RESEARCH STATEMENT

Wei Lu

Abstract

My long-term research goal is to provide a unified approach to solving natural language processing problems. To do that, 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. **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,

2. **Entity-level semantics**, where we proposed efficient models and algorithms based on several novel designs on hypergraphs that are capable of recognizing complex (e.g., nested) entities, 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 the conclusion that it is possible to design a unified approach for solving a wide range of structured prediction tasks. Specifically, 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 [Lu et al., 2008, Lu, 2014] 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 [Lu, 2014, Lu and Roth, 2015, Muis and Lu, 2016b, Li and Lu, 2017, Jie et al., 2017, Susanto and Lu, 2017b, Zou and Lu, 2018, 2019, Li and Lu, 2019, Xu et al., 2020]. We introduced to the community this framework through a popular tutorial at EMNLP 2017 [Lu, 2017]. Currently, I am further developing a more general understanding of the connections between different structured prediction models by extending the current framework, so that it can encompass a wider range of 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 Contributions

Sentence-level Semantics

Semantic parsing is the task of parsing a natural language sentence into its semantic representation. In 2008, we introduced the *hybrid tree* semantic parser to the community [Lu et al., 2008]. 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 [Lu, 2014]. Next, we introduced the neural component to the hybrid tree model in 2017, arriving at the neural hybrid tree semantic parser [Susanto and Lu, 2017b]. 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 [Jie and Lu, 2018]. The model is shown to be more robust and is particularly effective when dealing with certain languages such as Indonesian, and currently achieves 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 [Guo and Lu, 2018]. Though it employs a transition-based approach to structured prediction, the parser learns to construct a derivation process that 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 by Susanto and Lu [2017a] and Zou and Lu [2018].

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 by Lu et al. [2009]. 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 by Lin and Lu [2011]. We also worked on the problem of generating natural language sentences from graph-structured semantic representations (such as AMR and dependency structures), and proposed novel language generation models based on densely connected graph convolutional networks by Cao et al. [2019b]. The approach 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 fewer model parameters.

In 2011, the best paper committee comprising leading researchers unanimously selected our EMNLP 2011 paper by Lu and Ng [2011] as the only best paper in that conference.

### Entity-level Semantics

We proposed a novel hypergraph representation for named entity or mention recognition by Lu and Roth [2015]. 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 by Muis and Lu [2016c], a neural transition-based approach for efficiently recognizing nested entities, and another neural approach based on a novel segmental hypergraph representation for recognizing overlapping mentions by Wang and Lu [2018]. Leading labs such as Cornell NLP group and CMU NLP group also investigated this task and made extensive comparisons with our models by Katiyar and Cardie [2018] and Shibuya and Hovy [2020]. Specifically, Cornell NLP group also used a hypergraph representation similar to ours with “exactly the same” expressiveness.

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 by Lin and Lu [2015]. We also investigated how to effectively make use of the dependency information for improved named entity recognition. For example, within a semi-Markov conditional random fields framework, we performed provably efficient named entity recognition by Jie et al. [2017], and investigated ways to inject the dependency information into the recognition process that involves an LSTM by Jie and Lu [2019]. We also proposed new models for better modeling of incomplete annotations for named entity recognition by Jie et al. [2019]. Recently we also started looking into other fundamental problems such as relation extraction by Guo et al. [2019a], Nan et al. [2020] and joint entity and relation extraction by Wang and Lu [2020].

Recently, together with Alibaba and other collaborators, we participated in an international competition on “Complex Named Entity Recognition”, and achieved the best performance on 10 out of 13 tasks. The resulting paper by Wang et al. [2023] also received the Best Paper Award at SemEval 2022.

### Word-level Semantics

We proposed a few models for learning representations of words and graphs (where the nodes of the graph can be words). We proposed the GraRep model by Cao et al. [2015], a principled general approach for learning graph representations. It 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 by Cao et al. [2016]. 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 by Cao and Lu [2017], and also proposed another model exploiting sub-word (stroke) level information for learning Chinese word representations by Cao et al. [2018]. We further proposed models for learning word representations under a cross-domain setup by Yang et al. [2017], based on our proposed regularization-based domain adaptation framework by Lu et al. [2016].

As of today, based on Google Scholar, our GraRep paper is ranked the 1st amongst all papers published in CIKM over the past 5 years in terms of citation count.

---


2 See this link: [https://semeval.github.io/SemEval2022/awards](https://semeval.github.io/SemEval2022/awards)

3 See this link: [https://scholar.google.com/citations?hl=en&vq=eng_databases/informationsystems&view_op=list_hcore&venue=V-1Ng2019PUBJ.2020](https://scholar.google.com/citations?hl=en&vq=eng_databases/informationsystems&view_op=list_hcore&venue=V-1Ng2019PUBJ.2020)
References


https://istd.sutd.edu.sg/people/faculty/lu-wei


https://istd.sutd.edu.sg/people/faculty/lu-wei