Research Statement

Wei Lu
Singapore University of Technology and Design
25 February 2019

Abstract

In all my work, I followed a research approach that concentrates on the development of principled mathematical and algorithmic solutions to various fundamental NLP problems, with a specific focus on natural language semantics.

Specifically, I have made focused contributions to the following major areas: 1) entity-level semantics, where we worked out efficient models and algorithms based on several novel designs on hypergraphs that are capable of recognizing overlapping entities, 2) sentence-level semantics, where we proposed efficient algorithms for semantic parsing that can work robustly across different languages, and we also proposed models for language generation from semantic representations in the form of trees, forests and graphs, and 3) word-level semantics, where we proposed several principled mathematical approaches to learning word and graph representations, making use of network-level, node/word level and cross-domain information.

Background

Many of the problems within the field of natural language processing (NLP), such as chunking, tagging, word segmentation, named entity recognition, relation extraction, constituency parsing, dependency parsing, semantic parsing and targeted sentiment analysis, are essentially structured prediction problems. Such problems concern the prediction of a structured output from a given input. This observation motivates me to pursue a research direction on identifying a unified approach to solving a variety of NLP problems.

Over the years of research on designing models for different structured prediction problems, I have come to a conclusion that it is possible to design a unified approach for solving a wide range of structured prediction tasks. Specifically, after joining SUTD, in 2014, I figured out some connections between different structured prediction models, and developed and implemented a unified framework for structured prediction. This is a novel framework based on a specific type of directed hypergraphs that unifies various existing structured prediction models. Examples include classic models such as the linear-chain conditional random fields, the structural support vector machines, their latent variable, semi-Markov, and neural variants. It also encompasses tree structured prediction models such as parsing conditional random fields (parsing a sentence into its syntactic tree structure) based on weighted context-free-grammars, as well as the
latent variable variant that we proposed called hybrid trees [EMNLP’08] for semantic parsing (parsing a sentence into a tree-structured semantic representation). Various models for different applications developed in our research group are built based on the framework in the past few years [ACL’15, EMNLP’15a, NAACL’16, EMNLP’16c, AAAI’17a, AAAI’17b, EMNLP’17b, ACL’18a, ACL’18b, EMNLP’18a]. We introduced to the community this framework through a popular tutorial at EMNLP 2017 [EMNLP’17a], which attracted 95 registered participants. Currently, I am further developing a more general understanding on the connections between different structured prediction models by extending the current framework to encompass a wider range of structured prediction models (such as the transition-based structured prediction models).

While working on the unified structured prediction framework, I also focused on some concrete tasks within the field of NLP, with a specific focus on semantics. Some of the models involved in these tasks are implemented on top of the proposed structured prediction framework. We can broadly categorize the tasks into three main categories, focusing on semantic processing tasks at different levels within NLP.

In all my work, I followed a research approach that concentrates on the development of principled mathematical and algorithmic solutions to various fundamental NLP problems.

**Research Areas**

- **Named Entity / Mention Recognition** (*entity-level semantics*)

  We proposed a novel hypergraph representation for named entity / mention recognition [EMNLP’15a]. The model was built based on our proposed hypergraph-based structured prediction framework. To the best of our knowledge, it was the first hypergraph-based approach for this important information extraction task. One advantage of our approach is that it is able to capture overlapping entity mentions. While previous research efforts on handling nested named entities exist, we have demonstrated that our approach is effective while being able to offer a lower inference time complexity – it is linear in the number of words and linear in the number of semantic types (such as ORG, LOC, etc). We further proposed several novel models, including a model that is able to recognize discontinuous and overlapping entities [EMNLP’16c], a model based on a multigraph representation that labels gaps between words for extracting overlapping entities [EMNLP’17b], a neural transition-based approach for efficiently recognizing nested entities [EMNLP’18c], and another neural approach based on a novel segmental hypergraph representations for recognizing overlapping mentions [EMNLP’18d]. Currently our most recent model based on neural segmental hypergraphs achieves the state-of-the-art results on the benchmark datasets.
We have also done some other work related to entity recognition. Specifically, we introduced a neural approach to cross-domain named entity recognition where we used neural adaptation layers to adapt an existing named entity model trained on one domain for a new domain [EMNLP’18e]. We also introduced a model that makes use of dependency information within a semi-Markov conditional random fields framework to perform provably efficient named entity recognition [AAAI’17b].

- Semantic Parsing (sentence-level semantics)

Semantic parsing is the task of parsing a natural language sentence into its semantic representation. We introduced the hybrid tree semantic parser to the community in 2008 [EMNLP’08]. The parser assumes there exists a latent joint representation of words and semantics for a given sentence-semantics pair where efficient inference can be performed based on such a representation. The original parser was a generative model. We made some extensions to the parser to make it more effective. Specifically, we first introduced the discriminative version of the hybrid tree parser [EMNLP’14, ACL’15]. Next, we introduced the neural component to the hybrid tree model in 2017, arriving at the neural hybrid tree semantic parser [AAAI’17c]. In 2018, we introduced an alternative assumption on the latent representation for connecting words and semantics, and proposed a novel dependency-based hybrid tree model [EMNLP’18a]. The model is shown to be more robust and is particularly effective when dealing with certain languages such as Indonesian, and currently achieves the state-of-the-art results on benchmark datasets across a number of different languages.

We also additionally looked into the problem of parsing sentences into abstract meaning representations (AMR) using a transition-based model [EMNLP’18b]. Though it employs a transition-based approach to structured prediction, the parser learns to construct a derivation process which maps natural language words to basic semantic units. This process, which is inspired by the approach adopted by our hybrid tree semantic parsers, also yields a joint representation that contains both the words and semantics. Besides monolingual semantic parsing, we also looked into the task of multilingual semantic parsing under various different setups [COLING’14, ACL’17b, ACL’18a].

We also worked on the natural language generation task, which can be regarded as the inverse of the semantic parsing task. We first proposed a natural language generation model by inverting our hybrid tree semantic parser [EMNLP’09]. We next extended the hybrid tree model to support forest-structured semantic representations. It was then used by us to build a model which can generate natural language sentences from typed lambda calculus expressions [EMNLP’11]. The paper was also recognized as the only best paper at EMNLP 2011. Recently, we also worked on the problem of generating natural language sentences from graph-
structured semantic representations (such as AMR and dependency structures), and proposed a novel language generation model based on densely connected graph convolutional networks [TACL’19]. The model modifies the conventional graph convolutional neural network architecture used in the literature by introducing the novel dense connections, which was shown to be effective in alleviating the issue of training graph convolutional neural networks that involve many layers. The model yielded significant improvements over previous models on the language generation task while requiring significantly less model parameters.

- **Word / graph representation Learning (word-level semantics)**

  We proposed a few models for learning representations of words and graphs. We proposed the GraRep model [CIKM’15] for learning graph representations, which learns for each vertex in a weighted undirected graph a vector representation. Compared to existing approaches, our model is able to capture richer global structural information associated with the graph, and we also mathematically showed the connections between our approach and several previous classic approaches for learning graph and word representations. We then proposed another model based on neural networks for learning graph representations [AAAI’16]. While these models focus on learning representations based on graph information, we next looked into the problem of exploiting information within words (i.e., sub-word level information) for learning word representations. We designed a model that is able to exploit sub-word level information using convolutional neural networks for learning word representations for English [AAAI’17d], and also proposed another model exploiting sub-word (stroke) level information for learning Chinese word representations [AAAI’18b]. We further proposed models for learning word representations under a cross-domain setup [EMNLP’17c], based on our proposed regularization-based domain adaptation framework [EMNLP’16a].

Besides the above, we also made some focused contributions in a number of other areas. For example, we have some work related to sentiment analysis [AAAI’17a, AAAI’18a, ACL’18b]. Specifically, we worked on targeted sentiment analysis [AAAI’17a] where we need to jointly predict both the entities of interest and their sentiment from a given sentence. We also worked on aspect-level sentiment analysis [AAAI’18a] where a specific aspect target is provided and we need to predict its associated sentiment information. We also looked at a related task on negation scope extraction [ACL’18b], which aims to extract the negation scope (the part of the sentence being negated) with respect to a specific negation cue given in the text. We essentially regarded all these problems as prediction problems that involve latent structural information. For example, in the targeted sentiment analysis model, we assume there exists a latent text span that covers the entities, and it is the latent span that determines the sentiment associated with each entity. We therefore proposed and built a novel latent variable structured prediction model on top of our structured prediction framework, which
yielded the state-of-the-art results on the benchmark datasets. Similar ideas were used in our aspect-level sentiment analysis model and the negation scope extraction model, where we assume there exist some latent structural representations that is latent but crucial when making predictions.

We also worked on several interdisciplinary research topics. In one of them, we employed NLP techniques to perform textual inference to enhance the analogy-making experience in the design process [IDETC/CIE’17]. Together with researchers from the Singapore DSO national laboratories, we also created a large dataset consisting of cybersecurity related technical reports that are manually annotated with semantic information [ACL’17c] and organized a SemEval 2018 shared task based on the data [SemEval’18]. We also collaborated with colleagues from our EPD Pillar to employ the matrix product states methods introduced in the physics community to address a specific type of structured prediction problem – sequence-to-sequence learning where the input and outputs are linear-chain structures whose lengths are fixed [PR’18].

In the future, I plan to further extend our structured prediction framework so that it can encompass more structured prediction models. I also plan to apply the framework to more application domains within NLP and beyond.

**Selected Publications and Research Outputs**

[TACL’19] Zhijiang Guo, Yan Zhang, Zhiyang Teng and Wei Lu, “Densely Connected Graph Convolutional Networks for Graph-to-Sequence Learning”, Transactions of the Association for Computational Linguistics (accepted; decision type (a)).


