Transforming Natural Language Processing to Logical Forms

Nguyen Minh Le
Post Doctoral Fellow
Lab: Natural Language Processing Lab
Supervisor: Prof. Akira Shimazu

1. Purpose of the research

The purpose of our research is to focus on studying how natural language sentences could be transformed to its semantic representation like logical forms. Our main goal is to exploit machine learning on the corpus of sentences and their logical form to transform for a given new sentence. In addition, we also are interested in the task of how exploit structured prediction model to natural language processing applications. Our goal is also to construct a structured prediction model that allows both global and local context in its learning process. This machine learning framework then would be useful in exploiting it on the tasks of semantic parsing, information extraction, and machine translation. With the purpose of obtaining global context, we have chosen the topic modeling methods as the research direction to study.

2. Research Progress

In the COE project, I have conducted my research on transforming natural language (NL) sentences to its semantic representation such as logical form. Previously, I have built a structured prediction model for semantic parsing using Support Vector Machine. Although our model is successful for exploiting it on the corpus of the Robocup language, it still required a corpus of semantic annotation to a syntactic tree. This requirement therefore needs a human effort in creating such a corpus. In order to relax such constraints, we have designed a new structured prediction model which exploits only on the corpus of NL sentences and their semantic representation [1]. Our model is based on synchronous grammar models which have derived lexical rules using statistical machine translation techniques. We then construct a probabilistic synchronous grammar model using lexical rules and a context free semantic grammar of semantic representation outputs. Afterward, we applied online large-margin learning for estimating the parsing models. We have tested the model on both the Robocup and the Query Data based corpus. The results show significant improvement in comparison with previous works.

Beside that we have conducted our research on dependency parsing that exploits a new passive aggressive online large method for learning dependency parsing model. Our dependency parsing system has attended a dependency parsing shared task competition on the CONLL-2007 conference. The results showed that our system is better than the average result of those participating systems in CONLL-2007 shared task [2]. In addition, on the domain adaptation task our system attained the rank 4^{th}.

On the other hand, we are interested in the tasks of how structured prediction model could be able to apply to natural language processing applications. We have successfully formulated the conditional random field models for splitting clauses [3]. Our comparison with the state of the art works showed that the proposed system is more efficient than others in term of computational time, while the accuracy is comparable to the best results up to date.

In this time, we have contributed to the task of statistical machine translation by solving the problem of word ordering using a syntactic transformation model [4]. Our model showed that it significantly improves the phrase statistical machine models.

We also have conducted a research on modeling the topic of document using the large collection of text document. We applied latent semantic analysis to named entity recognition, which showed a significant improvement in comparison with named entity recognition not use topic modeling [5]. In addition, we also studied the uses Latent Direclet Allocation model for domain text classification. We mined the hidden topic model for a large collection available such as WikipeDiA for short and spare text classification. The results showed that topic modeling is able to use for improving such kinds of classification task [6].

With the purpose of using unlabeled data for improving the performance of learning based NLP application, we have designed a semi-supervised learning model for question classification [7] and a bootstrapping model for word sense disambiguation [8]. Our frameworks showed that using unlabeled data with ensemble learning model, the performance of either question classification or word-sense disambiguation are significant improved.

3. Future works

In our future work, we would like to study on how a semantic representation can be used to generate a natural language sentence. We would like to exploit synchronous grammar models for solving this generation task. We also would like to use a rich linguistic grammar such as Combinatory Categorical Grammar to represent the semantic output in order to enhance the performance of semantic parsing. On the other hand, the promising results of using topic modeling in text classification encourage us to exploit it on the domain of structured prediction such as semantic parsing tasks. In the future work, we would like to study how dependency parsing and semantic role labeling are able to joint. This aims at improving both the accuracy of dependency parsing and semantic role labeling. For this matter, we would like to incorporate semantic label into our current dependency parsing system.

Publications

1. **M.L. Nguyen and A. Shimazu**, "Online Large-Margin Structured Learning for Semantic Parsing", in submission.
2. **M.L. Nguyen**; **A. Shimazu; T.P. Nguyen; H.X. Phan**, *"A Multilingual Dependency Analysis System Using Online Passive-Aggressive Learning", Shared Task paper Proceedings of EMNLP-CONLL 2007, pp* 1149—1155
3. **V.V. Nguyen**, **M.L. Nguyen**, A. Shimazu, "Using Conditional Random Fields for Clause Splitting", 10th *Conference of the Pacific Association for Computational Linguistics (PACLING 2007)* pp. 58-65, 2007.
4. **P.T. Nguyen, A. Shimazu**, **M.L. Nguyen**, **V.V. Nguyen**, "A Syntactic Transformation Model for Statistical Machine Translation", *International Journal of Computer Processing of Oriental Languages* 20 (2), pp 1-21 (2007).
5. **H.X. Phan**, **S. Horiguchi**, **M.L. Nguyen**, **C.T. Nguyen**, "Semantic Analysis of Entity Context towards Open Named Entity Classification on the Web*", 10th Conference of the Pacific Association for Computational Linguistics (PACLING 2007)* pp. 137-144, 2007.
6. **H.X. Phan, S. Horiguchi, M.L. Nguyen, "**Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections", *to appear WWW 2008*.
7. **T.T. Nguyen**, **M.L. Nguyen**, **A. Shimazu**, "Using Semi-supervised Learning for Question Classification", *Journal Natural Language Processing* (to appear) 15(1) (2008)
8. **C. A. Le, A. Shimazu, V.N. Huynh**, **M.L. Nguyen**, "Semi-Supervised Learning Integrated with Classifier Combination for Word Sense Disambiguation", (to appear) *Computer Speech & Language* (2007)