top-down* discourse parser that can parse a document end-to-end from scratch, making it much more efficient and easily adaptable to new languages, domains and tasks by surpassing the expensive feature engineering step that often requires more time and domain/language expertise. Crucially, our parser generates a discourse tree from scratch without requiring discourse segmentation as a prerequisite; rather, it generates the EDUs as a by-product of parsing. Our novel formulation which is based on the pointing mechanism (§1.1), facilitates solving discourse segmentation and parsing tasks in a single unified neural model. Our parser achieves SoTA results on the benchmark datasets, while being faster than the existing ones. In another front, we propose SegBot [62], a general-purpose text segmentation model based on pointer networks, and show its effectiveness in both topic segmentation (*i.e.,* larger discourse units) and EDU segmentation.

Application to MT. Among other applications of discourse, Machine Translation (MT) has received a resurgence of interest lately. It is admitted that MT systems should consider phenomena that go beyond the current sentence to ensure consistency in the choice of lexical items and referring expressions, and that source-language coherence relations are also realized in the target language. Automatic MT evaluation is an integral part of the process of developing and tuning MT systems. Reference-based evaluation metrics compare the output of a system to one or more human (reference) translations, and produce a similarity score indicating the quality of the translation. The initial MT metrics approached similarity as a shallow word n-gram matching between the translation and the reference, with a limited use of linguistic information. BLEU is the best-known metric in this family, which has been used for years. However, it has been shown that BLEU and metrics akin to it are insufficient and unreliable for high-quality translation output.

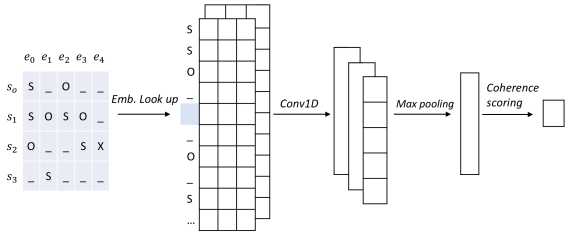
In [43, 26], we show that discourse information can be used to produce evaluation measures that improve over the SoTA in terms of correlation with human assessments. We conduct our research in four steps. First, we design a simple discourse-aware metric DR-LEX, which uses sub-tree kernel to compare RST trees generated with CODRA. We show that this metric helps to improve a large number of MT evaluation measures at the segment-level and at the system-level. Second, we show that tuning the weights in the linear combination of metrics using human assessed examples is a robust way to improve the effectiveness of the DR-LEX metric significantly. Third, we conduct an ablation study which helps us understand which elements of the RST tree have the highest impact on the quality of the evaluation measure. Interestingly enough, the *nuclearity* feature (i.e., the distinction between main and subordinate units) turns out to be more important than the discourse relation labels. Finally, based on these findings, we extend the tree-based representations and present the DISCOTK$_{party}$ metric, which makes use of a combination of discourse tree representations and many other metrics. The resulting combined metric with tuned weights scored best as compared to human rankings at the WMT14 Metrics task, both at the system and at the segment levels.

1.2.2 Coherence Modeling & Its Applications

Rather than parsing a discourse, the goal in coherence modeling is to build models that can distinguish a coherent text from incoherent ones. It has been a key problem in discourse analysis with applications in text generation, summarization, and coherence evaluation (*e.g.,* essay scoring). Inspired by formal theories of discourse, a number of coherence models have been proposed including the entity-based [8], syntax-based [73] and discourse relation based models [68]. The *entity grid* model [8] is one of the most popular models that has received much attention over the years. As exemplified in fig. 4a, the model represents a text by a grid that captures how grammatical roles of different discourse entities (nouns) change from one sentence to another in the text. The grid is then converted

	INVESTORS	MILLION	FUNDS	EQUIPMENT	CORP.	**PAPER**	SALE	TELECOMM.	LEASE	PROGRAM	CLEVELAND	RECEIVABLES	LEASES	DATA-PROCESS.	**LDI**	NON-RECOURSE
s_0	–	O	–	–	S	X	–	–	–	–	X	X	–	–	X	–
s_1	–	–	O	–	–	X	X	–	–	S	–	–	X	–	–	–
s_2	S	–	–	–	X	S	–	–	X	–	–	X	–	–	S	X
s_3	–	–	–	O	–	–	–	X	–	–	–	–	–	X	S	–

(a) Entity grid representation for a document. Grid cells correspond to grammatical roles of the entities in the sentences: subjects (S), objects (O), other (X), and absent (–).



(b) Neural entity grid model proposed in [85]. The model is trained using a pairwise ranking approach with shared parameters for positive and negative documents.

Figure 4: (a) Entity grid representation of a document [8]; (b) the neural entity grid model [85].

into a feature vector containing probabilities of local entity transitions, enabling ML models to measure the degree of coherence. In our work [85], we neuralized the traditional entity-grid model by using *distributed representations* of entity transitions and entity features. We also presented an end-to-end training method to learn task-specific high level features automatically in our model (fig. 4). In a follow-up work [48], we further improved the model by *lexicalizing* the entity transitions based on word embeddings, and adapted the model to asynchronous conversations by incorporating the underlying *conversational structure* in the grid representation and subsequently in feature computation. For this, we proposed a novel grid representation for asynchronous conversations, and adapted the convolution layer of the neural model accordingly.

According to [21], three factors collectively contribute to coherence: the organization of discourse segments, intention or purpose of the discourse, and attention or focused items. The entity-based approaches capture attentional structure, the syntax-based ones consider intention, and the organizational structure is largely captured by models that consider discourse relations and content (topic) distribution. In our later work [84], we proposed a *unified neural model* that incorporates sentence grammar (intentional structure), discourse relations, attention and topic structures in a single framework (fig. 5). We used an LSTM sentence encoder with explicit language model loss to capture the syntax. Inter-sentence discourse relations are modeled with a bilinear layer, and a lightweight convolution-pooling is used to capture the attention and topic structures. We evaluated our models on both local and global discrimination tasks on the benchmark dataset. Our results showed that our approach outperforms existing methods by a wide margin in both tasks.

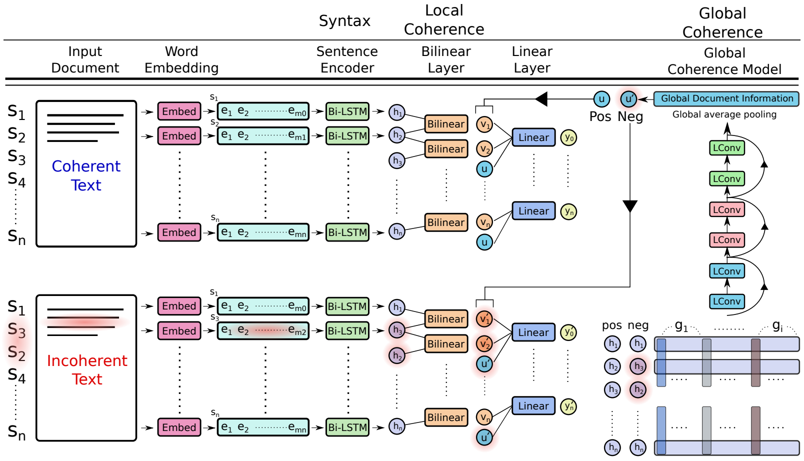


Figure 5: An overview of the coherence model proposed in [84].

Applications of Coherence Models. Despite continuous research efforts in developing novel coherence models, their usefulness in downstream tasks has largely been ignored. They have been evaluated in mainly two ways. The most common approach has been to evaluate them on *synthetic* discrimination tasks that involve identifying the right order of the sentences at the local and global levels. The other (rather infrequent) way has been to assess the impact of coherence score as an additional feature in downstream tasks like readability assessment and essay scoring. But since the concept of coherence goes beyond these constrained tasks and domains, so should the models. Given the recent advances in neural NLP methods, with claims of reaching human parity in machine translation, and fluency in summarization and language modeling, coherence evaluation of machine-generated texts is now more crucial than ever [60].

In our recent work [82], we showed through experiments that there is only a slight correlation between model performances on synthetic tasks and real-world use cases. Although models perform strongly in the synthetic tasks, they show poor performance and low correlations with human judgments on distinguishing coherent machine translations and system-generated summaries from incoherent ones. They also fail to perform well on the next utterance ranking task. However, we find that re-training the coherence models with task-specific data for machine translation evaluation leads to improved results and agreements with human judgments. This led us to infer that models trained on traditional synthetic tasks do not seem to be learning features that are useful for downstream applications. In our ongoing work, we are redesigning the training paradigm for coherence models, while evaluating them on more tasks.

1.2.3 Topic & Conversation Modeling

(a) Topic Modeling in Asynchronous Conversations. Coherence relations model *local* coherence between neighbouring text segments. At a larger and global level, a discourse exhibits a topic structure. For example, a news article about an earthquake may talk about the intensity, the damage, the aftershocks, and the casualties. Likewise, an email conversation about arranging a conference may discuss conference schedule, organizing committee, accommodation, and registration. **Topic segmentation** refers to the task of grouping the sentences into a set of non-overlapping topical segments, and **topic labeling** is the task of assigning short descriptions to the topical segments to facilitate interpretations of the topics.

Extensive research had been conducted in topic segmentation for monologue (*e.g.,* news articles) and for synchronous dialog (*e.g.,* meetings). However, no-one had studied this problem for *asynchronous* conversations (*e.g.,* emails). There was no reliable annotation scheme, no standard corpus, and no agreed-upon evaluation metrics available. During my PhD [33, 39, 36], I presented two new corpora of email and blog conversations annotated with topics, and proposed a complete computational framework and evaluation/agreement measures. For topic segmentation, I proposed two novel unsupervised models that exploit the fine-grained conversational structure beyond lexical information. I also proposed a novel graph-theoretic supervised topic segmentation model that combines lexical, conversational and topic features. For topic labeling, I proposed two novel guided random walk models that capture conversation specific clues from two different sources respectively: leading sentences and fine-grained conversational structure. In a follow-up work [77], we propose to generate *abstractive* topic labels with *textual entailment* and *aggregation* instead of simply extracting them.

(b) Conversation Disentanglement. Multiple ongoing conversations occur naturally, especially when the conversation involves more than two participants (fig. 6). A task related to topic segmentation is disentanglement, where the goal is identify individual conversations from an interleaved discussion thread. This can support users by providing online help. It is often considered as a prerequisite for downstream tasks such as utterance ranking and generation, summarization and question answering [37]. Prior methods rely mostly on handcrafted features that are dataset specific, which hinders generalization and adaptability. In our work [122], we proposed an end-to-end online framework that avoids time-consuming domain-specific feature engineering. We designed a novel way to embed the whole utterance that comprises timestamp, speaker, and message text, and proposed a custom attention mechanism that models disentanglement as a pointing problem while effectively capturing inter-utterance interactions in an end-to-end fashion. We also introduced a joint-learning objective to better capture contextual information. Our experiments showed that our method achieves SoTA performance in both link and conversation prediction tasks.

Time	Sp	Message Text
02:26	system	===zelot joined the channel
02:26	zelot	hi, where can i get some help in regards to issues with mount?
02:26	TuxThePenguin	After taking it out
02:26	hannasanarion	TuxThePenguin, try booting with monitors connected to motherboard
02:26	pnunn	TuxThePonguin, sounds like there is on board graphics as well, so try that without the card
02:27	hannasanarion	pnunn, right
02:27	pnunn	process of elimination.
02:27	TuxThePenguin	Along with Occam's Razor
02:27	Bashing-om	zelot: If you are on a supported release of 'buntu, this is a good place to ask.
02:27	TuxThePenguin	Any solution is most likely the simplest one
02:28	wllrt	I'm a emacs newb and looking to prevent rsi.

Figure 6: An excerpt of a conversation from the Ubuntu IRC corpus. Same color reflects same conversation.

(c) Utterance Ranking. *Retrieval-based* response generation that selects a suitable response from a pool of candidates (pre-existing human responses) has become a popular approach to framing dialog. Compared to the *generation-based* systems that generate novel utterances, retrieval-based systems produce fluent, grammatical and informative responses. Also compared to the traditional modular approach, it does not rely on dedicated modules for language understanding, dialog management, and generation, thus simplifying the system design. Prior methods typically aim to encode the context and the candidate responses in a joint semantic space by capturing short and long range dependencies, and then retrieve the most relevant response by matching the query representation against each candidate's representation through attentions. Most of these methods are however limited to only two-party conversations (mostly one conversation topic). As dialogue research progresses, it is necessary to study the more generic multi-party multi-turn scenario, which has become very common (*e.g.,* Slack, Whatsapp), and posits a unique set of challenges for the dialog models. For example, consider the conversation excerpt in fig. 6, where there are three ongoing conversation topics as highlighted by different color, and the participants can contribute to multiple topics simultaneously. An effective response selection method should model such complex conversational topic dynamics in the context, for which existing methods are deficient. It should match with its context in terms of the same conversation topic, while ignoring other non-relevant topics.

To address the these challenges in multi-party multi-turn dialog, in [116], we frame response selection as a dynamic topic tracking task with the intuition that the topic should remain the same as we go from the context to the response. Based on this new formulation, we propose a novel architecture (fig. 7) that can incorporate other related dialog tasks such as conversation disentanglement, enabling multi-task learning in a unified framework. Crucially, our formulation of the task needs to encode only two utterances at a time, thus allowing efficient encoding via large pretrained models

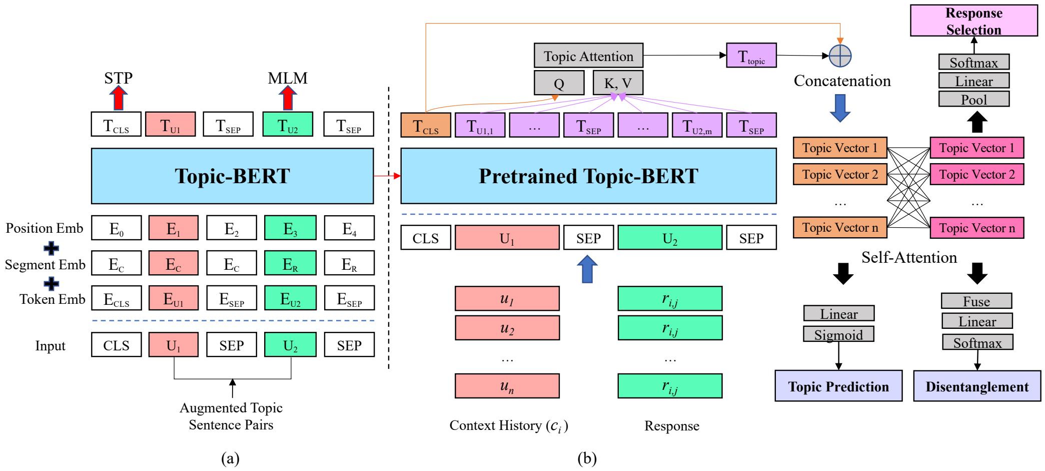


Figure 7: Overview of our Topic-BERT architecture [116]. (a) Topic-BERT pretraining with topic sentence pairs to incorporate utterance-utterance topic relationship. (b) Our multi-task framework which uses the pretrained Topic-BERT to enhance topic information in the encoded representations to support three downstream tasks – response selection as the main task while topic prediction and disentanglement as two auxiliary (optional) tasks.

like BERT. Furthermore, it facilitates pretraining of BERT-like models on topic related sentence pairs to incorporate topic relevance in pretraining, which can be done on large dialog corpora with self-supervised objectives, requiring no manual topic annotations, and can benefit not only response selection but also other dialog tasks. We evaluate the proposed models on the DSTC-8 Ubuntu IRC dataset, and show state-of-the-art results in both response selection and topic disentanglement outperforming the existing methods by a good margin.

(d) Dialog Act Recognition. Apart from the topic and coherence structures, a conversational discourse also exhibits a conversation structure; participants interact with each other by performing certain communicative acts like asking questions or requesting something, which are called **dialog acts**. Previous work on dialog act modeling has mostly focused on synchronous conversations. The dominant approaches use supervised sequence taggers like linear-chain CRFs to capture the conversational dependencies between the act types. However, modeling such conversational dependencies in asynchronous conversation is challenging, because the conversational flow often lacks sequential dependencies in its temporal order. In [33, 35], I proposed *unsupervised* conversational models, which are variants of Hidden Markov Models (HMMs). First, I demonstrated that the conversational models learn better sequential dependencies when they are trained on the sequences extracted from the finer conversational structure compared to when they are trained on the temporal order of the sentences. Further investigation shows that the simple unsupervised HMM tends to find topic clusters in addition to act clusters. To address this, I proposed an HMM+Mix model which not only explains away the topics, but also improves the act emission distribution by defining it as a mixture model.[4] In a follow-up work [44, 47], I proposed a class of *supervised* structured models in the form of CRFs defined over arbitrary graph structures of asynchronous conversations, while using an LSTM encoder to get the sentence representations. In [83], we advance the SoTA by proposing hierarchical LSTMs trained with word embeddings learned from a large unlabeled conversational corpus, and adapting the model with domain adversarial training to leverage the labeled data from synchronous domains by explicitly modeling the shift in the two domains (*e.g.,* meetings vs. forums).

1.3 Language Understanding Applications

1.3.1 Conversational Machine Reading

Significant progress has been made in teaching machines to read text and answer questions when the answer is directly expressed in the text (*e.g.,* SQuAD). However, in many situations such as interpreting rules to answer *Can I...?* or *Will I have to...?*, the text only gives a *recipe* to derive a final answer given the reader's background knowledge about the situation. This involves both the interpretation of instructions and reasoning based on the background knowledge. It can be further complicated due to missing information in the question in which case the reader has to ask further questions for clarification (fig. 8). This question answering scenario has been formalized as *conversational machine reading (CMR)* [98], where the machine (reader) needs to understand the knowledge base (KB) text, evaluate and keep track of the user scenario, ask clarification questions, and then make a final decision.

Existing approaches to CMR formalize the problem as two sub-tasks. The first is to make a decision among `Yes`, `No`, `Irrelevant`, and `Inquire` at each dialog turn given a rule text, a user scenario, an initial question and the current

[4] This work was conducted at Microsoft Research Asia, for which I was given the *"Microsoft Research Excellent Intern"* award.

dialog history. If one of Yes, No, or Irrelevant is selected, it implies that a final decision (Yes/No) can be made in response to the user's initial question, or stating the user's initial question is unanswerable (Irrelevant) according to the rule text. If the decision at the current turn is Inquire, it will then trigger the second task for follow-up question generation, which extracts an underspecified rule span from the rule text and generates a follow-up question accordingly.

In our work [19, 20], we identified two main drawbacks of existing methods. First, with respect to the reasoning of the rule text, they do not explicitly track whether a condition listed in the rule has already been satisfied as the conversation flows so that it can make a better decision. Second, with respect to the extraction of question-related rules, it is difficult for them to extract the most relevant text span to generate the next question.

In [19], we propose a new framework for CMR with a novel **E**xplicit **M**emory **T**racker (**EMT**), which explicitly tracks each rule sentence to make decisions and generate follow-up questions. It first segments the rule text into several rule sentences and allocates them into its memory. Then the initial question, user scenario, and dialog history are fed into EMT sequentially to update each memory module separately. At each turn, EMT predicts the entailment states (satisfaction or not) for every rule sentence, and makes a decision based on the current memory status. If the decision is Inquire, it extracts a rule span to generate a follow-up question by adopting a coarse-to-fine reasoning strategy (i.e., weighting token-level span distributions with its sentence-level entailment scores). Compared to previous methods which only consider entailment-oriented reasoning for decision making or follow-up question generation, EMT utilizes its updated memory modules to reason out these two tasks in a unified manner. Our results show that explicitly tracking rules with external memories boosts both the decision accuracy and the quality of generated follow-up questions. In addition to the performance improvement, EMT yields interpretability by explicitly tracking rules, which is visualized to show the entailment-oriented reasoning process of our model.

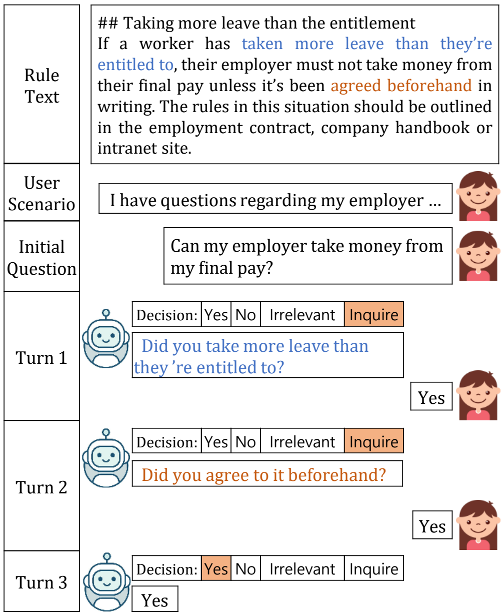

Rule Text
## Taking more leave than the entitlement
If a worker has taken more leave than they're entitled to, their employer must not take money from their final pay unless it's been agreed beforehand in writing. The rules in this situation should be outlined in the employment contract, company handbook or intranet site.
User Scenario
I have questions regarding my employer ...
Initial Question
Can my employer take money from my final pay?
Turn 1
Decision: Yes No Irrelevant Inquire
Did you take more leave than they 're entitled to?
Yes
Turn 2
Decision: Yes No Irrelevant Inquire
Did you agree to it beforehand?
Yes
Turn 3
Decision: Yes No Irrelevant Inquire
Yes

Figure 8: An example of Conversational Machine Reading tasks from ShARC [98].

Considering that interpretation of rule text and dialog is crucial for CMR, our follow-up work [20] proposes **DISCERN**: **Disc**ourse-Aware **E**ntailment **R**easoning **N**etwork(fig. 9). To better understand the logical structure of a rule text and to extract conditions from it, it first segments the rule text into clause-like elementary discourse units (EDUs) using a pre-trained discourse segmenter (§1.2.1). Each EDU is treated as a condition of the rule text, and our model estimates its entailment confidence scores over three states: ENTAILMENT, CONTRADICTION or NEUTRAL by reading the user scenario and dialog history. Then we map the scores to an entailment vector for each condition, and reason out the decision based on the entailment vectors and the logical structure of the rules. DISCERN is the first method to explicitly build the dependency between entailment states and decisions at each dialog turn. DISCERN outperforms EMT and achieves the SoTA results on the blind, held out test set of ShARC [98]. Specifically, DISCERN performs well on simple in-line conditions and conjunctions of rules while still needing improvements on understanding disjunctions.

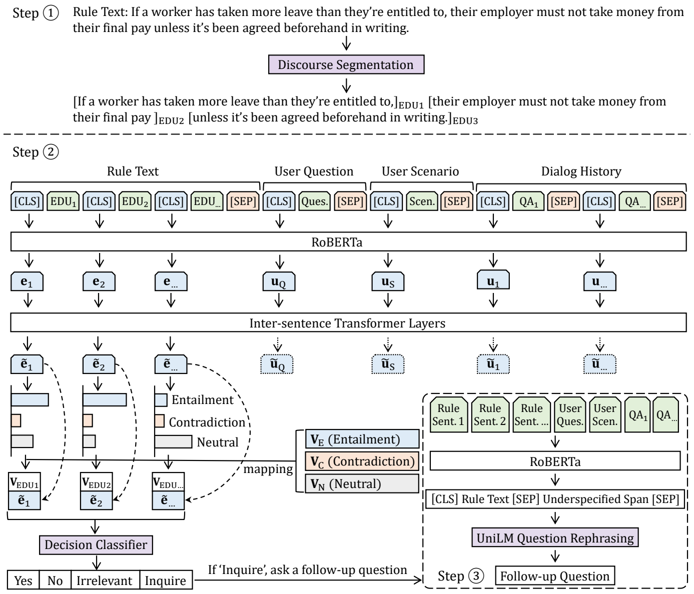

Step ①
Rule Text: If a worker has taken more leave than they're entitled to, their employer must not take money from their final pay unless it's been agreed beforehand in writing.
Discourse Segmentation
[If a worker has taken more leave than they're entitled to,]EDU1 [their employer must not take money from their final pay]EDU2 [unless it's been agreed beforehand in writing.]EDU3
Step ②
Rule Text
User Question
User Scenario
Dialog History
[CLS] EDU1 [CLS] EDU2 [CLS] EDU... [SEP] [CLS] Ques. [SEP] [CLS] Scen. [SEP] [CLS] QA1 [SEP] [CLS] QA... [SEP]
RoBERTa
e1 e2 e... uQ uS u1 u...
Inter-sentence Transformer Layers
ẽ1 ẽ2 ẽ... ũQ ũS ũ1 ũ...
Entailment
Contradiction
Neutral
mapping
VE (Entailment)
VC (Contradiction)
VN (Neutral)
VEDU1 ẽ1
VEDU2 ẽ2
VEDU... ẽ...
Decision Classifier
Yes No Irrelevant Inquire
If 'Inquire', ask a follow-up question
Rule Sent. 1
Rule Sent. 2
Rule Sent. ...
User Ques.
User Scen.
QA1
QA...
RoBERTa
[CLS] Rule Text [SEP] Underspecified Span [SEP]
UniLM Question Rephrasing
Step ③
Follow-up Question

Figure 9: Taking the segmented EDUs (conditions), user question, user scenario, and dialog history as inputs, DISCERNreasons out the decision among Yes, No, Irrelevant and Inquire. For Inquire decision, the question generation model asks a follow-up question.

1.3.2 Community Question Answering

Community question answering (cQA) is an evolution of the traditional question answering (QA), in the Web context, where users pose questions and then receive answers from other users. This setup is attractive as the anonymity on the

Web allows users to ask just about anything and then hope to get some honest answers from a number of people. On the negative side, there is no guarantee about the quality of the answers as people of very different background, knowledge, and with different motivation contribute answers to a given question.

In cQA, three tasks are of special relevance when a user poses a new question to the website (fig. 10): (*a*) determine whether a comment within a question-comment thread is a good answer to the question of that thread (*i.e., answer goodness*), (*b*) find related questions to the new question (*i.e., question-question similarity*), and (*c*) find relevant answers to the new question (i.e., *answer selection*). These tasks are interrelated as the information needed to answer a new question is usually found in the good comments of highly related questions.

q: "**How can I extend a family visit visa?**"

q_i: "Dear All; I wonder if anyone knows the procedure how I can extend the family visit visa for my wife beyond 6 months. I already extended it for 5 months and is 6 months running. I would like to get it extended for couple of months more.Any suggestion is highly appreciable.Thanks"

c^i_m: "You can get just another month's extension before she completes 6 months by presenting to immigration office a confirmed booking of her return ticket which must not exceed 7 months."

Figure 10: Example of three pieces of information in cQA problems: q is a newly-posed question, c^i_m denotes the m-th comment ($m \in \{1, 2, \ldots, M\}$) in the answer thread for the i-th potentially related question q_i ($i \in \{1, 2, \ldots, I\}$) retrieved from the forum.

In [34, 7, 45], we focused on task (*a*) , *i.e.,* classifying comments of an answer-thread as *good* vs. *bad* answers with respect to the thread question. This is a real problem, as a question can have hundreds of comments, the vast majority of which would not satisfy the users' information needs. Thus, finding the desired information in a long list of answers might be very time-consuming. The traditional approach learns a local classifier and uses it to predict for each comment separately. In contrast, we postulate that in a cQA setting, the answers from different users in a common thread are strongly interconnected and, thus, a joint answer selection model should be adopted to achieve higher accuracy. We model the thread-level dependencies in two different ways: (*i*) by designing specific features that are able to capture the dependencies between the answers in the same thread [7]; and (*ii*) by treating the task as **joint learning** (with global inference) over a fully-connected graph [34, 45]. I proposed two novel joint learning models that are on-line and integrate inference within learning. The first one jointly learns two *node-* and *edge*-level MaxEnt classifiers with stochastic gradient descent and integrates the inference step with loopy belief propagation. The second model is an instance of fully connected pairwise CRFs (FCCRF), which performs a global normalization of the functions. The FCCRF model significantly outperforms all other approaches and yields the best results on the task to date. Crucial elements for its success are the global normalization and an Ising-like edge potential.

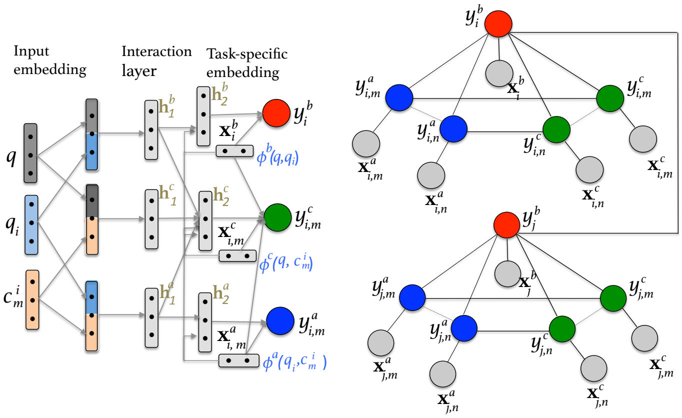


Figure 11: Graphical representation of our cQA model. On the left, we have three feed-forward neural networks to learn task-specific embeddings for the three cQA subtasks. On the right, a global CRF models intra- and inter-subtask dependencies.

Later in [46], we consider solving tasks (*b*) and (*c*) jointly with the help of task (*a*) in a joint **multi-task learning** framework (fig. 11). My approach has two steps. First, a deep neural net (DNN) in the form of a feed-forward neural network is trained to solve each of the three individual tasks, and the task-specific hidden layer activations are taken as embedded feature representations to be used in the second step. Then, a structured conditional model, a CRF, uses these embeddings and performs joint learning with global inference to exploit the dependencies between the different tasks. Previous work had mostly relied on recurrent or recursive architectures to propagate information through hidden layers, but had been disregarding the modeling strength of structured conditional models, which use global inference to model consistency in the output structure (*i.e.,* the class labels of all nodes in a graph). We explore the idea that combining simple DNNs with structured conditional models can be an effective and efficient approach for cQA subtasks that offers the best of both worlds. Our experimental results show that: (*i*) DNNs already perform very well on the question-question similarity and answer selection subtasks; (*ii*) strong dependencies exist between the subtasks under study, especially answer-goodness and question-question-relatedness influence answer-selection significantly; (*iii*) the CRFs exploit the dependencies between subtasks, providing sizeably better results that are on par or above the state of the art.

To allow effective access to the output of the cQA system, we also designed a web-based interactive cQA interface. The whole system was evaluated with real forum users and the findings were published in an IUI paper [31].

1.3.3 Data Augmentation for Sequence Labeling

Many tasks in NLP involve sequence labeling including syntactic and semantic tasks such as POS tagging and named entity recognition (NER). Neural models have outperformed traditional ML models on these tasks. However, they remain to be data hungry. Acquiring large annotated data can be expensive and prohibitive.

In our work [15], we propose a simple generation-based data augmentation method for low-resource sequence labeling tasks. Our method first linearizes the labeled sentences (fig. 12). Then a conditional language model (CLM) is trained on the linearized data and used to generate synthetic labeled data. Unlike employing weak taggers to label unseen data, our method unifies the processes of sentence generation and labeling using a CLM. Concretely, a word and its tag in a pair (*e.g.,* "B-PER Jose") are trained to be generated together. Our method does not require additional resources like gazetteer. Nevertheless, if unlabeled data or knowledge bases are available, it is also flexible to utilize these resources with a simple but effective conditional generation technique. Our method consistently outperforms the baselines in both supervised and semi-supervised settings on NER, POS tagging and target based sentiment analysis.

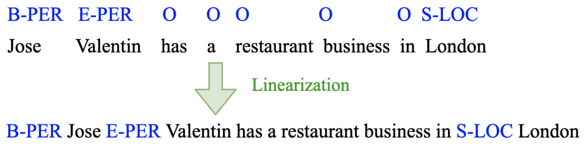

B-PER E-PER O O O O O S-LOC
Jose Valentin has a restaurant business in London
Linearization
B-PER Jose E-PER Valentin has a restaurant business in S-LOC London

Figure 12: An example of labeled sentence linearization. All words and their tags are paired up by inserting tags before (or after) the words (*O* tags removed).

1.3.4 Opinion Analysis

Fine-grained opinion mining involves: (*i*) identifying the opinion holder, (*ii*) identifying the target or aspect of the opinion, (*iii*) detecting opinion expressions, and (*iv*) measuring the intensity and sentiment of the opinion expressions. For example, in the sentence "John says, the hard disk is very noisy", John, the opinion holder, expresses a very negative opinion towards the target "hard disk" using the opinionated expression "very noisy". In an early work [71], we propose a general class of models based on Recurrent Neural Network (RNN) and word embeddings, that can be successfully applied to fine-grained opinion mining tasks without any task-specific feature engineering effort.

Our results on the task of opinion target extraction show that word embeddings improve the performance of state-of-the-art CRF models, when included as additional features. They also improve RNNs when used as pre-trained word vectors and fine-tuning them on the task gives the best results. A comparison between models demonstrates that RNNs outperform CRFs, even when they use word embeddings as the only features. Incorporating simple linguistic features into RNNs improves the performance even further. Our best results with LSTM RNN outperform the top performing system on the Laptop dataset and achieve the second best on the Restaurant dataset in SemEval-2014.

2 Multilingual Processing

With the advent of deep learning, NLP systems have seen remarkable advances in recent years. But they rely heavily on data-hungry models. Due to the availability of the data, these systems have been developed mostly for English and a handful of other high-resource languages like Chinese, French and German. However, there are more than 7,100 different languages, most of which have low/no resources (few/no labeled data, small/no Wikipedia, few/no online documents). Building systems only for the high-resource languages deprives a large part of the world population from language technologies. Fortunately, many of the languages do share a considerable amount of underlying structure at different linguistic levels (*e.g.,* vocabulary, word order). A significant part of my current research focuses on multilingual NLP, where the goal is to develop systems that perform well for diverse languages under low-resource conditions, while addressing questions of scientific interest about languages and their structural and functional properties. My interests lie in developing both supervised and unsupervised MT systems and general multilingual NLP models.

2.1 Machine Translation & Its Evaluation

Machine Translation (MT) has been considered as a flagship task in neural NLP that involves both language understanding and generation. There is a huge need for MT, both for humanity and for commerce. In our lab, we work on different aspects of MT research including novel model architectures, data augmentation, discourse or contextual MT, domain adaptation, bilingual dictionary induction, semi-supervised and unsupervised MT, and MT evaluation.

2.1.1 Supervised MT

Our research on supervised MT (*i.e.,* learning from parallel data) has focused on enhancements from the perspectives of both models and data, as I describe below.

(a) Tree-based Encoding Incorporating hierarchical structures like constituency trees has been shown to be effective for various NLP tasks. However, it is evident that SoTA sequence-based models like the Transformer [112] struggle to

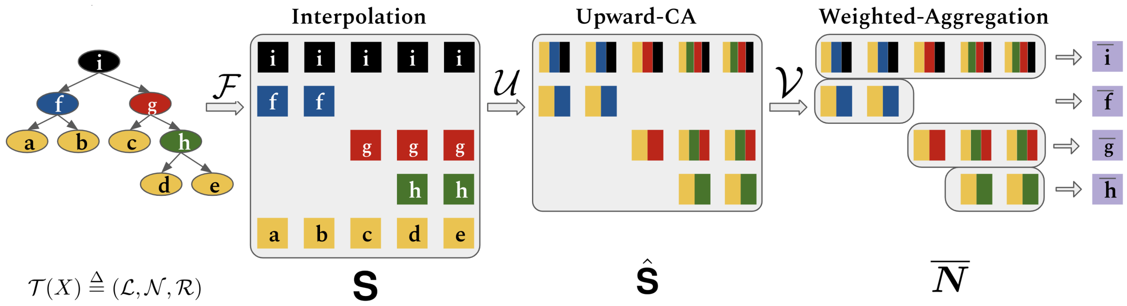


Figure 13: The hierarchical accumulation process of tree structures (best seen in colors). Given a parse tree, it is interpolated into a tensor $\mathbf{S}$, which is then accumulated vertically from bottom to top to produce $\hat{\mathbf{S}}$. Next, the (branch-level) component representations of the non-terminal nodes are combined into one representation as $\overline{\boldsymbol{N}}$ by weighted aggregation. Multi-colored blocks indicate accumulation of nodes of respective colors.

encode such structures inherently. On the other hand, dedicated models like the Tree-LSTM [108], while explicitly modeling hierarchical structures, do not perform as efficiently as the Transformer. In our ICLR-20 paper [92], we propose a novel attention-based method that encodes trees in a bottom-up manner and executes competitively with the Transformer at constant parallel time complexity. As attentions typically have query, key and value components, our model uses **hierarchical accumulation** to encode the *value* component of each non-terminal node by aggregating the hidden states of all of its descendants. The accumulation process is three-staged (fig. 13). First, we induce the value states of non-terminals with *hierarchical embeddings*, which help the model become aware of the hierarchical and sibling relationships between the nodes. Second, we perform an *upward cumulative-average* operation on each target node, which accumulates all elements in the branches originating from the target node to its descendant leaves. Third, these branch-level representations are combined into a new value representation of the target node by using *weighted aggregation*. Finally, the model proceeds to perform attention with *subtree masking* where the attention score between a non-terminal query and a key is activated only if the key is a descendant of the query.

We adopt our methods within the Transformer architecture and show improvements in the IWSLT'13 and WMT'14 English ↔ German and English ↔ French translation benchmarks. Our model also exhibits advantages over Tree-LSTM in classification tasks including Stanford Sentiment Treebank, IMDB Sentiment Analysis and Subject-Verb Agreement.

(b) Differentiable Attention Window. Rather than defining attentions over trees, we focus on attentions over local windows (*i.e.,* ngrams of tokens) in our subsequent work. Particularly, our work [88] focuses on improving attentions with differentiable windows. The key idea is to enable more *focused* attention, leveraging dynamic window selection for limiting (and guiding) the search space for the standard attention modules to work within. This can also be interpreted as performing a form of dynamic local attention. We make several key technical contributions. First, we formulate the dynamic window selection problem as a problem of learning a *discrete* mask, *i.e.,* binary values representing the window (fig. 14). By learning and composing left and right boundaries, we show that we are able to parameterize the (discrete) masking method. We then propose *soft* adaptations of the above mentioned, namely **trainable soft masking** and **segment-based soft masking**, which are differentiable approximations that can not only be easily optimized in an end-to-end fashion, but also inherit the desirable properties of discrete masking.

While these modules are task and model agnostic, we imbue the Transformer [112] model with our differentiable window-based attention. To this end, we propose two further variants, *i.e., multiplicative window attention* and *additive window attention* for improving the Transformer model. We evaluate our approach on a potpourri of NLP tasks, namely *machine translation, sentiment analysis, language modeling,* and *subject-verb agreement*. Extensive experimental results on these tasks demonstrate the effectiveness of our proposed method.

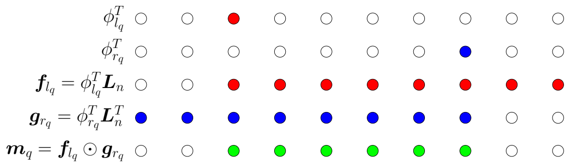

$\phi^T_{l_q}$
$\phi^T_{r_q}$
$\boldsymbol{f}_{l_q} = \phi^T_{l_q}\boldsymbol{L}_n$
$\boldsymbol{g}_{r_q} = \phi^T_{r_q}\boldsymbol{L}^T_n$
$\boldsymbol{m}_q = \boldsymbol{f}_{l_q} \odot \boldsymbol{g}_{r_q}$

Figure 14: Example of discrete masking with left and right boundary prediction models $\phi^T_{l_q}$ and $\phi^T_{r_q}$, and cumulative sums $\boldsymbol{f}_{l_q}$ and $\boldsymbol{g}_{r_q}$, and finally how the mask vector $\boldsymbol{m}_q$ can be derived from $\boldsymbol{f}_{l_q}$ and $\boldsymbol{g}_{r_q}$ for $l_q = 3$ and $r_q = 8$.

(c) Data Augmentation While the invention of novel architectures has been fundamental to MT progress, other *non-intrusive extensions* that do not modify the model architecture intensively like sub-word tokenization [104] to deal with out-of-vocabulary (OOV) problem or *back-translation* [103] to exploit extra monolingual data, have been crucial in advancing MT research. In our NeurIPS-20 paper [93], we propose **Data Diversification**, a simple but effective way to improve MT consistently and significantly. In this method, we first train multiple models on both backward (target→source) and forward (source→target) translation tasks. Then, we use these models to generate a diverse set of synthetic training data from both sides to augment the original data. Our method achieved the SoTA (at the time of publication) in the WMT'14 English-German and English-French translation tasks. Furthermore, it gives 1.0-2.0 BLEU gains in 4 IWSLT tasks (English↔German and English↔French) and 4 low-resource tasks (English↔Sinhala and English↔Nepali). We demonstrate that data diversification outperforms other related methods – knowledge distillation [58] and dual learning [117], and is complementary to back-translation in a semi-supervised setup. Our analysis further reveals that the method is correlated with ensembles of models and it sacrifices perplexity for better BLEU.

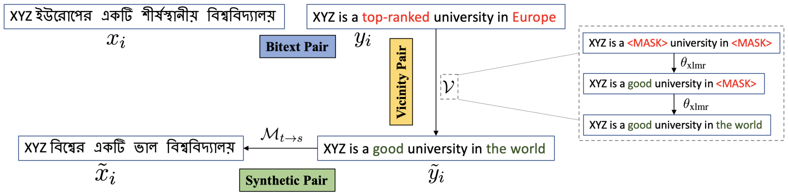

XYZ ইউরোপের একটি শীর্ষস্থানীয় বিশ্ববিদ্যালয়
x_i
XYZ is a top-ranked university in Europe
y_i
Bitext Pair
Vicinity Pair
$\mathcal{V}$
XYZ is a <MASK> university in <MASK>
θ_{xlmr}
XYZ is a good university in <MASK>
θ_{xlmr}
XYZ is a good university in the world
$\mathcal{M}_{t\rightarrow s}$
XYZ বিশ্বের একটি ভাল বিশ্ববিদ্যালয়
$\tilde{x}_i$
XYZ is a good university in the world
$\tilde{y}_i$
Synthetic Pair

Figure 15: Illustration of AugVic for Bengali-to-English translation. Here (x_i, y_i) is the original bitext pair, $\tilde{y}_i$ is a vicinal sample of y_i, and $(\tilde{x}_i, \tilde{y}_i)$ is a synthetic pair where $\tilde{x}_i$ is generated by a reverse intermediate translation system $\mathcal{M}_{t\rightarrow s}$. Right side of the figure shows the successive steps of vicinal sample generation.

Since the performance of Data Diversification depends on the performance of the base MT models, the improvements on low-resource languages are generally lower compared to high-resource ones, as the base MT models are much weaker in low-resource translation tasks. Back-translation (BT) has proved to be quite successful when sufficient in-domain monolingual data is available. However, when such data is scarce, which is indeed a common situation in low-resource settings, the success of BT is limited. BT can also suffer when there is a *domain mismatch* between the training and test domains. In a recent work [79], we propose **AugVic**, a novel method to **aug**ment **vic**inal samples around the bitext distribution (fig. 15). Instead of using extra monolingual data, it aims to leverage the vicinal samples of the original bitext, thereby enlarging its support to improve model generalization. With the goal of training a source-to-target NMT system, AugVic augments vicinal samples in the target language. The vicinal samples are generated by predicting the masked tokens of a target bitext sentence using a pretrained large-scale language model. To generate synthetic bitext data from these augmented vicinal samples through a reverse intermediate (target-to-source) model, we propose two different methods: the first one is based on the traditional BT, while the second one leverages the original source sentence as a guide. Our results show significant improvements over the bitext baselines with 2.76 BLEU gains on an average on eight different translation tasks without using any extra monolingual data. AugVic also complements traditional BT with additive gains when extra monolingual data is used. We also show AugVic's efficacy in bridging the gap between in-domain and out-of-domain performance in traditional back-translation with monolingual data.

(d) Discourse-based MT Thanks to the attention mechanism, NMT models such as the Transformer [112] can model much broader context. While initially translation was still done in a sentence-by-sentence fashion, researchers soon realized that going beyond that has become easier, and recent work has successfully exploited this. This is an exciting research direction as it can help address discourse phenomena such as anaphora, gender agreement, lexical consistency, and text coherence. However, it has been shown that even with the broader context, NMT models still fail on these aspects [102]. They tend to prefer a more typical alternative to a relatively rare but correct one (*e.g.,* French "*Il*" is often wrongly translated to the more common "*it*" than "*he*"). However, these seemingly trivial errors can erode translation to the extent that they can be easily distinguishable from human-translated texts [60].

There could be several reasons for why NMT models make such mistakes. In our work [52], we hypothesize that since almost all NMT models are trained with a conditional language model objective, it is clear that this objective alone is proving inadequate to capture all of the information available in the text. We therefore propose a class of conditional **generative-discriminative** hybrid losses that explicitly teach models what to generate and what not to generate. Specifically, we target the improvement of pronoun translation by focusing our fine-tuning efforts through our proposed objectives and also through leveraging the training data by extracting a subset of targeted fine-tuning data that the model has failed to learn correctly from. We show improvements both in general translation quality and in the pronoun translation without compromising on either, and we do this without any elaborate model architecture.

(e) Domain Adaptation for MT: Prior to the end-to-end NMT paradigm, a notably successful attempt on using neural networks for MT was the Neural Network Joint Model (NNJM) [14], which augments streams of source with target n-grams and learns a neural model over the vector representation of such streams. Impressive gains were achieved with NNJM used as an additional feature in the SMT decoder. In [42, 51], we extended NNJM for domain adaptation in order to leverage the huge amount of out-of-domain data coming from heterogeneous sources. We carry out our research in two ways: (*i*) we apply state-of-the-art domain adaptation techniques, such as mixture modeling and data selection using the NNJM, and (*ii*) we propose two novel methods to perform adaptation through instance weighting and weight readjustment in the NNJM framework. Our first method uses data dependent regularization in the loss function to perform (soft) data selection, while the second method fuses the in- and the out-domain models to readjust their parameters. Our evaluation on standard translation tasks demonstrates that the adapted models outperform the non-adapted baselines and the deep fusion model outperforms the other neural adaptation methods as well as phrase-table adaptation techniques. We also demonstrate that our methods are complementary to the existing methods.

2.1.2 Unsupervised and Semi-supervised MT

Although recent neural approaches to MT [112] have advanced the state of the art, they continue to rely heavily on large parallel data. Such large-scale parallel data is not always available, especially for low-resource languages. Therefore, the search for unsupervised and semi-supervised alternatives using monolingual data has been active. In this regard, our research spans both word-level and sentence-level translations as I briefly describe below.

(a) Word Translation & Cross-lingual Embeddings Most recent successful methods for Word Translation or Bilingual Lexicon Induction (BLI) are mapping-based, where a mapping function is learned to transform the word embeddings in a source language to the corresponding embeddings in the target language. This gives *cross-lingual word embeddings* (CLWE), where words with similar meanings are represented by similar vectors regardless of their actual language. CLWE enable comparing the meaning of words across languages, which is key to BLI and other multi-lingual applications such as unsupervised MT and multi-lingual retrieval. They also play a crucial role in knowledge transfer between languages (*e.g.,* from high to low resource languages) by providing a common representation space.

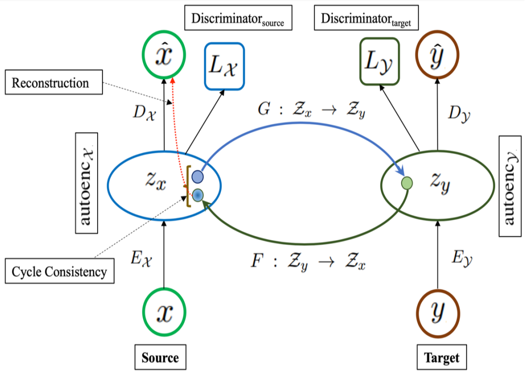


Figure 16: Adversarial autoencoder for CLWE.

Adversarial training has shown impressive success in learning CLWE without any parallel data (*i.e.,* unsupervised). However, recent work has shown superior performance for non-adversarial methods in more challenging language pairs. Also, most predominant methods learn a *linear* mapping function with the assumption that the word embedding spaces of different languages exhibit similar geometric structures (*i.e.,* approximately *isomorphic*). However, several recent studies have criticized this simplified assumption showing that it does not hold in general even for closely related languages. In our work [81, 80], we revisit adversarial training and propose a number of key improvements that yield more robust training and improved mappings. Our main idea is to learn the cross-lingual mapping in a projected latent space and add more constraints to guide the unsupervised mapping in this space. We accomplish this by proposing a novel **adversarial autoencoder** framework, where adversarial mapping is done at the (latent) code space as opposed to the original embedding space (fig. 16). This gives the model the flexibility to automatically induce the required geometric structures in its latent code space that could potentially yield better mappings. By mapping the latent vectors through adversarial training, our approach therefore departs from the isomorphic assumption.

In our adversarial training, not only the mapper but also the target encoder is trained to fool the discriminator. This forces the discriminator to improve its discrimination skills, which in turn pushes the mapper to generate indistinguishable translation. To guide the mapping, we include two additional constraints. Our first constraint enforces *cycle consistency* so that code vectors, after being translated from one language to another, and then translated back to their source space, remain close to the original vectors. The second constraint ensures *reconstruction* of the original input word embeddings from the back-translated codes. This grounding step forces the model to retain word semantics during the mapping process. Extensive experimentations with high- and low-resource languages from two different datasets show that our method achieves better performance than existing adversarial and non-adversarial approaches and is also competitive with the supervised system.

While not requiring any cross-lingual supervision makes the unsupervised methods attractive, a recent study [114] shows that these methods lack robustness and fail for a large number of languages. In our recent work [78], we propose LNMAP (**L**atent space **N**on-linear **Map**ping), a novel semi-supervised approach that uses minimal supervision from a seed dictionary, while leveraging semantic information from the monolingual word embeddings. LNMAP comprises two *autoencoders* (similar to fig. 16 but does not use adversarial training), which are first pre-trained independently in a self-supervised way to induce the latent code space of the respective languages. Then, we use a small seed dictionary to learn the non-linear mappings between the two code spaces. Our experiments with 15 different language pairs (in both directions) comprising high- and low-resource languages show significant improvements for LNMAP over SoTA in most of the tested scenarios. It is particularly very effective for low-resource languages; for example, using 1K seed dictionary, LNMAP yields about 18% absolute improvements on average over a SoTA supervised method.

(b) Unsupervised Sentence-level MT The goal of unsupervised (sentence-level) MT or UMT is to learn an MT model using only monolingual data. The standard UMT framework follows three main principles: model initialization, language modeling and iterative back-translation. Model initialization bootstraps the model with a knowledge prior like word-level transfer as described above. Language modeling, which takes the form of denoising auto-encoding (DAE) trains the model to generate plausible sentences in a language. Meanwhile, iterative back-translation (IBT) facilitates cross-lingual translation training by generating noisy source sentences for original target sentences. In our recent paper [94], we focus on a different aspect of the UMT framework, namely, its data diversification. If we look from this view, the DAE and IBT steps perform some form of data diversification to train the model. However, we conjecture that these diversification methods may have reached their limit as the performance does not improve further the longer we train the UMT models. In our work, we introduce a fourth principle to the standard framework: **Cross-model Back-translated Distillation** or CBD, with the aim to induce another level of diversification that the existing principles lack. CBD initially trains two UMT agents (models) using existing approaches. Then, one of the two agents translates the monolingual data from one language s to another t in the first level. In the second level, the generated data are back-translated from t to s by the *other agent*. In the final step, the synthetic parallel data created by the first and second levels are used to distill a supervised MT model. CBD is applicable to any existing UMT method and is more efficient than ensembling methods. CBD establishes the SoTA in the bilingual UMT tasks of WMT'14 English-French, WMT'16 English-German and WMT'16 English-Romanian. Without large scale pre-trained models and data, it shows consistent improvements of 1.0-2.0 BLEU compared to the baseline. It also boosts the performance on IWSLT tasks significantly. In our analysis, we explain with experiments why other similar variants and other alternatives from literature do not work well and cross-model back-translation is crucial for our method. We further demonstrate that CBD enhances the baselines by achieving greater diversity as measured by back-translation BLEU.

2.1.3 MT Evaluation

(a) Evaluation of Pronoun Translations As mentioned, the neural revolution in MT has made it easier to model larger contexts beyond the sentence-level, which can potentially help resolve some discourse-level ambiguities such as pronominal anaphora. Unfortunately, even when the resulting improvements are seen as substantial by humans, they remain virtually unnoticed by traditional automatic evaluation measures such as BLEU, as only a few words end up being affected. It has long been argued that as the quality of machine translation improves, there will be a singularity moment when existing evaluation measures would be unable to tell whether a given output was produced by a human or by a machine. Indeed, there have been recent claims that human parity has already been achieved, but it has also been shown that it is easy to tell apart a human translation from a machine output when going beyond the sentence level [60]. Overall, it is clear that there is a need for machine translation evaluation measures that look beyond the sentence level, and thus can better appreciate the improvements that a discourse-aware MT system could potentially bring.

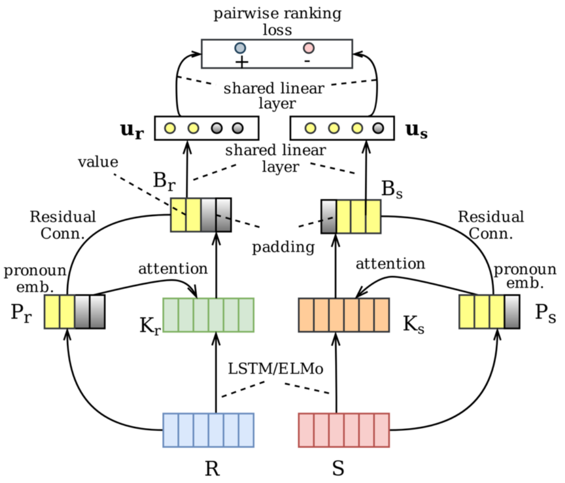


Figure 17: Our proposed framework to differentiate good pronoun translations from bad ones in context [53].

With this aim in mind, in [53], we contribute an extensive, targeted dataset that can be used as a test suite for pronoun translation, covering multiple source languages and different pronoun errors drawn from real system translations, for English. We further present a specialized evaluation measure (fig. 17) trained on this dataset. The measure performs pairwise evaluations: it learns to distinguish good vs. bad translations of pronouns, without being given specific signals of the errors. Our user study shows that the evaluation measure achieves high agreement with human judgments.

(b) Neural Pairwise MT Evaluation In another front [27, 28], earlier we presented a framework for MT evaluation using neural networks in a pairwise setting, where the goal is to select the better translation from a pair of hypotheses, given the reference translation. In this framework, lexical, syntactic and semantic information from the reference and the two hypotheses are embedded into small distributed vector representations, and fed into a multi-layer perceptron that models non-linear interactions between each of the hypotheses and the reference, as well as between the two hypotheses. We experiment with benchmark datasets from the WMT Metrics shared task, on which we obtain the best results published so far, with the basic network configuration. We also perform a series of experiments to analyze and understand the contribution of the different components of the network. We evaluate variants and extensions including, among others: fine-tuning of the semantic embeddings, and sentence-based representations modeled with recurrent neural networks. The proposed framework is flexible and generalizable, allows for efficient learning and scoring, and provides an MT evaluation metric that correlates with humans on par with the state of the art.

2.2 Multi-lingual Modeling

In recent years there has been an increase in the number of methods that attempt to learn general-purpose multilingual representations, which aim to capture shared knowledge across languages. Jointly trained deep contextualized multilingual LMs such as mBERT, XLM-R and mBART, coupled with supervised fine-tuning in the source (English) language, have been quite successful in transferring linguistic and task knowledge from one language to another without using any task labels in the target language (called **zero-shot transfer**). The joint pre-training with multiple languages allows these models to generalize across languages. Despite their effectiveness, recent studies have also highlighted one crucial limiting factor for successful cross-lingual transfer. They all agree that the cross-lingual generalization ability of the model is limited by the (lack of) structural similarity between the source and target languages. For example, for transferring mBERT from English, [54] report about 23.6% accuracy drop in Hindi (structurally dissimilar) compared to 9% drop in Spanish (structurally similar) in cross-lingual natural language inference (XNLI). The difficulty level of transfer is further exacerbated if the (dissimilar) target language is low-resourced, as the joint pretraining step may not have seen many instances from this language in the first place.

One attractive way to improve cross-lingual generalization is to perform *data augmentation*, and train the model on examples that are similar but different from the labeled data in the source language. Back-translation [103] has been a successful method but requires parallel data to train effective machine translation systems, acquiring which can be more expensive for low-resource languages than annotating the target language data. In our work [5], we propose UXLA (fig. 18), a robust **u**nsupervised **cross-lingual** **a**ugmentation framework for improving cross-lingual generalization of multilingual LMs. UXLA augments data from the unlabeled training examples in the target language as well as from the virtual input samples generated from the vicinity distribution of the source and target language sentences. With the augmented data, it performs simultaneous *self-learning* with an effective *sample distillation* to learn a strongly adapted cross-lingual model from noisy (pseudo) labels for the target language task. We propose novel ways to generate virtual sentences using a multilingual masked LM, and get reliable task labels by simultaneous multilingual co-training. This co-training employs a two-stage co-distillation process to ensure robust transfer to dissimilar and/or low-resource languages. We perform extensive experiments on three diverse zero-resource cross-lingual transfer tasks – XNER, XNLI, and PAWS-X, and across many (14 in total) language pairs comprising languages that are similar/dissimilar/low-resourced. UXLA yields impressive results on XNER, setting SoTA in all tested languages and outperforming the baselines by a good margin. The relative gains for UXLA are particularly higher for structurally dissimilar and/or low-resource languages: 28.54%, 16.05%, and 9.25% absolute improvements for Urdu, Burmese, and Arabic, respectively. For XNLI, with only 5% labeled data in the source, it gets comparable results to the baseline that uses all the labeled data, and surpasses the standard baseline by 2.55% on average when it uses all the labeled data in the source. We also have similar findings in PAWS-X. We provide a comprehensive analysis of the factors that contribute to UXLA's performance.

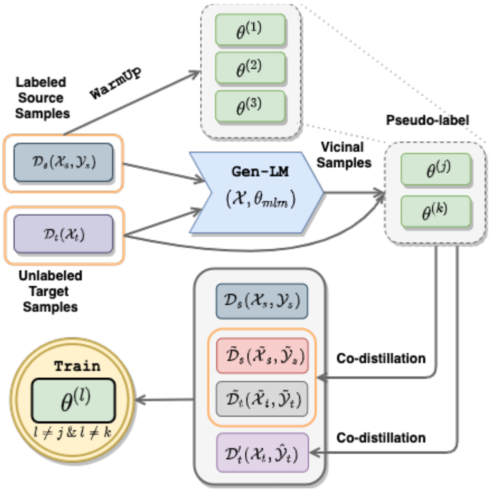

$\theta^{(1)}$
$\theta^{(2)}$
$\theta^{(3)}$
Labeled Source Samples
WarmUp
Pseudo-label
Vicinal Samples
$\mathcal{D}_s(\mathcal{X}_s, \mathcal{Y}_s)$
Gen-LM
$(\mathcal{X}, \theta_{mlm})$
$\theta^{(j)}$
$\theta^{(k)}$
$\mathcal{D}_t(\mathcal{X}_t)$
Unlabeled Target Samples
$\mathcal{D}_s(\mathcal{X}_s, \mathcal{Y}_s)$
$\tilde{\mathcal{D}}_s(\tilde{\mathcal{X}}_s, \tilde{\mathcal{Y}}_s)$
Co-distillation
$\tilde{\mathcal{D}}_t(\tilde{\mathcal{X}}_t, \tilde{\mathcal{Y}}_t)$
Train
$\theta^{(l)}$
$l \neq j \,\&\, l \neq k$
Co-distillation
$\mathcal{D}'_t(\mathcal{X}_t, \hat{\mathcal{Y}}_t)$

Figure 18: Training flow of UXLA. After training the base task models $\theta^{(1)}$, $\theta^{(2)}$, and $\theta^{(3)}$ on source labeled data $\mathcal{D}_s$ (**WarmUp**), we use two of them ($\theta^{(j)}$, $\theta^{(k)}$) to **pseudo-label** and **co-distill** the unlabeled target language data ($\mathcal{D}'_t$). A pretrained LM (**Gen-LM**) is used to generate new vicinal samples for both source and target languages, which are also pseudo-labeled and co-distilled using the two task models ($\theta^{(j)}$, $\theta^{(k)}$) to generate $\tilde{\mathcal{D}}_s$ and $\tilde{\mathcal{D}}_t$. The third model $\theta^{(l)}$ is then progressively trained on these datasets: $\{\mathcal{D}_s, \mathcal{D}'_t\}$ in epoch 1, $\tilde{\mathcal{D}}_t$ in epoch 2, and all in epoch 3.

In a concurrent work [123], we propose a novel fine-tuning method based on co-training that aims to learn more generalized semantic equivalences as complementary to multilingual language modeling (*e.g.,* masked LM) using the unlabeled data in the target language. We also propose an adaption method based on contrastive learning to better capture the semantic relationship in the parallel data, when a few translation pairs are available. We report significant gains compared to directly fine-tuning multilingual pre-trained models and other semi-supervised alternatives.

In another recent work [69], we consider a low-resource setting for cross-lingual NER, where there is limited source-language training data and no target-language train/dev data. We first introduce a novel labeled sequence translation method to translate the training data to the target language as well as to other languages (fig. 19). Compared with exiting methods, our labeled sentence translation approach leverages placeholders for label projection, which effectively avoids many issues faced during word alignment, such as word order change, entity span determination, noise-sensitive similarity metrics and so on. This allows us to finetune the LM based NER model on multilingual data rather than on the source-language only. Note however that this instance-based transfer add limited semantic variety to the training set, since they only translate entities and the corresponding contexts to a different language. To add more diversity in the training data, we extend our previously proposed method [15] for monolingual data augmentation to multilingual data augmentation (fig. 12). Particularly, we train conditional LMs on multilingual labeled data and then use it to generate more synthetic multilingual training data. Through empirical experiments, we observe that when fine-tuning pretrained multilingual LMs for low-resource cross-lingual NER, translations to more languages can also be used as an effective data augmentation method, which helps improve performance of both the source and the target languages.

Previously in [6], we proposed a zero-resource transfer framework with two encoders – one for the source language and the other for the target. Our source model was based on a bidirectional LSTM-CRF, which we transfer to a target model in two steps. We first project the mono-lingual word embeddings to a common space through *word-level* adversarial training. The word-level mapping yields initial cross-lingual links between two languages but does not take any NER information into account. Transferring task information in the cross-lingual setup is specifically challenging because languages vary in the word order. To tackle this, we propose an *augmented fine-tuning* method with parameter sharing and feature augmentation, and jointly train the target model in supervision of the source model.

Labeled sentence in the source language:
[PER Jamie Valentine] was born in [LOC London].

1. Translate sentence with placeholders:
src: PER0 was born in LOC1.
tgt: PER0 nació en LOC1.

2. Translate entities with context:
PER0
src: [Jamie Valentine] was born in London.
tgt: [Jamie Valentine] nació en Londres.

LOC1
src: Jamie Valentine was born in [London].
tgt: Jamie Valentine nació en [Londres].

3. Replace placeholders with translated entities:
[PER Jamie Valentine] nació en [LOC Londres].

Figure 19: An example of labeled sentence translation.

Prior Work on Cross-lingual cQA. In our prior work [50, 76], we studied the problem of question-question similarity reranking in community Question Answering (cQA), when the input question can be either in English or in Arabic, and the questions it is compared to are always in English. We start with a simple language-independent representation based on cross-language word embeddings, which we input into a feed-forward multilayer neural network to classify pairs of questions, (English, English) or (Arabic, English), regarding their similarity. Furthermore, we explore the question of whether adversarial training can be used to improve the performance of the network when we have some unlabeled examples in the target language (fig. 40). In our setup, the basic task-solving network is paired with another network that shares the internal representation of the input and tries to decide whether the input example comes from the source (English) or from the target (Arabic) language. The training of this language discriminator network is adversarial with respect to the shared layers by using gradient reversal during backpropagation, which makes the training to maximize the loss of the discriminator rather than to minimize it. The main idea is to learn a high-level abstract representation that is discriminative for the main classification task, but is invariant across the input languages.

We show that using the unlabeled data for adversarial training allows us to improve the results by a sizable margin in both directions, *i.e.,* when training on English and adapting the system with the Arabic unlabeled data, and vice versa. Moreover, the resulting performance is comparable to the best monolingual English systems at SemEval. We also compare our unsupervised model to a semi-supervised model, where we have some labeled data for the target language. To the best of our knowledge, ours is the first to show promising results with adversarial training for cross-lingual representation learning, moreover, we are not aware of any work on cross-language question reranking for cQA prior to our work.

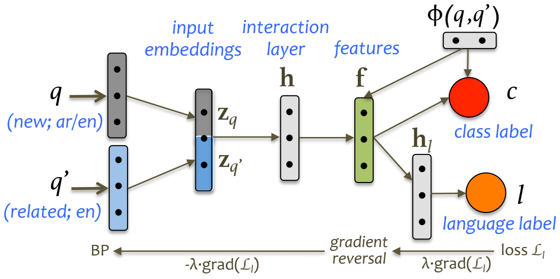


Figure 20: Architecture of CLANN for the question to question similarity problem in cQA.

3 Language Generation

Text generation has been a core problem in NLP. Thanks to the advances in neural architectures such as Transformers [112], models are now capable of generating texts that are of better quality than before. However, modern text generations systems still suffer from a number of key issues including dull and repetitive generation, hallucinated output and toxic output. Our work aims at mitigating these issues of neural text generation as we describe below.

3.1 Mitigating Degeneration & Coverage Issues in Text Generation Applications

Despite the countless efforts that have been made to improve neural architectures for language generation, models trained with the standard Maximum Likelihood Estimation (MLE) objective are known to prefer generating dull and highly repetitive texts, a problem known as *degeneration*. For instance, in *open-ended generation* tasks, such as story continuation or open dialogue generation, it has been observed that even with large pre-trained language models (LMs) like GPT-2 [96], high frequency tokens largely dominate the generation [119]. Similar observation has been reported in *directed generation* tasks such as summarization, image captioning and machine translation. In our recent paper [66], based on the known observation that the text generation models trained with MLE objective tend to generate repetitive tokens or phrases, we introduce a novel method called **ScaleGrad** for neural text generation training, by directly maneuvering the gradients to make the model learn to use novel tokens during training. Our method is training based (as opposed to decoding), which aims to address the *fundamental modeling* problem, that is, the token-level distribution predicted by the generation model. We conduct experiments with different neural architectures including

LSTM and Transformer across different tasks in opened-ended and directed text generation. Through extensive analysis we demonstrate that ScaleGrad consistently improves the generation quality according to both human evaluation and automatic metrics. Compared to other training based methods, ScaleGrad is architecturally simpler and easier to fit into current neural models, while possessing a wide applicability to different text generation tasks.

For directed generation tasks aimed at recovering the source message either fully or a compressed version of it (*e.g.,* summarization, MT), a major shortcoming of the existing encoder-decoder architectures is that they could keep covering some parts in the source while ignoring the other important concepts, thus resulting in less comprehensive coverage. Existing methods for improving the neural coverage either require extra parameters and loss to furnish the model with better learning capacity or place a specific bound on the sum of attention scores. On the other hand, monotone nondecreasing submodular objectives have been shown to be ideal for content selection in *extractive* text summarization and *statistical* MT [65]. Despite their appropriateness, submodular functions for content selection have so far been ignored in neural text generation models. In our work [29], we define a class of novel attention mechanisms called **diminishing attentions** with submodular functions and in turn, prove the submodularity of the effective neural coverage. The submodular maximization problem is generally approximated by greedy selection. However, it is not suited to optimizing attention scores in auto-regressive generation systems. We therefore put forward a simplified yet principled and empirically effective solution. By imposing *submodularity* on the coverage enforced by the decoder states on the encoder states, our diminishing attention method enhances the model's awareness of previous steps, leading to more comprehensive overall coverage of the source and maintaining a focus on the most important content when the goal is to generate a compressed version of the source (*e.g.,* text summarization). We further enhance our basic diminishing attention and propose **dynamic diminishing attention** to enable dynamically adapted coverage. Our results highlight the benefits of submodular coverage. Our diminishing attention mechanisms achieve SoTA results on three diverse directed text generation tasks, abstractive summarization, neural machine translation (NMT) and image-paragraph generation spanning across two modalities, three neural architectures and two training strategy variations.

3.2 Other Related Research on Text Generation

Transfer Learning for Summarization Recent advances in summarization have been mostly driven by the availability of large-scale datasets and by the introduction of large pretrained models. Creating data for every new domain is infeasible and expensive. Thus, the ability to transfer large pretrained models to new domains with little or no in-domain data is desirable, especially as such models make their way into real-world applications. In [17], we build on recent work in pretrained models and improve its zero- and few-shot capability by encoding characteristics of the target summarization dataset in unsupervised, intermediate fine-tuning data. We view the summarization process as a function of subaspects, which determine the output. We focus on the subaspects of *extractive diversity*, determined by how well an extractive model performs on the data, *compression ratio* between the source document and summary, and, the *lead bias* for news domains. We assume knowledge of the target domain such as the size of input documents and the desired summaries, and the abstractiveness of summaries, all of that can be treated as prior knowledge if the task is to be well-defined. We propose **WikiTransfer**, where we encode this knowledge into Wikipedia data by extracting *pseudo-summaries* of the desired length and filtering examples based on the desired level of abstraction. We use this (unsupervised) data to fine-tune a model to learn characteristics of the target dataset. We show that this method improves zero-shot domain transfer over transfer from other domains, achieving SoTA in unsupervised abstractive summarization. We also demonstrate the benefits of WikiTransfer in few-shot settings, and show additional improvements when applying it with data augmentation and a regularization term for training with potentially noisy augmented data. We show robustness in these settings and analyze differences in performance in both automatic and human assessments.

Early Work on Unsupervised Models for QA & Summarization In my M.Sc. [32], I investigated unsupervised methods to automatically answer both simple and complex questions. Simple questions (*e.g.,* "Who is the president of USA?") require small snippets of text as answers and are easier to answer than complex questions (*e.g.,* "Describe the after-effects of cyclone Cindy?") which entail richer information needs and require synthesizing information from multiple documents. My work on complex QA was published in a JAIR article [12] and in one conference paper [11]. I approached the task as a query-focused multi-document summarization and employed an extractive approach to select a subset of the original sentences as the answer. I experimented with one simple vector space model and two statistical unsupervised models for computing the importance of the sentences. The performance of these approaches depends on the features used and the weighting of these features. I extracted different kinds of informative features for each sentence, and use a gradient descent search to learn the feature-weights from a development set. I first showed that the tree kernel features based on the syntactic and shallow semantic trees of the sentences improve the performance of these models significantly, then I showed that with a large feature set and the optimal feature-weights, my unsupervised models perform as good as SoTA systems with the advantage of not requiring any human annotated data for training. In a separate but related work [10], I show that the syntactic and shallow semantic tree kernels can also improve the performance of the random walk model for answering complex questions.

3.3 Controllable Generation

Although large-scale LMs are able to learn the distribution of their training set well enough to generate realistic text, simply imitating the distribution of the training data during generation has many drawbacks; large-scale text training sets are crawled from the web which is imbued with toxicity, bias, hate, and misinformation. Methods for better controlling or filtering generation are valuable for making LMs trained on such data safer and more generally useful for downstream applications. Existing approaches to controlling LMs have limitations. For example, Class-conditional LMs (CC-LMs) such as CTRL [55] are limited in controlling what *not* to generate (*e.g.,* toxicity). Another approach is to use discriminators to steer generation, but existing methods to do this are very computationally intensive.

In [59], we present **GeDi** as an algorithm for efficiently guiding generation from large LMs to make them safer and more controllable. Our proposed method uses CC-LMs as generative discriminators (GeDis) to guide language generation towards desired attributes. We use GeDis to compute classification likelihoods for all candidate next tokens during generation using Bayes rule, saving many thousand-fold in computation as compared with using a standard (non-generative) discriminator to compute this for large vocabulary sizes. We then show how these likelihoods can guide generation from large language models via weighted decoding and filtering (fig. 21). Our experimental results verify the ability of GeDi to control generation in a variety of settings while maintaining linguistic quality on par with strong LMs. GeDi trained on sentiments of movie reviews can generate book text with a positive or negative tone better than or equivalently to SoTA baselines. It is able to significantly reduce the toxicity of GPT-2 and GPT-3 generation, without sacrificing linguistic quality. GeDi trained on a dataset of only 4 topics can generalize to new control codes zero-shot, allowing them to guide generation towards a wide variety of topics. It is also very computationally efficient for both training and inference compared to existing methods.

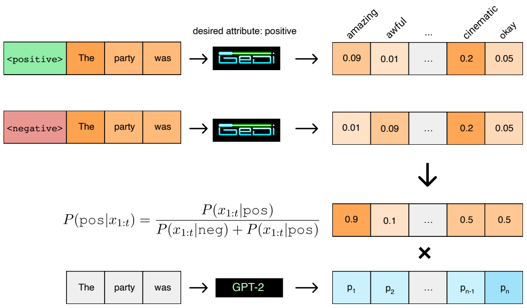


Figure 21: An example of how GeDi-guided generation uses Bayes rule to efficiently compute classification probabilities for possible next tokens at each generation step with only element-wise operations. These classification probabilities can then be used to guide generation from a language model (*e.g.,* GPT-2) to achieve attribute control across domains.

4 Robust & Fair NLP

Neural NLP systems have achieved outstanding performance on benchmark datasets. Many of these research advances have led to production systems for applications like MT, QA, speech recognition, and dialog. However, these models are only as good as the data they are trained on and they fail catastrophically or amplify discrimination against minority demographics when exposed to input from outside the training distribution. This lack of robustness exposes concerning limitations in existing models' language understanding capabilities, and creates problems when such systems are deployed to real users. With the aim to make NLP systems robust and ethical, we also work on different aspects of robust and fair NLP and I describe those in this section.

(a) L2 and Dialectal Variations Current NLP models seem to be trained with the implicit assumption that everyone speaks fluent (often U.S.) Standard English, even though two-thirds (>700 million) of the English speakers in the world speak it as a second language (L2). Even among native speakers, a significant number speak a dialect like African American Vernacular English (AAVE) rather than Standard English. These World Englishes exhibit variation at multiple levels of linguistic analysis. Therefore, putting these models directly into production without addressing this inherent bias puts them at risk of committing linguistic discrimination by performing poorly for many speech communities (*e.g.,* AAVE and L2 speakers). This could take the form of either failing to understand these speakers, or misinterpreting them. For example, the recent mistranslation of a minority speaker's social media post resulted in his wrongful arrest [30].

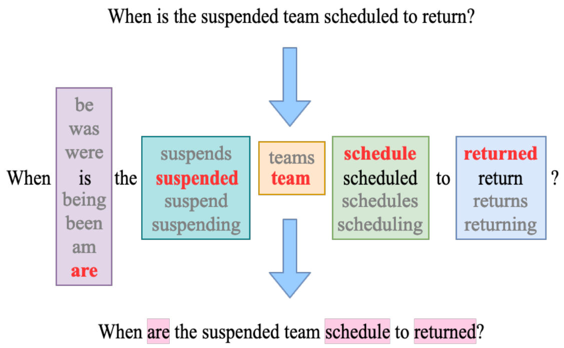


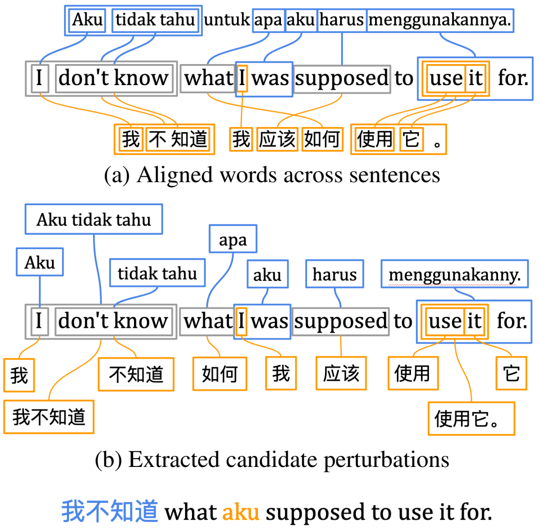
Figure 22: MORPHEUS looks at each noun, verb, or adjective in the sentence and selects the inflected form (marked in **red**) that maximizes the target model's loss. To maximize semantic preservation, it only considers inflections belonging to the same universal part of speech as the original word.

Since L2 (and many L1 dialect) speakers often exhibit variability in their production of inflectional morphology, we argue that NLP models should be robust to inflectional perturbations in order to minimize their chances of propagating linguistic discrimination. Particularly, in [110], we propose **Morpheus**, a method for generating plausible and semantically similar adversaries by perturbing the inflections in the clean examples (fig. 22). We demonstrate its effectiveness on multiple machine comprehension and translation models, including BERT and Transformer. We also show that adversarially fine-tuning the model on an adversarial training set generated via weighted random sampling is sufficient for it to acquire significant robustness, while preserving performance on clean examples.

Many extant NLP systems use a combination of a whitespace and punctuation tokenizer followed by a data-driven subword tokenizer such as byte pair encoding (BPE; [104]). However, a purely data-driven approach may fail to find the optimal encoding, both in terms of vocabulary efficiency and L2/dialectal generalization. This could make the neural NLP systems more vulnerable to inflectional perturbations. In our follow-up work [111], we propose Base-InflecTion Encoding or **BITE**, which uses morphological information to help the data-driven tokenizer use its vocabulary efficiently and generate robust token sequences. In contrast to morphological segmentors, we reduce inflected forms to their base forms before reinjecting the inflection information into the encoded sequence as special symbols. This approach gracefully handles the canonicalization of words with nonconcatenative morphology while generally allowing the original sentence to be reconstructed. We demonstrate its effectiveness at making NLP systems robust to non-standard inflection use while preserving performance on Standard English examples. Crucially, simply fine-tuning the pretrained model for the downstream task after adding BITE is sufficient. Unlike adversarial training, BITE does not enlarge the dataset and is more computationally efficient. We also show that BITE helps BERT generalize to dialects unseen during training and also helps Transformer-big converge faster for MT tasks.

(b) Robustness to Code-mixing As mentioned (§2), the massive multilingual models have demonstrated impressive cross-lingual transfer abilities: simply fine-tuning them on task data from a high resource language such as English after pretraining on monolingual corpora was sufficient to manifest such abilities. However, transferring from one language to another is insufficient for NLP systems to understand multilingual speakers in an increasingly multilingual world. In many multilingual societies, it is common for multilingual interlocutors to produce sentences by mixing words, phrases, and even grammatical structures from the languages in their repertoires, a phenomenon known as *code-mixing*. Hence, it is crucial for NLP systems serving multilingual communities to be robust to code-mixing if they are to understand and establish rapport with their users or defend against adversarial polyglots. Although gold standard data is important for definitively evaluating code-mixed text processing ability, such datasets are expensive to collect and annotate.

In our recent work [111], we posit that performance on appropriately crafted adversaries could act as a lower bound of a model's ability to generalize to the distribution simulated by said adversaries. We propose two strong black-box adversarial attacks targeting the cross-lingual generalization ability of massive multilingual representations (fig. 23), demonstrating their effectiveness on SoTA models for NLI and QA. We also propose an efficient adversarial training scheme that takes the same number of steps as standard supervised training and show that it creates more language-invariant representations, improving accuracy in the absence of lexical overlap.



(a) Aligned words across sentences

(b) Extracted candidate perturbations

(c) Final multilingual adversary

Figure 23: Bumblebee's three key stages of adversary generation: (a) Align words in the matrix (English) and embedded sentences (top: **Indonesian**, bottom: **Chinese**); (b) Extract candidate perturbations from embedded sentences; (c) Construct final adversary by maximizing the target model's loss.

(c) Reliability Testing Rigorous testing is critical to ensuring an NLP system works as intended (functionality) when used under real-world conditions (reliability). A lack of rigorous testing, coupled with machine learning's (ML) implicit assumption of identical training and testing distributions, may inadvertently result in systems that are harmful and discriminate against minorities, who are often underrepresented in the training data. This can take the form of misrepresentation of or poorer performance for people with disabilities, specific gender, ethnic, age, or linguistic groups. Examples include GPT-3 agreeing with suggested suicide [97], the mistranslation of an innocuous social media post resulting in a minority's arrest [30], and biased grading algorithms that can negatively impact a minority student's future [18]. Many of such potential harms can be mitigated by detecting them early and preventing the offending model from being put into production. Hence, in addition to being mindful of the biases in the NLP pipeline and holding creators accountable via audits, in our position paper [109], we argue for the need to evaluate an NLP system's reliability in diverse operating conditions. We reformulate adversarial attacks as *dimension-specific*, worst-case tests that can be used to approximate real-world variation. We contribute a reliability testing framework — **DOCTOR** — that translates safety and fairness concerns of NLP systems into quantitative tests. We demonstrate how testing dimensions for

DOCTOR can be drafted for a specific use case. Finally, we discuss the policy implications, challenges, and directions for future research for reliability testing.

(d) Robustness of NMT to Natural Noise It has been shown that NMT models are quite brittle against small input perturbations. Real noises can exhibit in many forms such as spelling and grammatical errors, homophones replacement, new words or even a valid word used in a less or unfamiliar context. In the presence of such noise, not only can the word embeddings of perturbations cause irregularities with the local context, but the contextual representation of other words may also get affected, a process we call *noise propagation*. In [120], we propose a method called **Context-Enhanced Reconstruction (CER)** to minimize this noise propagation and reduce the irregularities in contextual representation. To reduce the sensitivity of contextual words towards noisy words in the encoder, we inject made-up words randomly to the source side of the training data to break the text naturalness. We then use a Noise Adaptation Layer (NAL) to enable a more stable contextual representation by minimizing the reconstruction loss. In the decoder, we add perturbations with a semantic constraint and apply the same reconstruction loss. Unlike adversarial examples which are crafted to cause the target model to fail, our perturbation process does not have such a constraint and does not rely on a target model. Our input perturbations are randomly generated, representing any types of noises that can be observed in real-world usage. This makes the perturbation process generic, easy and fast. Experimental results on Chinese-English and French-English translation show significant improvements over the baselines for various domains.

5 Interdisciplinary Research

In addition to my own NLP research, I have been collaborating with researchers from other disciplines. These include computer vision (§5.1), speech (§5.4), social computing (§5.2), database and data mining (§5.3), and health (§5.5).

5.1 Multimodal (Image-Text) NLP

Vision and language are two of the most fundamental channels for humans to perceive the world and to act based on that. It has been a long-standing goal in AI to build machines that can jointly understand (and generate) vision and language data. Our work on multi-modal NLP spans learning general purpose cross-modal representations (§5.1.1) for visual-language tasks and improving the visual-language tasks with novel methods (§5.1.2).

5.1.1 Multimodal Representation Learning

Vision-and-language pre-training (VLP) has emerged to be an effective approach to learn general purpose vision-language representations. However, existing methods have several limitations. They lack the ability to model complex and fine-grained interactions between image and text. Most methods rely on a pre-trained object detector for image feature extraction, which is both annotation-expensive and computation-expensive. Finally, the datasets used for pre-training mostly consist of noisy image-text pairs collected from the Web. The widely used pre-training objectives such as masked language modeling (MLM) are prone to over-fitting to the noisy text.

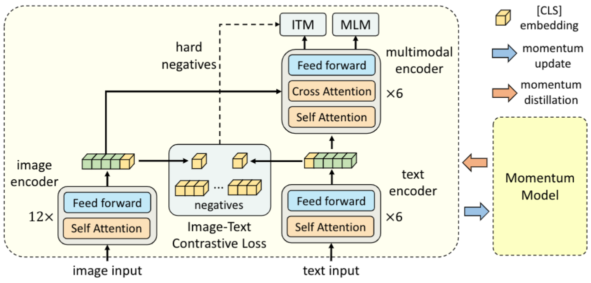
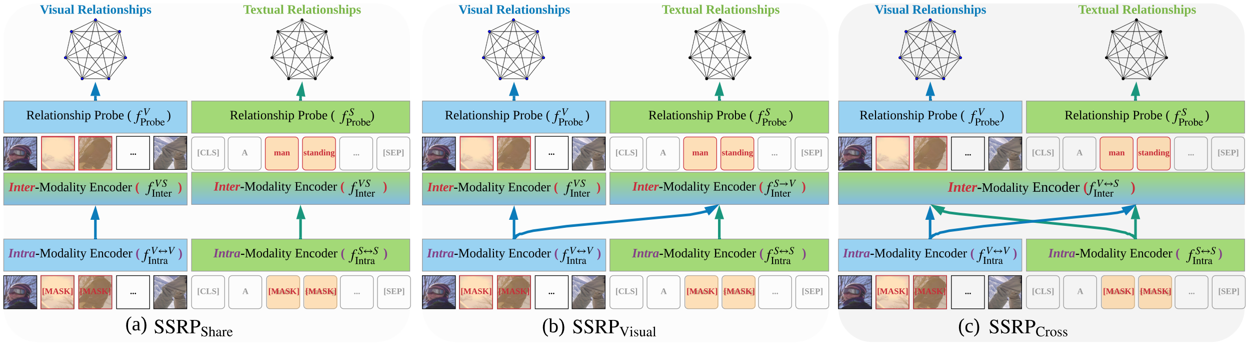


Figure 24: Framework of ALBEF.

In our NeurIPS-21 work [63], we propose ALign BEfore Fuse (ALBEF), a new vision-language representation learning framework to address the above limitations (fig. 24). ALBEF contains a transformer-based image encoder and a text encoder (first 6 layers of BERT), and a multimodal encoder (last 6 layers of BERT with additional cross-attention layers). We pre-train ALBEF by jointly optimizing three objectives: (*i*) an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion, (*ii*) an image-text pairwise matching loss at the multimodal encoder (using in-batch hard negatives mined through contrastive similarity), and (*iii*) an MLM loss at the multimodal encoder. To learn from noisy data, we propose momentum distillation, where we use a momentum model (a moving-average version of the base model) to generate pseudo-targets (additional supervision) for both image-text contrastive learning and masked language modeling. We also provide theoretical explanations from the perspective of mutual information maximization, showing that momentum distillation can be interpreted as generating views for each image-text pair. ALBEF achieves state-of-the-art performance on multiple vision-language downstream tasks such as image-text retrieval, visual question answering (VQA), and natural language visual reasoning (NLVR).

Self-supervised Visual Relationship Learning. Visual graph representations such as *scene graphs* that describe object relationships in images have become crucial for high-level computer vision tasks that need complex reasoning such as image captioning, image retrieval and visual reasoning. Despite great progress, current visual relationship models still rely on human-annotated relationship labels. Due to the combinatorics involved — two objects and one


Visual Relationships
Textual Relationships
Relationship Probe (f^V_{Probe})
Relationship Probe (f^S_{Probe})
[CLS]
A
man
standing
...
[SEP]
Inter-Modality Encoder (f^{VS}_{Inter})
Inter-Modality Encoder ($f^{S\rightarrow V}_{\text{Inter}}$)
Inter-Modality Encoder ($f^{V\leftrightarrow S}_{\text{Inter}}$)
Intra-Modality Encoder ($f^{V\leftrightarrow V}_{\text{Intra}}$)
Intra-Modality Encoder ($f^{S\leftrightarrow S}_{\text{Intra}}$)
[MASK]

(a) $\text{SSRP}_{\text{Share}}$ (b) $\text{SSRP}_{\text{Visual}}$ (c) $\text{SSRP}_{\text{Cross}}$

Figure 25: Overview of the three proposed types of SSRP frameworks, each of which consists of three types of modules: intra-modality encoder, inter-modality encoder and relationship probe.

relationship between them, where objects and relationships each have different types — relationships are numerous and have a long-tailed distribution, and thus, it is difficult to collect enough annotations to sufficiently represent important but less frequently observed relationships.

In our NeurIPS-20 paper [25], we propose a novel, self-supervised relationship probing (SSRP) method to discover relations between objects from the model's representation space (fig. 25). Our approach is based on two simple observations: (*i*) when we slightly change the images, the relative visual relations in those images remain unchanged; (*ii*) relations mentioned in image descriptions are visually observable in the corresponding image. Our approach relies on three modules, each consisting of a set of layers. In the first module, implicit intra-modal relationships are modeled using transformer encoders. In the second module, cross-modal learning allows for implicit relationship information to be leveraged across modalities. In the third module, relationships between visual and textual entities are represented explicitly as latent variables via a technique we call *relationship probe*. All modules are trained using self-supervision, with a first stage relying on masked LM to train the first two modules, and a second stage relying on contrastive learning and linguistic dependency trees as supervisory signals to train the relationship probe network. Our approach addresses issues with existing visual relationship models: it relies on self-supervision rather than explicit supervision, it explicitly models relationships as latent variables, and it leverages cross-modal learning but allows a single modality as input at prediction time. Our experiments demonstrate that our method can benefit both vision and vision-language tasks including Natural Language for Visual Reasoning (NLVR), Visual QA (VQA, GQA), and image captioning.

5.1.2 Visual-Language Tasks

(a) Cross-modal Retrieval. Cross-modal retrieval is the task to retrieve the images (*resp.* texts) that are relevant to a given textual (*resp.* image) query. The fundamental challenge in this task is to learn a common representation shared by data from different modalities. The common approach to learning such cross-modal embedding space is to first encode individual modalities into their respective features, and then map them into a common semantic space, which is often optimized via a ranking loss that encourages the similarity of the mapped features of ground-truth image-text pairs to be greater than that of any other negative pair. Although the feature representations in the learned common space have been successfully used to describe high-level semantic concepts of multi-modal data, they are not sufficient to retrieve images with detailed local similarity (*e.g.,* spatial layout) or sentences with word-level similarity.

In our CVPR-18 paper [22], we propose to incorporate generative models into textual-visual feature embedding for cross-modal retrieval. In particular, in addition to the conventional cross-modal feature embedding at the global semantic level, we also introduce an additional cross-modal feature embedding at the local level, which is *grounded* by two generative models: image-to-text and text-to-image. Figure 26 illustrates the concept of our proposed cross-modal feature embedding with generative models at high level, which includes three learning steps: *look*, *imagine*, and *match*. Given a query in image or text, we first *look* at the query to extract an *abstract* representation. Then, we *imagine* what the target item (text or image) in the other modality should look like, and get a more concrete *grounded* representation. We accomplish this by asking the representation of one modality (to be estimated) to generate the item in the other modality, and comparing the generated items with gold standards. After that, we *match*

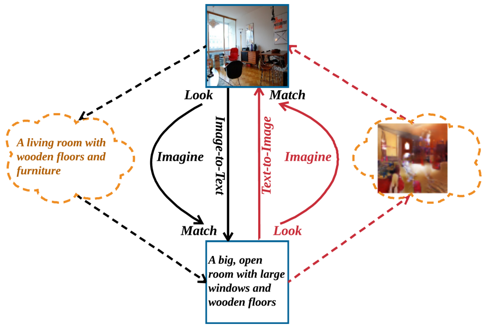

Look
Match
Image-to-Text
Text-to-Image
Imagine
Imagine
Match
Look
A living room with wooden floors and furniture
A big, open room with large windows and wooden floors

Figure 26: Conceptual illustration of our cross-modal feature embedding with generative models. The cross-modal retrievals (*Image-to-Text* & *Text-to-Image*) are shown in different colors. The two blue boxes are cross-modal data, and the generated data are shown in two dashed yellow clouds.

the right image-text pairs using the relevance score which is calculated based on a combination of *grounded* and *abstract* representations. We conduct extensive experimentations on the benchmark dataset, MSCOCO. Our empirical results demonstrate that the combination of the *grounded* and *abstract* representations can significantly improve the state-of-the-art performance on cross-modal image-caption retrieval.

(b) Image Captioning. Despite the impressive results achieved by deep learning in automatic image captioning, one performance bottleneck is the availability of large paired datasets because neural image captioning models are generally *annotation-hungry*, requiring a large amount of annotated image-caption pairs to achieve effective results. However, in many applications and languages, such large-scale annotations are not readily available, and are expensive and slow to acquire. In these scenarios, unsupervised methods that can generate captions from unpaired data or **semi-supervised** methods that can exploit paired annotations from other domains or languages are highly desirable. In our ECCV-18 paper [23], we pursue the latter research avenue, where we assume that we have access to image-caption paired instances in one language (Chinese), and our goal is to transfer this knowledge to a target language (English) for which we do not have such image-caption paired datasets. We also assume that we have access to a separate source-target (Chinese-English) parallel corpus to help us with the transformation. In other words, we wish to use the source language (Chinese) as a pivot language to bridge the gap between an input image and a caption in the target language (English).

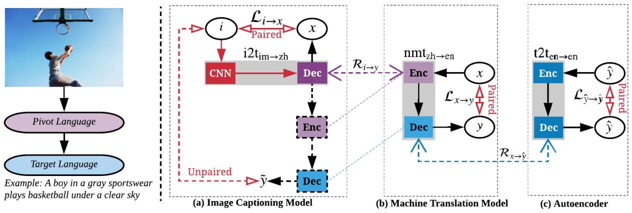

Pivot Language
Target Language
Example: A boy in a gray sportswear plays basketball under a clear sky
$\mathcal{L}_{i\to x}$
Paired
i
x
i2t$_{\text{im}\to\text{zh}}$
CNN
Dec
Enc
Dec
Unpaired
$\tilde{y}$
$\mathcal{R}_{i\to y}$
nmt$_{\text{zh}\to\text{en}}$
Enc
Dec
x
y
$\mathcal{L}_{x\to y}$
Paired
t2t$_{\text{en}\to\text{en}}$
Enc
Dec
$\hat{y}$
$\hat{y}$
$\mathcal{L}_{\hat{y}\to\hat{y}}$
Paired
$\mathcal{R}_{x\to\hat{y}}$
(a) Image Captioning Model
(b) Machine Translation Model
(c) Autoencoder

Figure 27: Illustration of our image captioning model with a pivot language. It first transforms an image into latent pivot sentences, from which our machine translation model generates the target caption.

The concept of using a pivot language (usually high-resource like English) as an intermediary language has been studied previously in machine translation (MT). Although related, image captioning with the help of a pivot language is fundamentally different from MT, since it involves putting together two different tasks – captioning and translation. In addition, the pivot-based *pipelined* approach to MT suffers from two major problems when it comes to image captioning. First, the conventional pivot-based MT methods assume that the datasets for source-to-pivot and pivot-to-target translations come from the same (or similar) domain(s) with similar styles and word distributions. However, as it comes to image captioning, captions in the pivot language (Chinese) and sentences in the (Chinese-English) parallel corpus are quite different in styles and word distributions. Second, the errors made in the source-to-pivot translation get propagated to the pivot-to-target translation module in the pipelined approach.

In [23], we present an approach that can effectively capture the characteristics of an image captioner from the source language and align it to the target language using another source-target parallel corpus (fig. 27). Specifically, our pivot-based image captioning framework comprises an image captioner *image-to-pivot*, an encoder-decoder model that learns to describe images in the pivot language, and a *pivot-to-target* translation model, another encoder-decoder model that translates the sentence in the pivot language to the target language, and these two models are trained on two separate datasets. We tackle the variations in writing styles and word distributions in the two datasets by adapting the language translation model to the captioning task. This is achieved by adapting both the encoder and the decoder of the pivot-to-target translation model. In particular, we regularize the word embeddings of the encoder (of the pivot language) and the decoder (of the target language) models to make them similar to image captions. We also introduce a joint training algorithm to connect the two models and enable them to interact with each other during training. The results show that our approach yields substantial gains over the baselines.

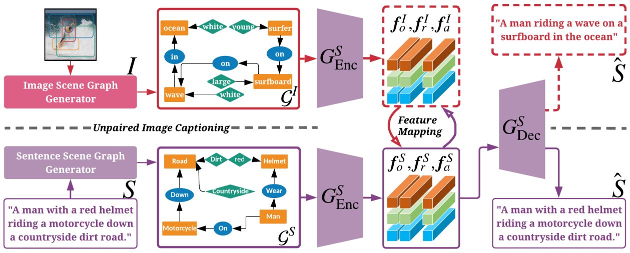


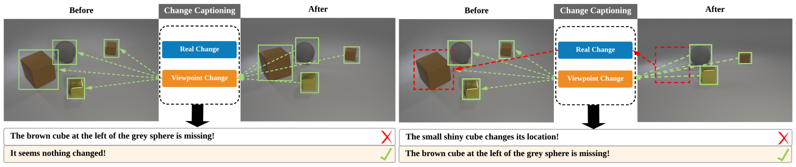
Image Scene Graph Generator
I
ocean
white
young
surfer
in
on
on
large
surfboard
wave
white
$\mathcal{G}^I$
G^S_{Enc}
f^I_o, f^I_r, f^I_a
"A man riding a wave on a surfboard in the ocean"
$\hat{S}$
Unpaired Image Captioning
Feature Mapping
G^S_{Dec}
Sentence Scene Graph Generator
S
"A man with a red helmet riding a motorcycle down a countryside dirt road."
Road
Dirt
red
Helmet
Countryside
Down
Wear
Man
Motorcycle
On
$\mathcal{G}^S$
G^S_{Enc}
f^S_o, f^S_r, f^S_a
$\hat{S}$
"A man with a red helmet riding a motorcycle down a countryside dirt road."

Figure 28: Illustration of our graph-based image captioning method. Our model consists of one visual scene graph detector (Top-Left), one fixed off-the-shelf scene graph language parser (Bottom-Left), a scene graph encoder G^S_{Enc}, a sentence decoder G^S_{Dec}, and a feature mapping module.

Inspired by the success of unsupervised MT, in our later (ICCV-19) work [24], we focus on **unsupervised** (or unpaired) image captioning. However, unlike unsupervised neural MT (§2.1.2) where the encoders can be shared across source and target languages, due to the different structures and characteristics of image and text modalities, the encoders of image and sentence cannot be shared to connect the two modalities. The critical challenge in unpaired image captioning is therefore the gap of information misalignment in images and sentences, so as to fit the encoder-decoder framework. To address this, we propose a *scene graph* based method that exploits the rich semantic information captured by scene graphs. Our framework comprises an image scene graph generator, a sentence scene graph generator, a scene graph encoder, a sentence decoder, and a feature alignment module that maps the features from image to sentence modality (fig. 28). We first extract the sentence scene graphs from the sentence corpus and train the scene graph encoder and the sentence decoder on the text modality. To align the scene graphs between images and sentences, we use CycleGAN to build the data correspondence between the two modalities. Specifically, given the unrelated image and sentence scene graphs, we first encode them with the scene graph encoder trained on the sentence corpus. Then, we perform unsupervised cross-modal mapping for feature level alignments with CycleGAN. By mapping the features, the encoded image scene graph is pushed close to the sentence modality, which is then used effectively as input to the sentence decoder to generate meaningful sentences. Our experimental results demonstrate the effectiveness of our proposed model in producing quite promising image captions. The comparison with recent unpaired image captioning methods validates the superiority of our method.

(c) Image Change Captioning. Inspired by the research development from dense object detection to image captioning, which generates natural language descriptions (instead of object labels) to describe the salient information in an image, the *change captioning task* has recently been proposed to depict the salient differences between images. Arguably, captions describing the changes are more accessible (hence preferred) for users, compared to the map-based labels (*e.g.,* pixel-level binary maps as change labels). Despite the progress, the existing change captioning methods cannot handle the viewpoint change properly. As shown in fig. 30, the viewpoint change in the images can overwhelm the actual object change leading to incorrect captions. Handling viewpoint changes is more challenging as it requires the model to be agnostic of the changes in viewpoints from different angles, while being sensitive to other salient changes.

In our ECCV-20 paper [107], following the prevailing architecture of a visual encoder plus a sentence decoder, we propose a novel viewpoint-agnostic image encoder, called Mirrored Viewpoint-Adapted Matching (M-VAM) encoder, for the change captioning task. Our main idea is to exhaustively measure the feature similarity across different regions in the two images so as to accurately predict the changed and unchanged regions in the feature space. The changed and unchanged regions are formulated as probability maps, which are used to synthesize the changed and unchanged features. We further propose a Reinforcement Attention Fine-tuning (RAF) process to allow the model to explore other caption choices by perturbing the probability maps. Our model outperforms the SoTA change captioning methods by a large margin in both Spot-the-Diff and CLEVR-Change datasets. Extensive experimental results also show that our method produces more robust prediction results for the cases with viewpoint changes.


Before
Change Captioning
After
Real Change
Viewpoint Change
The brown cube at the left of the grey sphere is missing! ✗
It seems nothing changed! ✓
Before
Change Captioning
After
Real Change
Viewpoint Change
The small shiny cube changes its location! ✗
The brown cube at the left of the grey sphere is missing! ✓

(a) Only viewpoint change happen

(b) Both object change and viewpoint change happen

Figure 29: Current change captioning methods can be influenced by the viewpoint change between image pairs, which results in generating wrong captions. To address this problem, we propose a novel framework that aims at explicitly separating viewpoint changes from real changes.

(d) Video Captioning. Different from image captioning, in the video captioning task, temporal information is more significant since the actions, event dynamics, and the surroundings in a video can hardly be fully represented by a single image. Prior methods mainly focus on encoding the temporal relationship in a video to predict a correct sentence, where the spatial feature of a frame is typically aggregated or pooled into one feature vector representing the frame; it is also computationally prohibitive to keep all the spatial features for all the frames. An RNN is used to compose the sequence of the frame features into a video representation, followed by another RNN as a language decoder to generate the words. Despite its success, the Seq2Seq based video captioning methods still suffer from some key limitations. First, existing methods assume that a video only narrates a single story and has few scenery changes, and thus they usually apply an RNN encoder to encode a video along the forward direction (from start to end), or the backward direction, or both. In the presence of noisy information at the beginning/end of a sequence, this approach will have a negative influence on encoding the key information that appears after/before the noise (fig. 30). This motivated us to come up with the novel idea: *how about watching a video twice: first predict the key frame and then encode the video based on that?* Another key problem is that the models disregard spatial features of the frames. Without the spatial information, we can hardly distinguish the location of the entities. Thus, to generate better captions, we need to incorporate spatial information while still limiting the dimensions of the visual features to a level that is computationally feasible.

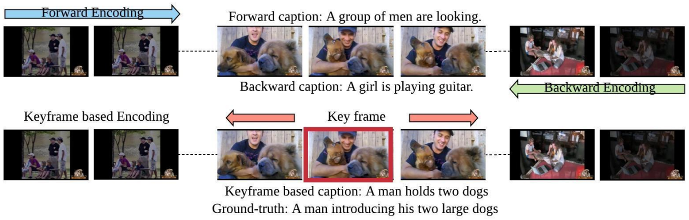
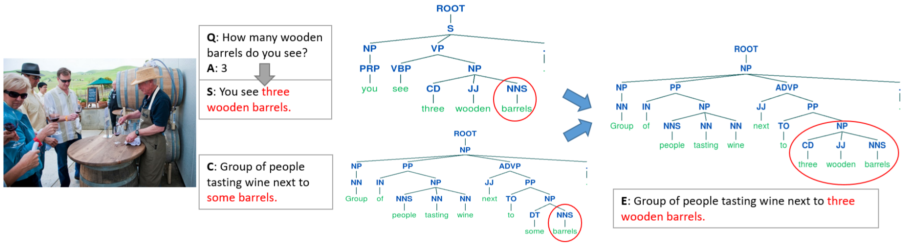

Forward Encoding
Forward caption: A group of men are looking.
Backward caption: A girl is playing guitar.
Backward Encoding
Keyframe based Encoding
Key frame
Keyframe based caption: A man holds two dogs
Ground-truth: A man introducing his two large dogs

Figure 30: Overall idea of the proposed video refocusing encoder. Top: video captioning with forward and backward video encoding. Bottom: the proposed key frame based video encoding.

In [106], we introduce a novel video captioning model, where we focus on the **visual encoding** part for generating a better video representation. We propose a simple spatial feature obtained by average pooling across different frame regions and integrate it into the captioning model. To tackle the issues with the noisy frames, we introduce a novel video refocusing encoder, where we encode each video twice. The first time is to select the key frame of a video by a key frame prediction network, and the second time is to re-encode the video centered at the key frame by two opposite directional RNN encoders. With no additional annotation, we further propose a novel reinforcement-learning based training method to jointly train the video refocusing encoder with the captioning model in an end-to-end manner. We test out the method on two widely used benchmark video captioning datasets, and we achieve results that rival SoTA methods and even outperform them in most cases.

In our follow-up work [105], we focus on the **language decoding** part of the Seq2Seq framework. We found that the prevailing RNN architectures, such as LSTMs, make mistakes by linking two words that do not appear together in the video but appear together frequently in the data. For example, "playing" appears often with "man" although it can also be found infrequently with some other entities, like "dog", "cat", and "baby" in other videos. In addition, when a word (*e.g.,* woman) appears in the data more frequently than another word (*e.g.,* motorcycle), the decoder tends to predict the former than the latter when both occur in a video, resulting in a caption like "a woman is riding a woman" instead of "a woman is riding a motorcycle". We propose a boundary-aware hierarchical language decoder for video captioning. It consists of a high-level GRU based language decoder, working as a global (caption-level) language model, and a low-level GRU based language decoder, working as a local (phrase-level) language model. The key novelty lies in the introduction of a binary gate into the low-level GRU language decoder, named Binary Gated Recurrent Unit (B-GRU), to detect phrasal boundaries according to language information and feed them back to the high-level language decoder to generate a global understanding of the currently generated sentence segments. To further improve the performance, we also incorporate another task of video prediction in a multi-task learning framework with a shared attention model for the two tasks.

(e) Visual Question Answering (VQA). Most existing methods perform VQA by utilizing the attention mechanism and combining the features from the two modalities for predicting answers. Although promising performance has been reported, there is still a huge gap for humans to truly understand the model decisions without any explanation for them. The visual justification through *attention visualization* is implicit and it cannot entirely reveal what the model captures from the attended regions for answering the questions. There could be many cases where the model attends to the right regions but predicts wrong answers. It has also been shown that attentions can be misleading. What's worse, the visual justification is not accessible to visually impaired people who are the potential users of the VQA techniques. Therefore, in our ECCV-18 paper [64], we explore textual explanations to compensate for these weaknesses of visual attention in VQA. Another crucial advantage of textual explanation is that it elaborates and enhances the predicted answer with more relevant information. Unfortunately, although textual explanations are desired for both model interpretation and effective communication in natural contexts, little progress has been made in this direction, partly because almost all the public datasets do not provide explanations for the annotated answers.



Q: How many wooden barrels do you see?
A: 3

S: You see three wooden barrels.

C: Group of people tasting wine next to some barrels.

E: Group of people tasting wine next to three wooden barrels.

Figure 31: An example of the pipeline to fuse the question (Q), the answer (A) and the relevant caption (C) into an explanation (E). Each QA pair is converted into a statement (S). The statement and the most relevant caption are both parsed into parse trees, which are then aligned by the common node. The subtree including the common node in the statement is merged into the caption tree to obtain the explanation.

We address the above limitations of existing VQA systems by introducing VQA-E (VQA with Explanations), where the models are required to provide a textual explanation for the predicted answer. We conduct our research in two steps. First, to foster research in this area, we construct a new dataset with textual explanations for the answers. The VQA-E dataset is automatically derived from the popular VQA v2 dataset by synthesizing an explanation for each image-question-answer triple. The VQA v2 dataset is one of the largest VQA datasets with over 650k question-answer pairs, and more importantly, each image in the dataset is coupled with five descriptions from MSCOCO captions. Although these captions were written without considering the questions, they do include some QA-related information and thus exploiting these captions could be a good initial point for obtaining explanations free of cost. We further explore several simple but effective techniques to synthesize an explanation from the caption and the associated question-answer pair (fig. 31). To relieve concern about the quality of the synthesized explanations, we conduct a comprehensive user study to evaluate a randomly selected subset of the explanations. The user study results show that the explanation quality is good for most question-answer pairs while being a little inadequate for the questions asking for a subjective response or requiring common sense. Overall, we believe the newly created dataset is good enough to serve as a benchmark for the proposed VQA-E task.

To show the advantages of learning with textual explanations, we also propose a novel VQA-E model, which addresses both the answer prediction and the explanation generation in a multi-task learning architecture. Our dataset enables us to train and evaluate the VQA-E model, which goes beyond a short answer by producing a textual explanation to justify and elaborate on it. Through extensive experiments, we find that the additional supervisions from explanations can help the model better localize the important image regions and lead to an improvement in the accuracy of answer prediction. Our VQA-E model outperforms the state-of-the-art methods in the VQA v2 dataset.

5.2 NLP for Social Media

I have worked on tweet classification problems to support crisis situations (§5.2.1), and advised a visiting student for her work on fact (or claim) verification (§5.2.2).

5.2.1 Crisis Computing

During the onset of a crisis situation (*e.g.,* earthquake, flood), rapid analysis of messages posted on microblogging platforms such as Twitter can help humanitarian organizations gain situational awareness, learn about urgent needs, and to direct their decision-making processes accordingly. However, time-critical analysis of such big crisis data brings challenges to machine learning techniques, especially to supervised learning methods. The scarcity of labeled data, particularly in the early hours of a crisis, delays the learning process. Traditional approaches use *batch learning* with hand engineered features like cue words and TF-IDF vectors. This approach has three major limitations. First, in the beginning of a disaster situation, there is no labeled data available for training for that particular event. Later, the labeled data arrives in minibatches depending on the availability of volunteers. Due to the discrete word representations and the variety across events, traditional classification models perform poorly when trained on previous (out-of-domain) events. Second, training a classifier from scratch every time a new minibatch arrives is infeasible. Third, extracting the right features for each disaster related classification task is time consuming and requires domain knowledge.

Deep neural networks (DNNs) are ideally suited for disaster response with big crisis data. They are usually trained with *online learning* (*e.g.,* Stochastic Gradient Descent or SGD) and have the flexibility to adaptively learn from new batches of labeled data without requiring to retrain from scratch. Due to their distributed word representation, they generalize well and make better use of the previously labeled data from other events to speed up the classification process in the beginning of a disaster. DNNs obviate the need for manually crafting features and automatically learn latent features as distributed dense vectors, which generalize well.

(a) Supervised Model. In [87, 86], we proposed convolutional neural networks (CNN) for the classification tasks in a disaster situation. CNN captures the most salient n-gram information by means of its convolution and max-pooling operations. On top of the typical CNN, we propose an extension that combines multilayer perceptron with a CNN. We present a series of experiments using different variations of the training data – event data only, out-of-event data only and a concatenation of both. Experiments are conducted for binary (*useful* vs. *not useful*) and multi-class (e.g., *donations*, *sympathy*, *casualties*) classification tasks. Empirical evaluation shows that our CNN models outperform non-neural models by a wide margin in both classification tasks in all scenarios. In the scenario of no event data, the CNN model shows substantial improvement of up to 10 absolute points over several non-neural models. Our variation of the CNN model with multilayer perceptron performed better than its CNN-only counter part. Another finding is that blindly adding out-of-event (a prior crisis event) data either drops the performance or does not give any noticeable improvement over the event only model. To reduce the negative effect of large out-of-event data and to make the most out of it, we apply two simple domain adaptation techniques – (*i*) weight the out-of-event labeled tweets based on their closeness to the event data, (*ii*) select a subset of the out-of-event labeled tweets that are correctly labeled by the event-based classifier. Our results show that the latter results in a better classification model.

(b) Semi-supervised Domain-adapted Model. Although obtaining a large amount of labeled data at the beginning of a crisis event (*e.g.,* Earthquake) is infeasible to train an effective DNN-based classifier, in most cases, we can have access to a good amount of labeled and abundant unlabeled data from past similar events (*e.g.,* Floods) and event-specific unlabeled data. In such situations, we need methods that can leverage the labeled and unlabeled data in a past event (we refer to this as a *source* domain), and that can adapt to a new event (we refer to this as a *target* domain) without requiring any labeled data in the new event. In other words, we need models that can do *domain adaptation* to deal with the distribution drift between the domains and *semi-supervised* learning to leverage the unlabeled data in both domains.

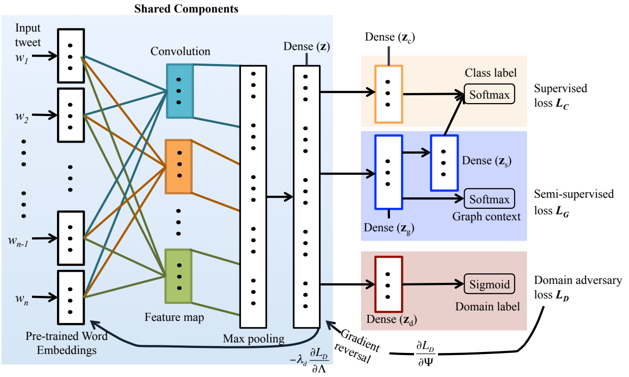


Figure 32: Architecture of the domain adversarial network with graph-based semi-supervised learning. The shared components part is shared by supervised, semi-supervised and domain classifier.

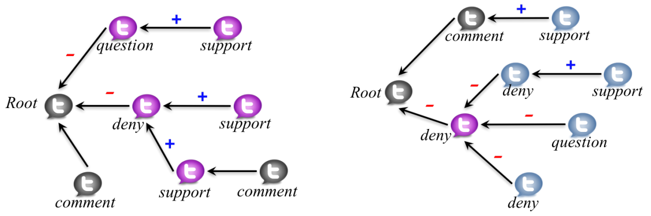
In our follow-up work [3, 4], we extend our method proposed in [86], proposing a novel model that performs domain adaptation and semi-supervised learning within a single unified deep learning framework (fig. 32). In this framework, the basic task-solving network (a CNN in our case) is put together with two other networks – one for semi-supervised learning and the other for domain adaptation. The semi-supervised component learns internal representations (features) by predicting contextual nodes in a graph that encodes *similarity* between labeled and unlabeled training instances. The domain adaptation is achieved by using a domain discriminator, which is a binary classifier that tries to decide whether the input example comes from the source or from the target domain. The training of this domain discriminator network is *adversarial* with respect to the shared layers by using gradient reversal during backpropagation, which makes the training *maximize* the loss of the discriminator rather than to minimize it. The overall idea is to learn high-level abstract representation that is discriminative for the main classification task, but is invariant across the domains. We propose an SGD algorithm to train the components of our model simultaneously. The effectiveness of the proposed approach is shown using two real-world Twitter datasets on scenarios where there is only unlabeled data in the target domain (or event).

5.2.2 Fact Checking (Rumor Detection)

The increasing popularity of social media has drastically changed how our daily news is produced, disseminated and consumed. The latest Pew Research statistics show that 70% of American adults at least occasionally get news on social media.[5] Without systematic moderation, a large volume of information based on false or unverified claims (*e.g.,* fake news, rumors, propaganda, etc.) can proliferate online. Such misinformation poses unprecedented challenges to information credibility, which traditionally relies on fact-checkers to manually assess whether specific claims are true or not. Despite the increased demand, the effectiveness and efficiency of human fact-checking is handicapped by the volume and fast pace of the noteworthy claims being produced on a daily basis. Therefore, it is an urgent need to automate the process and ease the human burden in assessing the veracity of claims.

Earlier approaches to automatic fact verification (or rumor detection) use recurrent neural networks (RNN) to capture the dynamic temporal characteristics of rumor diffusion. This method however oversimplifies the structural information associated with message propagation that can provide useful clues indicative of rumors. Propagation structures have been shown to be conducive to false rumor detection. In our work [75], we propose a tree-structured *recursive neural network* (RvNN) based on rumor propagation tree structures. The semantics of post content and the response relationships among the posts are jointly captured in the RvNN via the recursive feature learning process along the tree structure.

To illustrate our intuition, Figure 40 exemplifies the propagation trees of two rumors in our dataset, one being false and the other being true. Structure-insensitive methods typically relying on the relative ratio of different stances in the text cannot do well when such aggregated relativity is unclear as in this example. However, it can be seen that when a post denies a false rumor, it tends to spark supportive or affirmative replies confirming the denial; in contrast, denial to a true rumor tends to trigger questioning or denying utterances in the replies. This observation suggests a more general hypothesis that the repliers tend to disagree with (or question) those who support a false rumor or deny a true rumor, and agree with those who deny a false rumor or support a true rumor. Meanwhile, rather than directly responding to the source post (*i.e.,* the root post), a reply is usually responsive to its immediate ancestor, suggesting an obvious local characteristic of the interaction. The RvNN model naturally utilizes such structural properties for learning to capture rumor indicative signals and strengthen the representations by recursively aggregating the signals along different branches.



(a) False rumor

(b) True rumor

Figure 33: Propagation trees of two rumorous source tweets. Nodes may express stances on their parent as commenting, supporting, questioning or denying. The edge arrow indicates the direction from a response to its responded node, and the polarity is marked as '+' ('-') for support (denial). The same node color indicates the same stance on the veracity of the root node (*i.e.,* source tweet).

We propose two variants based on the standard RvNN, *i.e.,* a *bottom-up* (BU) model and a *top-down* (TD) model, which represent the propagation tree structure from two different angles, in order to recursively visit the nodes and combine their representations following distinct directions. With such basic architectures, the node features can be hierarchically refined by the recursion following the tree structure. Consequently, it can be expected that the discriminative features will be embedded into the learned representations more effectively. However, a potential issue of this model is that all responding posts are treated equally during the recursion, which may amplify the noise in the tree-structured representation learning. For example, the commenting posts in Figure 40 should have been less important due to their weak opinions towards the source claim. Previous studies have found that rumor detection can benefit from taking into

[5] https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

account the different stances expressed in responding posts. In this work, we introduce a novel method to improve our basic RvNN-based models by learning to automatically attend on those most evidential posts that express specific stances. Inspired by the success of neural attention, we propose specific attention mechanisms to encourage our model to be focused on such responsive posts in the tree during the bottom-up/top-down recursion.

We conduct extensive experiments on four real-world microblog datasets of different languages and demonstrate that 1) the proposed RvNN-based method yields outstanding improvements over the state-of-the-art baselines by a large margin; 2) the attention on most evidential posts over the propagation tree is effective; and 3) our method performs particularly well on early rumor detection which is crucial for timely intervention and debunking.

A recent trend in automatic fact verification is to establish more objective tasks and **evidence-based verification** solutions, which focus on the use of evidence obtained from more reliable sources, *e.g.,* encyclopedia articles, verified news, etc., as an important distinguishing factor. In the Fake News Challenge[6], the body text of an article is used as evidence to detect the stances relative to the claim made in the headline. The Fact Extraction and VERification (FEVER) task requires extracting evidence from Wikipedia and synthesizing information from multiple documents to verify the claim. Along the same line, in our work [74], we propose an approach to claim verification by using representation learning to embed sentence-level evidences based on coherence modeling and natural language inference (NLI).

c:	The test of a 5G cellular network is the cause of unexplained bird deaths occurring in a park in The Hague, Netherlands. *Verdict*: ***False***
s_1:	**[*Contradict*]**: Lots of tests going on with it in the Netherlands, but there haven't been test done in The Haque during the time that the mysterious starling deaths occurred.
s_2:	**[*Contradict*]**: One such test did occur in an area generally near Huijgenspark, but it took place on 28 June 2018.
s_3:	**[*Entail*]**: It's not clear whether tests with 5G have been carried out again, but so far everything points in the direction of 5G as the most probable cause.
s_4:	**[*Neutral*]**: Between Friday, 19 Oct and Saturday, 3 Nov 2018, 337 dead starlings and 2 dead common wood pigeons were found.
s_5:	**[*Entail*]**: The radiation created on the attempt of 5G cellular networks are not harmful only for birds but also for humans too.
s_6:	**[*Neutral*]**: 5G network developers promise faster data rates in addition to reduce energy and financial cost.
s_7:	**[*Neutral*]**: Parts of the park are blocked and dogs are no longer allowed to be let out, the dead birds are always cleaned up as quickly as possible.

Figure 34: Sentences topically coherent (s_1–s_4) and not coherent (s_5–s_7) with each other relative to the claim c, where their semantic entailment relations with c are shown.

The example in Figure 34 illustrates our general idea: given a claim and its relevant articles, we try to embed into the claim-specific representation those evidential sentences (*e.g.,* s_1–s_4) that are not only topically coherent among themselves considering the claim, but could also semantically infer the claim based on textual entailment relations such as *entail, contradict*, and *neutral*. We hypothesize that sentence-level evidence can convey more complete and deeper semantics, thus providing stronger NLI capacity between claim and evidence, which would result in better claim-specific representations for more accurate fact-checking decisions. To this end, we propose an end-to-end hierarchical attention network for sentence-level evidence embedding that aims to attend on important sentences (*i.e.,* evidence) by considering their topical coherence and semantic inference strength. Our model can determine the verdict of a claim more reasonably with evidential sentences embedded into the learned claim representation. Meanwhile, with the help of attention, crucial evidence can be highlighted and referred for better interpretability of the verdict. We use a co-attention mechanism to model sentence coherence and integrate the coherence- and entailment-based attentions into our proposed hierarchical attention framework for better evidence embedding. We experimentally confirm that our method is much more effective than several SoTA claim verification models using three public benchmark datasets.

5.3 NLP for Database & Data Mining

5.3.1 Deep Entity Resolution

Entity resolution (ER), a fundamental problem in data integration, has been extensively studied for 70+ years, from different aspects and in many domains such as health care, e-commerce, data warehouses, and many more. Despite the great efforts, there is still a long journey ahead in democratizing ER. Adding to the difficulty is the rapidly increasing size, number, and variety of sources of big data. A typical ER pipeline consists of four main steps: (*i*) labeling entity pairs as either matching or non-matching pairs; (*ii*) learning rules/ML models using the labeled data; (*iii*) blocking for reducing the number of comparisons; and (*iv*) applying the learned rules/ML models.

The major challenge of current solutions in democratizing ER is that each step needs human-in-the-loop. Even a "simple" step, such as step *i*, which is thought to be trivial, turned out to be difficult in practice. Moreover, the human resources required in each step might be different – knowing what (step *i*) is easier than telling why (step *ii*) or how (step *iii*). In practice, step *i* is tedious because humans can only label up to several hundred (or a few thousand) entity pairs

[6] http://www.fakenewschallenge.org/

and are error-prone. Intuitively, the hope to reduce this effort is to have a "prior knowledge" about what values would most likely match. Regardless of using rule- or ML-based methods, step *ii* requires experts to provide (domain-specific) similarity functions from a large pool (*e.g.,* SimMetrics). In addition, experts may also need to specify the thresholds. Ideally, this step needs a unified metric that can decide different cases of matched entities, from both syntactic and semantic perspectives. For step *iii*, a blocking function is typically defined over a few attributes, *e.g.,* country and gender in a table about demographic information, without a holistic view over all attributes or the semantics of the entities.

In [16], we present DeepER, a system for democratizing ER that needs much less labeled data by considering prior knowledge of matched values, captures both syntactic and semantic similarities without feature engineering, and provides an automated and customizable blocking method that takes a holistic view of all attributes – all of these targets are achieved by gracefully using distributed representations (DRs) of tuples. DRs of tuples is an extension of DRs of words (word embeddings). We present two methods for effectively computing DRs of tuples by composing the DRs of all the tokens within all attribute values of a tuple. The first method is a simple averaging of the tokens' DRs while the second uses uni- and bi-directional LSTM to convert each tuple into a DR. We introduce an end-to-end approach to tune the DRs that is customized for a specific ER task which improves the performance of DeepER (fig. 35). We propose two efficient and effective blocking algorithms based on the DRs for tuples and locality sensitive hashing, which takes the semantic relatedness of all attributes into account. DeepER shows superior performance compared to a SoTA ER solution as well to published methods on several benchmark datasets from citations, products, and proteomics. Finally, the proposed blocking delivers outstanding results under different conditions.

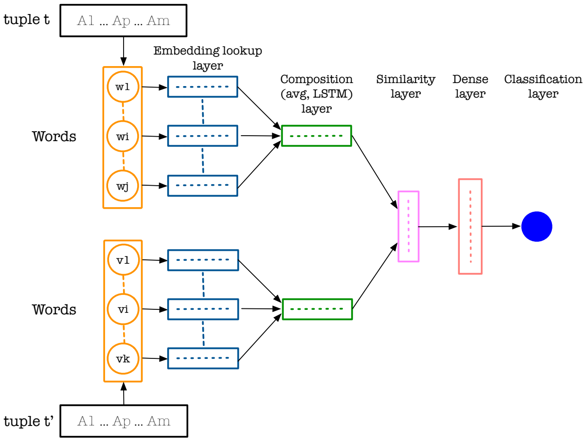
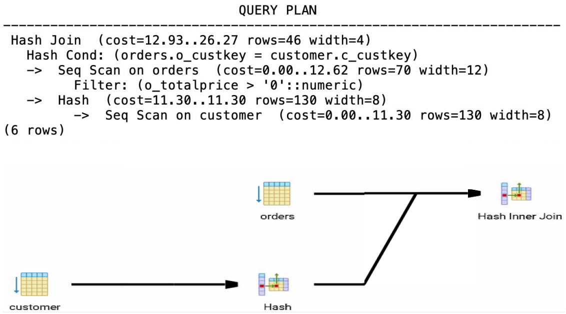


Figure 35: Deep Entity Resolution Framework.

5.3.2 NLG for Database Education

A key learning goal of learners taking a database course is to understand how SQL queries are processed in an RDBMS in practice. A relational query engine produces a query execution plan (QEP), which represents an execution strategy of an SQL query. Most commercial RDBMS expose the QEP of an SQL query using visual or textual format (fig. 36). Unfortunately, comprehending these textual formats to understand the query execution strategies of SQL queries in practice is daunting for learners. On the other hand, the visual format is relatively more user-friendly but hides important details. We advocate that an intuitive natural language (NL) description of a QEP can greatly facilitate learners to comprehend how an SQL query is executed by an RDBMS.

The majority of NL interfaces for RDBMS, however, have focused either on translating NL sentences to SQL queries or narrating SQL queries in an NL sentence. Scant attention has been paid for generating NL descriptions of QEPs. Natural language generation (NLG) for QEPs is challenging from several fronts. First, deep neural language generation methods that are very successful in NLP, rely on massive training sets of labeled examples. Such training sets in our context are prohibitively expensive to create as they demand database experts to translate thousands of QEPs of a wide variety of SQL queries. Second, ideally we would like to generate NL descriptions of QEPs using one application-specific dataset (*e.g.,* movies) and then use it for other applications (*e.g.,* hospital). That is, the NLG framework should be generalizable.



```
                               QUERY PLAN
----------------------------------------------------------------------
 Hash Join  (cost=12.93..26.27 rows=46 width=4)
   Hash Cond: (orders.o_custkey = customer.c_custkey)
   ->  Seq Scan on orders  (cost=0.00..12.62 rows=70 width=12)
         Filter: (o_totalprice > '0'::numeric)
   ->  Hash  (cost=11.30..11.30 rows=130 width=8)
         ->  Seq Scan on customer  (cost=0.00..11.30 rows=130 width=8)
(6 rows)
```

Figure 36: A QEP and its visual tree representation.

In [72, 115], we present a novel end-to-end system called LANTERN (naturaL lANguage descripTion of quERy plaNs) to generate NL descriptions of QEPs. Given an SQL query and its QEP, it automatically generates an NL description of the key steps undertaken by the underlying RDBMS to execute the query. To this end, instead of mapping an entire QEP to its NL description, we focus on mapping the set of physical operators in an RDBMS to corresponding NL descriptions and then stitch them together to generate the description of a specific QEP. Any RDBMS implements a small number of physical operators to execute any SQL query. Hence, although there can be numerous QEPs, they are all built from a small set of physical operators. Consequently, it is more manageable to label these operators and generate an NL description of any QEP from them. This also allows us to generalize LANTERN to handle any application-specific database as the relations, attributes, and predicates can simply be used as placeholders in describing a physical operator.

Lastly, it makes our framework orthogonal to the complexities of SQL queries as they are all executed by a small set of physical operators. We present a flexible declarative framework called POOL for succinctly specifying NL descriptions of physical operators in an RDBMS. We first develop a rule-based framework called RULE-LANTERN to generate an NL description of a QEP by leveraging the specified descriptions of physical operators. We observe from our engagements with learners that although rule-based approaches have high accuracy, it makes the descriptions of QEPs monotonous, leading to boredom. In fact, this is consistent with psychological theories that repetition of messages can lead to annoyance and boredom. To address this issue, we develop a novel deep learning-based language generation framework called NEURAL-LANTERN that infuses language variability in the generated description by exploiting a group of paraphrasing tools and pretrained language models (*e.g.,* BERT). Importantly, it addresses the challenge of training data generation by first generating a large number of random queries based on schema information and actual values in the database and then utilizing RULE-LANTERN and the paraphrasing tools to generate a large number of NL descriptions of the physical operators. We built LANTERN on top of PostgreSQL and SQL Server. Our exhaustive experimental study with real learners demonstrates the superiority of LANTERN compared to existing QEP formats of commercial RDBMS.

5.3.3 Aspect-based Neural Recommender

With the shift towards an increasingly digital lifestyle, recommender systems play a critical role in helping consumers to find the best product or service amongst a variety of options. Some of the most widely used recommendation systems rely on the Collaborative Filtering (**CF**) technique, which utilizes past interaction data such as ratings, purchase logs, or viewing history, to model user preferences and item features. However, a major limitation of CF techniques is its inability to provide reliable recommendations to users with few ratings, or recommend items with limited ratings, *i.e.,* the well-known *cold start* problem. Recent recommender systems have considered another valuable source of information which is readily available in many review websites: free-text reviews. More often than not, users provide an accompanying review to *explain* why they liked or disliked that particular product or service. For example, a review may include the user's opinions on the various *aspects* of an item, such as its price, performance, quality, etc. By focusing on these salient factors, we can better infer both the preferences of a specific user (*e.g.,* User *X* prefers a restaurant with outdoor seating) and the properties of an item (*e.g.,* Restaurant *Y* is famous for its seafood dishes).

Different users may emphasize more on different aspects throughout their interactions with these items. For example, some user may like a particular ***restaurant*** for its *food*, while another user frequents the same restaurant due to its cozy *ambiance*. Similarly, a user may prioritize the *storyline* when choosing a ***horror movie***, but pays more attention to the *cast* when evaluating an ***action movie***. Understandably, the importance of each aspect largely depends on both the user and item in question, and being able to capture such dynamic and fine-grained interactions between users and items would be invaluable in determining why some user may prefer an item over the other. In [121], we propose a novel neural recommender system which performs aspect-based representation learning for users and items by designing an attention mechanism to focus on the relevant parts of these reviews while learning the representation of aspects on the task. Furthermore, we estimate aspect-level user and item importance in a joint manner using the idea of co-attention, which allows us to model the finer-grained interactions between users and items (fig. 37). We conduct extensive experiments on 25 benchmark datasets from *Amazon* and *Yelp* to evaluate our proposed model against several SoTA baselines. We investigate how the different components in our proposed model contribute to its effectiveness.

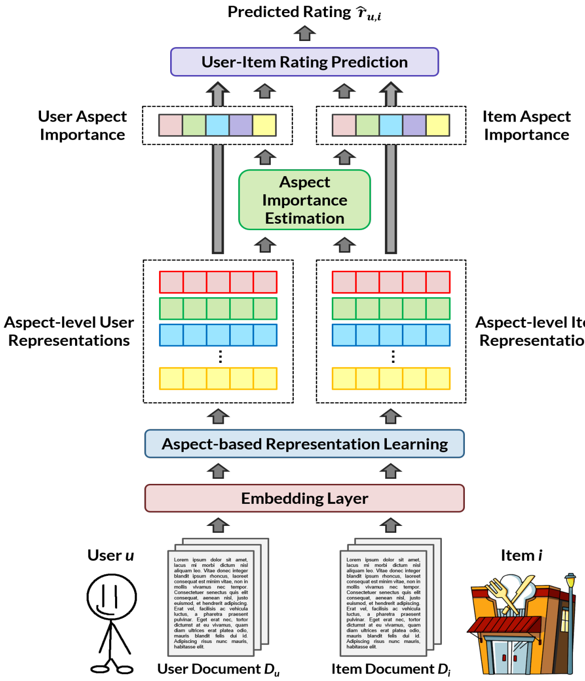

Predicted Rating $\hat{r}_{u,i}$
User-Item Rating Prediction
User Aspect Importance
Item Aspect Importance
Aspect Importance Estimation
Aspect-level User Representations
Aspect-level Item Representations
Aspect-based Representation Learning
Embedding Layer
User u
Item i
User Document D_u
Item Document D_i

Figure 37: Overall architecture of the proposed model

5.3.4 Sentence Representation for Text Mining

In earlier work, we focus on learning distributed representation of sentences for text mining tasks that involve clustering, classifying, or ranking sentences. Most prior work on learning sentence representations by and large considers only the content of a sentence and disregards the relations among sentences. In our work [99, 100], we propose a series of novel models for learning latent representations of sentences (i.e., Sen2Vec) that consider the content of a sentence as well as inter-sentence relations. We first represent the inter-sentence relations with a language network and then use the network to induce contextual information into the content-based Sen2Vec models. Two different approaches are introduced to exploit the information in the network. Our first approach *retrofits* (already trained) Sen2Vec vectors with respect to the network in two different ways: (*i*) using the adjacency relations of a node, and (*ii*) using a stochastic sampling method which is more flexible in sampling neighbors of a node. The second approach uses a regularizer to encode the information in the network into the existing Sen2Vec model. Experimental results show that our proposed models outperform existing methods in three fundamental tasks — *classification*, *clustering*, and *ranking*, demonstrating the effectiveness of our approach.

5.4 Speech Recognition

Our work has explored unsupervised modeling of speech (§5.4.1), online ASR (§5.4.2) and speech transformer (§5.4.3).

5.4.1 Unsupervised Speech Processing

Our interest in unsupervised speech processing stems from the desire to depart from expert based, fully supervised automatic speech recognition systems to the decipher-based scenario, where unlabeled speech and non-parallel text are available. In this scenario, a machine would have to learn to read and listen from scratch without correspondences between speech and text. Unsupervised representation learning can be seen as tackling the *listening* part of the larger problem. Another motivating factor for our work is unsupervised spoken language acquisition – the problem of discovering discrete linguistic structure from speech. The problem of acoustic unit discovery (AUD) falls under this category. The task is to cluster similar sounding acoustic segments, thereby discovering sound units that occur frequently in a speech corpus. A lower dimensional structured latent space can make the problem easier by reducing the number of parameters needed to build an AUD clustering model.

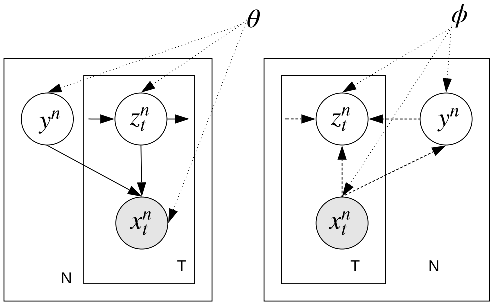


Figure 38: Proposed generative model on the left and the corresponding inference model on the right. $\boldsymbol{z_t}$ and $\boldsymbol{y}$ are latent random variables, $\boldsymbol{x_t}$ is the observed random variable, T is the number of frames in the acoustic segment, $\boldsymbol{x_{1:T}}$. Model and posterior parameters are θ and ϕ respectively. $\boldsymbol{y}$ and $\boldsymbol{z}$ encodes information present at different time scales in the speech signal.

In [57], we propose a novel generative model, the factorial deep markov model (FDMM) that learns disentangled and intepretable representations from speech without supervision. At a high level, the FDMM is just a variational auto-encoder (VAE) which, in addition to the usual encoder and decoder neural nets, has a transition neural net that models the *Markovian* dynamics in the latent space. The model is trained using Stochastic Variational Inference (SVI), an optimization-based approximate inference method. We evaluate our model on speaker verification, dialect identification and domain mismatched ASR tasks and show that it successfully encodes content and style/domain information in two independent latent variables.

5.4.2 Preventing Early Endpointing in Online ASR

With the development of end-to-end neural models for ASR, high attention is given to approaches in deploying these models for online speech recognition. Despite the recent progress, online ASR is known to have an early endpointing problem. There are mainly two categories of studies dealing with early endpointing. Voice-activity-detection scans input audio frames for long silence interval to stop decoding. However, using silence detection for ASR endpointing may not be ideal since they are essentially two different tasks. Furthermore, silence detection ignores acoustic cues or speaking rhythm. Another line of research is end-of-query detection. To train the ASR and endpointing jointly, a special '⟨/s⟩' token signaling the end of the utterance was incorporated into the label sequence for prediction. However, it requires performing a forced alignment between the transcript and speech beforehand to obtain the ground truth endpoint. In [127], we introduce a novel approach to address the early endpointing that neither relies on different types of silence nor obtains the ground truth endpoint with forced alignment. We leverage our ScaleGrad technique [66], which was originally proposed to mitigate the text degeneration issue (§3.1). We adapt it to discourage the early generation of '⟨/s⟩'. A scaling term is added to encourage the model to learn to keep generating non-'⟨/s⟩' tokens by directly maneuvering the gradient of the training loss. Our method is effective and can be jointly applied with other techniques discussed above. Experiments show that our model outperforms the baseline by a good margin.

5.4.3 Speech Transformers

Recently, the Transformer model [112] has been successfully introduced for ASR. As an end-to-end model, the Transformer not only combines the acoustic model, pronunciation dictionary and LM in a unified neural framework, it is also well known for its fast computation speed and ability to learn long range relationships. The encoder transforms audio signals into high-level representations, from which the decoder generates the text sequences in an auto-regressive manner one token at a time. In our work, we propose a series of improvements to the basic speech transformer model.

(b) Speaker Adaptation. Although ASR performance has greatly improved with the transformer, *speaker mismatch* between training and test data generally degrades the performance. Previous studies to address the speaker mismatch problem can be categorized into feature adaptation and model adaptation. Feature adaptation works on acoustic features, either by normalizing acoustic features to be speaker-independent, or by bringing auxiliary speaker related knowledge (*e.g., i*-vector) into the acoustic model. Model adaptation estimates the speaker-dependent parameters from speaker-independent model parameters with additional adaptation data.

In our work [125, 128], we propose we propose a unified speaker adaptation approach consisting of feature adaptation and model adaptation. For feature adaptation, we employ a speaker-aware persistent memory model which generalizes better to unseen test speakers by making use of speaker i-vectors to form a persistent memory. We concatenate speaker i-vectors to speech utterance, and apply this to each encoder layer, thus forming a persistent memory through the depth of the encoder (fig. 39). Different from prior work which learns for each speech time step, our method learns utterance level speaker knowledge. For model adaptation, we use a novel gradual pruning method to adapt to target speakers without changing the model architecture. We gradually prune less contributing parameters on model encoder to a certain sparsity level, and use the pruned parameters for speaker adaptation, while freezing the unpruned parameters to keep the original model performance. On the Librispeech dataset, our proposed approach brings 2.74-6.52% word error rate (WER) reduction (relative) on general speaker adaptation. On target speaker adaptation, our method outperforms the baseline by up to 20.1% and surpasses the finetuning based baseline by up to relative 8.62%.

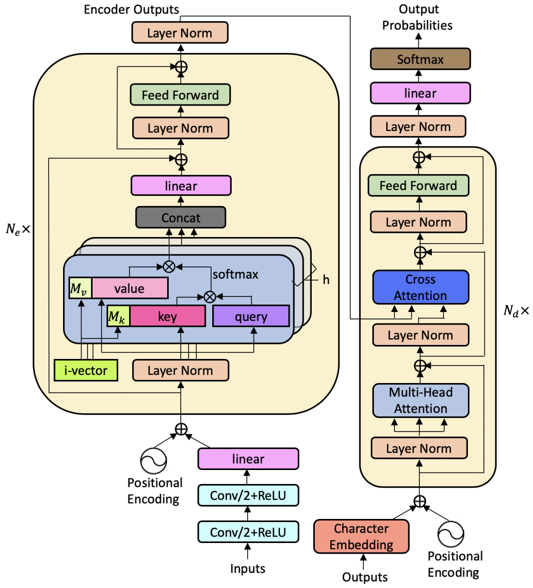


Figure 39: Speaker aware persistent memory. M_k and M_v from speaker i-vectors are concatenated to key and value.

(a) Improved Cross Attention with Monotonic Alignment. The default cross-attention module attends to the entire input utterance and obtains corresponding attention weights for decoding. However, when it comes to ASR, the same method may not work well, as monotonic alignment between text output and speech input is a characteristic of ASR, and has been studied using various techniques. In order to achieve better alignments between output and input for ASR, in [124], we propose a straightforward yet effective cross attention biasing technique for the Transformer model that takes output-input alignments into consideration without adding additional parameters to encoder hidden states. We take advantage of cross attention weights as a reference of output-input alignment to be used in current cross attention computation. We apply a Gaussian mask on attention weights centered at the alignment position. Additionally, we introduce a regularizer which regularizes alignment between output and input to encourage monotonicity.

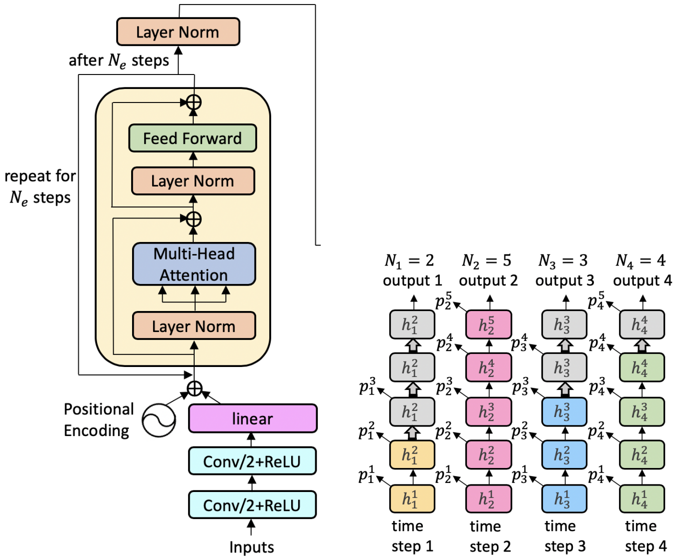


Figure 40: (Left) Encoder of the Universal Speech Transformer; Decoder has similar N_d repeated layers (not shown for brevity). (Right) Example of adaptive computation with 4 input time steps. Maximum number of layers L is 5 here. Illustration is applicable to both encoder and decoder.

Since lower layers of the Transformer capture more acoustic and local information, we apply our cross attention biasing on lower layers of the Transformer model, and leave the cross attention at higher layers to attend to entire speech input to capture global information. Our results on LibriSpeech 100h dataset show that our proposed model yields 14.5%-25.0% relative word error rate (WER) reductions.

(b) Universal Speech Transformer. The fixed numbers of encoder and decoder layers in the transformer model limit its computation capability. On one hand, compared with LSTMs, which have iterative or recursive computation, speech transformer loses the recurrent inductive bias, which is helpful to tackle tasks of varying complexity. Each input speech time step goes through the same and fixed numbers of encoder and decoder layers to compute the final output, regardless of the fact that different speech time steps differ in phoneme obscurity and noise level, and thus may require different computational resources. On the other hand, determining the number of encoder and decoder layers requires careful tuning for each dataset to achieve optimal performance.

In our work [126], we extend the basic idea of universal transformer [13] to the ASR task. It has a transformer-like architecture and uses a dynamic per-position halting mechanism to choose the required number of layers for each input time step dynamically, which exactly addresses the issues with the speech transformer analyzed above. To our knowledge, this is the first work regarding dynamic encoder and decoder depth in ASR. The recurrent nature of universal transformer best suits the needs of recognizing phonemes with different complexity and noise level, at the same time dynamically learning the encoder and decoder depth, which relieves the burden of tuning depth-related hyperparameters. However, the universal transformer model has two problems when applied on ASR. First, it adds the depth embedding and positional embedding repeatedly for each layer, which dilutes the acoustic information carried by hidden representations. Second, it performs a partial update of hidden vectors between layers, which is less efficient compared to the full update given the same number of updates. To tackle these two problems, we remove the depth embedding and only add the positional embedding once at the transformer encoder front-end, and we replace the partial update of hidden representations between layers with a full update. On three benchmark datasets, our model outperforms the baseline by 3.88% -13.7%, and achieves better results with much less computation cost compared to a very deep transformer model. From the experimental results, it can be seen that the number of encoder layers required varies among different input time steps and different datasets, which further substantiates the value of dynamic depth over fixed depth for datasets with varying complexity.

5.5 Deep Learning for Health

5.5.1 *activity2vec*: Representation Learning for Activity Time-Series

Physical activity and sleep are crucial to health and wellbeing. With the increasing popularity of wearable devices like *Fitbit*, which collect detailed data about the body's movements, there is an increased interest in using actigraphy for detecting sleep-related disorders and tracking longitudinal changes in the subject's condition. Although much lower in fidelity than clinical devices, the availability of wearables provides a novel opportunity, owing to its non-intrusive and real-time capabilities. However, only a minuscule proportion of the population has both their clinical and wearables data available. Hence, any approach towards using activity signals should utilize unsupervised learning. In addition, an important aspect is that information in actigraphy signals depends on the *subjects* and their *environments*, such as their routines and surroundings along with measurement errors owing to device design. In [1], we propose a new method *activity2vec* that addresses these challenges. Our method is an unsupervised representation learning model that learns *distributed* representations for activity signals spanning over a time segment (*e.g.,* at a day level) in a subject invariant manner. We use two public datasets to evaluate our approach against baselines on four disorder prediction tasks (Sleep Apnea, Diabetes, Insomnia, Hypertension). Using a linear classifier (logistic regression), we show that our proposed representation learning method outperforms the baseline time-series methods, with day-level representations performing the best. The linear classifier with our learned features performs at par with the convolution neural network baseline trained end-to-end on the tasks. We also demonstrate the effectiveness of inducing subject invariant features.

5.5.2 Sleep Stage & Quality Prediction

Outside of the wake state, sleep can be divided into three stages: *Rapid Eye Movement* (REM), Light sleep and Deep sleep.In [2], we make the first attempt to use CPAP-available flow signal to identify sleep stages automatically. CPAP users can know about their sleep health by learning about their sleep states, while health-care providers can track longitudinal sleep health and overall success of CPAP therapy (fig. 41).

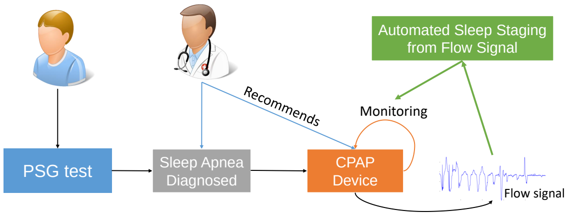


Figure 41: An application use case of our model. A patient undergoes Polysomnography (PSG) to ascertain the sleep disorders and is diagnosed with Sleep Apnea. Healthcare provider recommends CPAP therapy that involves a CPAP device. Flow signal can be obtained from the device daily for monitoring purposes. By adding the automated sleep staging step, we can help healthcare providers with the means for continuous monitoring of the patient.

We propose a new neural network architecture based on chain-structured CRF that explicitly models the temporal dynamics in the sleep states, over a CNN to learn high-level abstract features from CPAP flow signals and an *RNN* to encode temporal context in these features. The entire Neural CRF (CNN-RNN-CRF) network is trained for sleep staging in an end-to-end fashion. Our Neural CRF method shows a substantial improvement over the state-of-art when applied to the CPAP flow signal for sleep staging. Further, we improve the performance using a class distribution cost-sensitive prior to deal with the imbalanced distribution of sleep stages and using a domain dependent regularization over the CRF parameters.

In a separate study [101], we explore deep learning models for sleep quality prediction using actigraphy (wearables) data. In one setting, we first perform human activity recognition (HAR) on raw sensor data, and then feed HAR output into both conventional and deep learning models to perform sleep quality prediction. In the other setting, we employ several deep learning models directly on the raw wearable sensor data without performing HAR or any other feature extraction. Our results show that using a time-batched LSTM RNN on the raw wearables data improves the sleep quality prediction by an additional 10% with an overall AUC of 0.97 compared to the state-of-the-art non-deep learning approaches, which itself shows a 15% improvement over the current clinical practice. Moreover, utilizing deep learning on raw data eliminates the need for data pre-processing and simplifies the overall workflow to analyze actigraphy data for sleep and physical activity research. From an application impact perspective, the proposed approach promises a very high-fidelity screening test for sleep disorders directly from wearables data, potentially replacing the need for an inconvenient and expensive visit to a sleep laboratory for an evaluation.

5.5.3 NLP for ICU Management

Patients admitted into the intensive care unit (ICU) are monitored by different instruments on their bedside, which measure different vital signals about patient's health. During their stay, doctors visit the patient intermittently for check-ups and make *clinical notes* about the patient's health and physiological progress. These notes can be perceived as *summarized expert knowledge* about the patient's state. Predicting the condition of patients during their ICU stay can help plan better resource usage for patients that need it most in a cost-effective way. Prior studies have focused exclusively on modeling the problem using the time series signals from medical instruments. Expert knowledge from doctor's notes has been ignored in the literature.

In [56], we use clinical notes in addition to the time-series data for improved prediction on benchmark ICU management tasks (fig. 42). While the time-series data is measured continuously, the doctor notes are charted at intermittent times. This creates a new challenge to model continuous time series and discrete time note events jointly. We propose such a multimodal deep neural network that comprises of recurrent units for the time-series and convolution network for the clinical notes. We demonstrate that adding clinical notes improves the performance on in-hospital mortality prediction, modeling decompensation, and length of stay forecasting tasks.

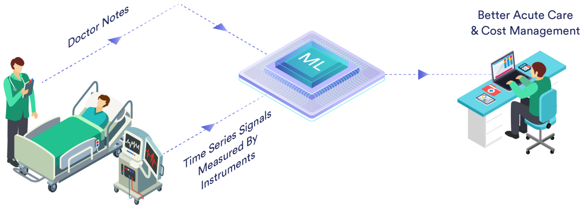


Figure 42: Doctor notes compliment physiological signals for better ICU management.

6 Ongoing & Future Work

In the future, I plan to explore the following research directions.

6.1 A Multilingual, Multitask Model that Learns Continually

I envision that future NLP models should be flexible enough to adapt to new languages (and dialects) and new tasks in each of these languages as they arrive, while preventing catastrophic forgetting (fig. 43). In a recent work [95] (under review), we propose a continual learning framework for learning new relations from texts with the help of a dynamic memory and data augmentation. In our ongoing work, we are exploring the emerging idea of *prompt tuning* [61] for continual few-shot learning, where we use a fixed large-scale generative language model for all tasks and append only a short sequence of task-specific tokens (*a.k.a.* prompts) to the input sequence. The parameters of the prompt tokens are tuned during training on the downstream tasks. We are exploring this idea for both language understanding and generation tasks.

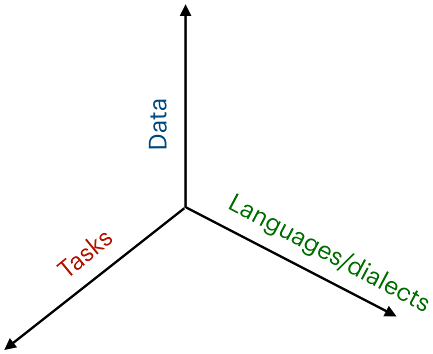


Figure 43: Multilingual, Multitask Continual Learning.

6.2 Improving Coherence, Factual Correctness and Reasoning in NLG

The SoTA neural text generation models suffer from three main limitations. First, the generated output lacks coherence as found in our recent study [82]. Second, the models tend to hallucinate, *i.e.,* they generate facts that are not faithful to the source and not consistent with the previously generated texts. Finally, they lack reasoning capabilities when the task requires logical or numerical reasoning. The latter two limitations are quite crucial for applications like data2text and summarization. I aim to address these problems of neural text generation models in two ways. First, by building reliable evaluation measures for these aspects, and then by proposing new methods to improve on these measures.

6.3 NLP for Programming

The rapid growth of code-related platforms such as Stack Overflow and GitHub has led to huge amounts of rich, open-source data containing programs associated with natural language. These include code with comments, questions and answers with code snippets, and communications between software developers. This opens up new opportunities (and challenges) for researchers to develop effective tools to assist programming.

Various research problems have emerged in this domain including pre-training of large-scale language models, NLP for code repair and fixing bugs, understanding natural language descriptions in code, code synthesis, automatic generation of code comments and documentation, etc. In our recent work [118], we propose CodeT5, a unified pre-trained encoder-decoder Transformer model that better leverages the code semantics conveyed from the developer-assigned identifiers. Our model is unified in that it builds on a unified framework to seamlessly support both code understanding and generation tasks, and it employs a unified format of task control codes to allow for multi-task learning. In future, I would like to focus on developing scalable methods that utilize code structures (*e.g.,* Abstract Syntax Tree), program-level (vs. function-level) code sequences, and source-code constraints for better understanding and generation of code.

6.4 Robust, Fair and Explainable NLP

Last but not the least, I would like to put significant effort into making the models robust, fair and explainable. Previously, we have explored robust training and encoding methods [110, 111] and debiasing methods like GeDi [59]. Our ongoing work explores the *causality* theory and its principles to address these issues.

Remark on Evaluating Computer Science Research

It is a common perception that the number of archival journals authored by a candidate is often considered as a major factor in promotion and tenure decisions in science and engineering fields. However, in computer science, especially in the areas of NLP, ML and AI, the preferred means of publication has been one of a number of selective annual conferences. Not only are these papers carefully peer-reviewed (typically by 3 to 5 referees), but the competition for acceptance is often overwhelming with extremely low acceptance rates (10-25%); so conference papers are highly valued in our field. The highly competitive peer-review process ensures a paper's novelty and quality. In fact, the ACL, EMNLP and NAACL conferences have much higher h5 Google Scholar citation index (157, 132, 105, respectively) compared to the journals in NLP (link). Likewise, the conferences ICLR, NeurIPS (formerly, NIPS), ICML, and AAAI are the top ranked venues in AI with h5-indices of 253, 245, 204 and 157, respectively (link).

References

[1] Karan Aggarwal, Shafiq Joty, Luis Fernandez-Luque, and Jaideep Srivastava. Adversarial unsupervised representation learning for activity time-series. In *Thirty-Third AAAI Conference on Artificial Intelligence*, AAAI'19, pages 834 – 841, Honolulu, Hawaii, September 2019. AAAI.

[2] Karan Aggarwal, Swaraj Khadanga, Shafiq Joty, Louis Kazaglis, and Jaideep Srivastava. A structured learning approach with neural conditional random fields for sleep staging. In *IEEE Big Data 2018*. IEEE, 2018.

[3] Firoj Alam, Shafiq Joty, and Muhammad Imran. Domain Adaptation with Adversarial Training and Graph Embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'18, pages 1077—-1087, Melbourne, Australia, 2018. Association for Computational Linguistics.

[4] Firoj Alam, Shafiq Joty, and Muhammad Imran. Graph based semi-supervised learning with convolutional neural networks to classify crisis related tweets. In *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM'18, pages 556 – 559, Stanford, California, June 2018. AAAI.

[5] M Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. UXLA: A Robust Unsupervised Data Augmentation Framework for Cross-Lingual NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, ACL'21, pages 1978–1992, Bangkok, Thailand, 2021. ACL.

[6] Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. Zero-Resource Cross-Lingual Named Entity Recognition. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'20, pages 7415–7423, New York, USA, September 2020. AAAI.

[7] Alberto Barrón-Cedeño, Simone Filice, Giovanni Da San Martino, Shafiq Joty, Lluís Màrquez, Preslav Nakov, and Alessandro Moschitti. Thread-level information for comment classification in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 687–693, Beijing, China, July 2015. Association for Computational Linguistics.

[8] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

[9] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 1171–1179. Curran Associates, Inc., 2015.

[10] Yllias Chali and Shafiq Joty. Improving the performance of the random walk model for answering complex questions. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, ACL'08, pages 9–12, Columbus, Ohio, 2008. ACL.

[11] Yllias Chali and Shafiq Joty. Selecting sentences for answering complex questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 304–313, Honolulu, Hawaii, 2008. ACL.

[12] Yllias Chali, Shafiq Joty, and Sadid Hasan. Complex question answering: Unsupervised learning approaches and experiments. *Journal of Artificial Intelligence Research*, 35(1):1–47, 2009.

[13] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.

[14] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL'14, pages 1370–1380, Baltimore, USA, 2014. ACL.

[15] Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. Daga: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, pages 6045—-6057, Punta Cana, Dominican Republic, 2020. ACL.

[16] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and Nan Tang. Distributed representations of tuples for entity resolution. In *The Forty-fourth International Conference on Very Large Data Bases*, volume 11 of *VLDB-2018*, pages 1454 – 1467, Rio de Janeiro, Brazil, August 2018.

[17] Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'21, Mexico City, Mexico, 2021. ACL.

[18] Todd Feathers. Flawed algorithms are grading millions of students' essays. *Vice*, 2019.

[19] Yifan Gao, Chien-Sheng Wu, Shafiq Joty, Caiming Xiong, Richard Socher, Irwin King, Michael Lyu, and Steven C.H. Hoi. Emt: Explicit memory tracker with coarse-to-fine reasoning for conversational machine reading. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL'20, pages 935—-945, Seattle, USA, 2020. ACL.

[20] Yifan Gao, Chien-Sheng Wu, Jingjing Li, Shafiq Joty, Steven C.H. Hoi, Caiming Xiong, Irwin King, and Michael Lyu. Discern: Discourse-aware entailment reasoning network for conversational machine reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, pages 2439—-2449, Punta Cana, Dominican Republic, 2020. ACL.

[21] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[22] Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *Computer Vision and Pattern Recognition*, CVPR'18, Spotlight, Salt Lake City, UTAH, USA, 2018. IEEE.

[23] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *European Conference on Computer Vision*, ECCV'18, pages xx–xx, Munich, Germany, 2018. Springer.

[24] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the International Conference on Computer Vision*, ICCV'19, pages 10323–10332, Seoul, Korea, 2019. IEEE.

[25] Jiuxiang Gu, Jason Kuen, Shafiq Joty, Jianfei Cai, Vlad Morariu, Handong Zhao, and Tong Sun. Self-Supervised Relationship Probing. In *2020 Conference on Neural Information Processing Systems*, NeurIPS'20, Vancouver, Canada, 2020.

[26] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 687–698, Baltimore, USA, June 2014. ACL.

[27] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing*, ACL'15, pages 805–814, Beijing, China, 2015. ACL.

[28] Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Machine translation evaluation with neural networks. *Computer Speech Language (Special Issue on Deep Learning for Machine Translation)*, 45:C:180–200, 2017.

[29] Simeng Han, Xiang Lin, and Shafiq Joty. Resurrecting Submodularity for Neural Text Generation. *arXiv (* not peer reviewed)*, 2020.

[30] Alex Hern. Facebook translates 'good morning' into 'attack them', leading to arrest. *The Guardian*, 2017.

[31] Enamul Hoque, Shafiq Joty, Lluís Màrquez, and Giuseppe Carenini. CQAVis: Visual Text Analytics for Community Question Answering. In *Proceedings of the 2017 international conference on Intelligent user interfaces*, IUI '17, page to appear, Limassol, Cyprus, 2017. ACM.

[32] Shafiq Joty. *Answer Extraction for Simple and Complex Questions*. PhD thesis, University of Lethbridge, December 2008.

[33] Shafiq Joty. *Discourse Analysis of Asynchronous Conversations*. PhD thesis, University of British Columbia, Vancouver, December 2013.

[34] Shafiq Joty, Alberto Barrón-Cedeño, Giovanni Da San Martino, Simone Filice, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Global thread-level inference for comment classification in community question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 573–578, Lisbon, Portugal, 2015. ACL.

[35] Shafiq Joty, Giuseppe Carenini, and Chin-Yew Lin. Unsupervised Modeling of Dialog Acts in Asynchronous Conversations. In *Proceedings of the twenty second International Joint Conference on Artificial Intelligence*, IJCAI'11, Barcelona, Spain, 2011.

[36] Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond T. Ng. Exploiting Conversation Structure in Unsupervised Topic Segmentation for Emails. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, EMNLP'10, pages 388–398, Massachusetts, USA, 2010. ACL.

[37] Shafiq Joty*, Giuseppe Carenini*, Raymond Ng, and Gabriel Murray. Discourse processing and its applications. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, ACL'19, pages 1–6, Florence, Italy, 2019.

[38] Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. A Novel Discriminative Framework for Sentence-Level Discourse Analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 904–915, Jeju Island, Korea, 2012. ACL.

[39] Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. Topic Segmentation and Labeling in Asynchronous Conversations. *Journal of Artificial Intelligence Research*, 47:521–573, 2013.

[40] Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41:3:385–435, 2015.

[41] Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 486–496, Sofia, Bulgaria, 2013. ACL.

[42] Shafiq Joty, Nadir Durrani, Hassan Sajjad, and Ahmed Abdelali. Domain Adaptation Using Neural Network Joint Model. *Computer Speech Language (Special Issue on Deep Learning for Machine Translation)*, 45:C:161–179, 2017.

[43] Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. Discourse Structure in Machine Translation Evaluation. *Computational Linguistics*, 43:4:683–714, Dec 2017.

[44] Shafiq Joty and Enamul Hoque. Speech Act Modeling of Written Asynchronous Conversations with Task-Specific Embeddings and Conditional Structured Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, pages 1746–1756. ACL, 2016.

[45] Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Joint Learning with Global Inference for Comment Classification in Community Question Answering. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'16, San Diego, California, 2016.

[46] Shafiq Joty, Lluís Màrquez, and Preslav Nakov. Joint multitask learning for community question answering using task-specific embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP'18, pages 4196 – 4207, Brussels, Belgium, 2018.

[47] Shafiq Joty and Tasnim Mohiuddin. Speech Act Modeling of Written Asynchronous Conversations: A Neural CRF Approach. *Computational Linguistics (Special Issue on Language in Social Media, Exploiting discourse and other contextual information)*, 44(4):859 – 894, 2018.

[48] Shafiq Joty*, Tasnim Mohiuddin*, and Dat Nguyen*. Coherence Modeling of Asynchronous Conversations: A Neural Entity Grid Approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'18, pages 558—-568, Melbourne, Australia, 2018. Association for Computational Linguistics.

[49] Shafiq Joty and Alessandro Moschitti. Discriminative reranking of discourse parses using tree kernels. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2049–2060, Doha, Qatar, October 2014. ACL.

[50] Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. Cross-language learning with adversarial neural networks: Application to community question answering. In *Proceedings of The SIGNLL Conference on Computational Natural Language Learning*, CoNLL'17, pages 226–237, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[51] Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP'15, pages 1259–1270, Lisbon, Portugal, 2015. ACL.

[52] Prathyusha Jwalapuram, Shafiq Joty, and Youlin Shen. Pronoun-targeted finetuning for nmt with hybrid losses. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, page 2267–2279, Punta Cana, Dominican Republic, 2020. ACL.

[53] Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP'19, page 2964–2975, Hong Kong, 2019. ACL.

[54] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual {bert}: An empirical study. In *International Conference on Learning Representations*, 2020.

[55] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[56] Swaraj Khadanga, Karan Aggarwal, Shafiq Joty, and Jaideep Srivastava. Using clinical notes with multimodal learning for icu management. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP'19, page 6432–6437, Hong Kong, 2019. ACL.

[57] Sameer Khurana, Shafiq Joty, Ahmed Ali, and James Glass. A fatorial deep markov model for unsupervised disentangled representation learning from speech. In *International Conference on Acoustics, Speech, and Signal Processing*, ICASSP'19, pages 6540 – 6544, Brighton, UK, September 2019. IEEE.

[58] Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics.

[59] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. Gedi: Generative discriminator guided sequence generation. *ACL ARR May*, 2021.

[60] Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In *EMNLP*, 2018.

[61] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *CoRR*, abs/2104.08691, 2021.

[62] Jing Li, Aixin Sun, and Shafiq Joty. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, IJCAI-ECAI-2018, pages 4166 – 4172, Stockholm, Sweden, July 2018.

[63] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *the 2021 Conference on Neural Information Processing Systems*, NeurIPS'21 (under review), Online, 2021.

[64] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *European Conference on Computer Vision*, ECCV'18, pages xx–xx, Munich, Germany, 2018. Springer.

[65] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[66] Xiang Lin and Shafiq Joty. Straight to the Gradient: Learning to Use Novel Tokens for Neural Text Generation. In *In Thirty-eighth International Conference on Machine Learning*, ICML'21, Virtual, 2021.

[67] Xiang Lin*, Shafiq Joty*, Prathyusha Jwalapuram, and Saiful Bari. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'19, page 4190–4200, Florence, Italy, 2019. ACL.

[68] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 997–1006, Portland, Oregon, 2011. Association for Computational Linguistics.

[69] Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, ACL'21, Bangkok, Thailand, 2021. ACL.

[70] Linlin Liu*, Xiang Lin*, Shafiq Joty, Simeng Han, and Lidong Bing. Hierarchical pointer net parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP'19, pages 1007—-1017, Hong Kong, 2019. ACL.

[71] Pengfei Liu, Shafiq Joty, and Helen Meng. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP'15, pages 1433–1443, Lisbon, Portugal, 2015.

[72] Siyuan Liu, Sourav S Bhowmick, Wanlu Zhang, Shu Wang, Wanyi Huang, and Shafiq Joty. Neuron: Query execution plan meets natural language processing for augmenting db education (demo). In *Proceedings of 45th ACM SIGMOD International Conference on Management of Data*, SIGMOD'19, page 1953–1956, Amsterdam, The Netherlands., July 2019. ACM.

[73] Annie Louis and Ani Nenkova. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1157–1168, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[74] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. Sentence-level evidence embedding for end-to-end claim verification with hierarchical attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'19, page 2561–2571, Florence, Italy, 2019. ACL.

[75] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. An Attention-based Rumor Detection Model with Tree-structured Recursive Neural Networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4:42):1–28, 2020.

[76] Giovanni Da San Martino, Salvatore Romeo, Alberto Barron-Cedeno, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Cross-language question re-ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'17, pages 1145–1148, Tokyo, Japan, September 2017. ACM.

[77] Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Shafiq Joty. Towards topic labeling with phrase entailment and aggregation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'13, pages 179–189, Atlanta, Georgia, June 2013. ACL.

[78] Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. LNMap: Departures from Isomorphic Assumption in Bilingual Lexicon Induction Through Non-Linear Mapping in Latent Space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, pages 2712—-2723, Punta Cana, Dominican Republic, 2020. ACL.

[79] Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. Augvic: Exploiting bitext vicinity for low-resource nmt. In *Proceedings of the Findings of 59th Annual Meeting of the Association for Computational Linguistics*, ACL'21 Findings, pages xx—-xx, Bangkok, Thailand, 2021. ACL.

[80] Tasnim Mohiuddin and Shafiq Joty. Revisiting adversarial autoencoder for unsupervised word translation with cycle consistency and improved training. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'19, page 3857–3867, Minneapolis, USA, 2019. ACL.

[81] Tasnim Mohiuddin and Shafiq Joty. Unsupervised Word Translation with Adversarial Autoencoder. *Computational Linguistics (presented at ACL-2020)*, 46(2):1 – 32, 2020.

[82] Tasnim Mohiuddin*, Prathyusha Jwalapuram*, Xiang Lin*, and Shafiq Joty*. Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks. In *Proceedings of the European Chapter of the ACL*, EACL'21, pages x—-x, Kyiv, 2021. ACL.

[83] Tasnim Mohiuddin*, Thanh-Tung Nguyen*, and Shafiq Joty*. Adaptation of hierarchical structured models for speech act recognition in asynchronous conversation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'19, page 1326–1336, Minneapolis, USA, 2019. ACL.

[84] Han-Cheol Moon*, Tasnim Mohiuddin*, Shafiq Joty*, and Chi Xu. A unified neural coherence model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP'19, page 2262–2272, Hong Kong, 2019. ACL.

[85] Dat Nguyen and Shafiq Joty. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL'17, pages 1320–1330, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[86] Dat Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Robust classification of crisis-related data on social networks using convolutional neural networks. In *Proceedings of the Eleventh International Conference on Web and Social Media*, ICWSM'17, pages 632–635, Montréal, Québec, Canada, May 2017. AAAI.

[87] Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. Applications of online deep learning for crisis response using social media information. In *4th international workshop on Social Web for Disaster Management*, volume abs/1610.01030 of *SWDM'16*, Indianapolis, USA, 2016.

[88] Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. Differentiable window for dynamic local attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL'20, pages 6589—-6599, Seattle, USA, 2020. ACL.

[89] Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. Efficient constituency parsing by pointing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL'20, pages 3284–3294, Seattle, USA, 2020. ACL.

[90] Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. A conditional splitting framework for efficient constituency parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, ACL'21, pages xx—-xx, Bangkok, Thailand, 2021. ACL.

[91] Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. Rst parsing from scratch. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL'21, Mexico City, Mexico, 2021. ACL.

[92] Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi, and Richard Socher. Tree-structured attention with hierarchical accumulation. In *International Conference on Learning Representations*, ICLR-20, 2020.

[93] Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. Data Diversification: An Elegant Strategy for Neural Machine Translation. In *2020 Conference on Neural Information Processing Systems*, NeurIPS'20, Vancouver, Canada, 2020.

[94] Xuan-Phi Nguyen, Shafiq Joty, Thanh-Tung Nguyen, Wu Kui, and Ai Ti Aw. Cross-model back-translated distillation for unsupervised machine translation. In *In Thirty-eighth International Conference on Machine Learning*, ICML'21, Virtual, 2021.

[95] Chengwei Qin and Shafiq Joty. Continual Few-shot Relation Learning via Data Augmentation and Embedding Space Regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP'21 (Under Review), page x–x, Punta Cana, Dominican Republic, 2021. ACL.

[96] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *Open-AI Blog*, 2019.

[97] Anne-Laure Rousseau, Clément Baudelaire, and Kevin Riera. Doctor GPT-3: hype or reality? *Nabla Technologies Blog*, 2020.

[98] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[99] Tanay Saha, Shafiq Joty, and Mohammad Hasan. Con-s2v: A generic framework for incorporating extra-sentential context into sen2vec. In *Proceedings of The European Conference on Machine Learning Principles and Practice of knowledge discovery in databases*, ECML-PKDD'17, Macedonia, Skopje, September 2017. Springer.

[100] Tanay Saha, Shafiq Joty, Naeemul Hassan, and Mohammad Hasan. Regularized and retrofitted models for learning sentence representation with context. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, CIKM'17, Singapore, November 2017. ACM.

[101] Aarti Sathyanarayana, Shafiq Joty, Luis Fernandez-Luque, Ferda Ofli, Jaideep Srivastava, Ahmed Elmagarmid, Shahrad Taheri, and Teresa Arora. Sleep Quality Prediction From Wearable Data Using Deep Learning. *JMIR mHealth and uHealth (JMU)*, 4(4)(e125), 2016.

[102] Rico Sennrich. Why the time is ripe for discourse in machine translation. 2018.

[103] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[104] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, 2016.

[105] Xiangxi Shi, Jianfei Cai, Jiuxiang Gu, and Shafiq Joty. Video Captioning with Boundary-Aware Hierarchical Language Decoding and Joint Video Prediction. *Neurocomputing*, pages 347–356, 2020.

[106] Xiangxi Shi, Jianfei Cai, Shafiq Joty, and Jiuxiang Gu. Watch it twice: Video captioning with a refocused video encoder. In *Proceedings of the 27th ACM International Conference on Multimedia*, ACMMM'19, page 818–826, Nice, France, 2019. ACM.

[107] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *European Conference on Computer Vision*, ECCV'20, Virtual, 2020.

[108] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics.

[109] Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A. Bennett, and Min-Yen Kan. Reliability testing for natural language processing systems: An adversarial perspective. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, ACL'21, page 4153–4169, Bangkok, Thailand, 2021. ACL.

[110] Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. It's morphin' time! combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL'20, pages 2920—-2935, Seattle, USA, 2020. ACL.

[111] Samson Tan, Shafiq Joty, Lav R. Varshney, and Min-Yen Kan. Mind Your Inflections! Improving NLP for Non-Standard English with Base-Inflection Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, pages 5647—-5663, Punta Cana, Dominican Republic, 2020. ACL.

[112] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[113] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc., 2015.

[114] Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4406–4417, Hong Kong, China, November 2019. Association for Computational Linguistics.

[115] Weiguo Wang, Sourav S Bhowmick, Hui Li, Shafiq Joty, and Siyuan Liu. Towards enhancing database education: Natural language generation meets query execution plans. In *Proceedings of 2021 ACM SIGMOD International Conference on Management of Data*, SIGMOD'21, pages x – x, Xi'an, Shaanxi, China, June 2021. ACM.

[116] Weishi Wang, Shafiq Joty, and Steven C.H. Hoi. Response selection for multi-party conversations with dynamic topic tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, pages 6581—-6591, Punta Cana, Dominican Republic, 2020. ACL.

[117] Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Multi-agent dual learning. In *International Conference on Learning Representations*, 2019.

[118] Yue Wang, Weishi Wang, Shafiq Joty, and Steven Hoi. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *The 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP'21 (under review), Online, 2021. ACL.

[119] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020.

[120] Weiwen Xu, AiTi Aw, Yang Ding, Kui Wu, and Shafiq Joty. Addressing the vulnerability of nmt in input perturbations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track)*, NAACL'21, Mexico City, Mexico, 2021. ACL.

[121] Chin Yao, Kaiqi Zhao, Shafiq Joty, and Gao Cong. Aspect-based neural recommender. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM'18, pages 147 – 156, Turin, Italy, October 2018. ACM.

[122] Tao Yu and Shafiq Joty. Online conversation disentanglement with pointer networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, pages 6321—-6330, Punta Cana, Dominican Republic, 2020. ACL.

[123] Tao Yu and Shafiq Joty. Effective fine-tuning methods for cross-lingual adaptation. In *the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP'21 (under review), Online, 2021. ACL.

[124] Yingzhu Zhao, Chongjia Ni, Cheung-Chi LEUNG, Shafiq Joty, Eng Siong Chng, and Bin Ma. Cross attention with monotonic alignment for speech transformer. In *21st Annual Conference of the International Speech Communication Association*, Interspeech'20, pages 5031 – 5035, Shanghai, China, October 2020. IEEE.

[125] Yingzhu Zhao, Chongjia Ni, Cheung-Chi LEUNG, Shafiq Joty, Eng Siong Chng, and Bin Ma. Speech transformer with speaker aware persistent memory. In *21st Annual Conference of the International Speech Communication Association*, Interspeech'20, pages 1261 – 1265, Shanghai, China, October 2020. IEEE.

[126] Yingzhu Zhao, Chongjia Ni, Cheung-Chi LEUNG, Shafiq Joty, Eng Siong Chng, and Bin Ma. Universal speech transformer. In *21st Annual Conference of the International Speech Communication Association*, Interspeech'20, pages 5021 – 5025, Shanghai, China, October 2020. IEEE.

[127] Yingzhu Zhao, Chongjia Ni, Cheung-Chi Leung, Shafiq Joty, Eng Siong Chng, and Bin Ma. Preventing early endpointing for online automatic speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, ICASSP'21, Brighton, UK, May 2021. IEEE.

[128] Yingzhu Zhao, Chongjia Ni, Cheung-Chi LEUNG, Shafiq Joty, Eng Siong Chng, and Bin Ma. A unified speaker adaptation approach for asr. In *the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP'21 (under review), Online, 2021. ACL.