Abstract

Text parsing is a core research topic in natural language processing (NLP). The output representation of text parsing is considered as vital seeds for NLP applications. Although text parsing has been studied comprehensively and intensively for decades, it is still a tough but attractive field. This research focuses on developing a novel parser for Abstract Meaning Representation (AMR). AMR is a powerful semantic representation at the sentence level. This representation features many aspects of semantic such as name entity, co-reference, and semantic relations. AMR represents the meaning of a text in form of a labeled root directed acyclic graph. Many applications based on AMR have been introduced with great improvement thank the robustness of AMR. Therefore, improving the accuracy of AMR parser would be a great contribution to the success of those applications including text generation, text extraction.

Abstract Meaning Representation parsing is a challenging task due to the complicated combination of many subtasks and the structure of AMR graph. AMR covers a wide range of semantic representation from co-reference, named entity and semantic relations. Hence, facing many tough subtasks in a single combined problem is the most arduous difficulty in AMR parsing. Additionally, the requirement of the graph-structure output is the second problem which restricts the utilization of many advanced deep learning model. Moreover, the existing AMR corpora are considered as small however it contains so large vocabulary that make the corpora sparse. The sparsity of the dataset is an addressed tough issue for any learning system.

Prior successful works utilized transition-based system which converts the dependency tree into AMR graph, however, this method is approaching the limit because of the simplicity of the model. Recent studies in AMR parsing turn to deep learning which has proved its robustness in many tasks. This thesis presents our study on developing an AMR parser using deep learning. First, we introduced a framework for text-to-AMR (AMR parsing) and AMR-to-text (AMR generation). Second, we combined convolutional sequence to sequence with our proposed linearization algorithm for AMR parsing. Third, we published an AMR dataset which we expect to encourage studies in the legal domain.

Firstly, we generalized AMR parsing and AMR generation tasks into a framework which would be an essential tool for addressing problem and evaluation in AMR-related works. This framework contains two components: a graph conversion algorithm and a neural machine translation model. This framework separates graph conversion and translation model so that we can easily assess the contribution of these factors. Secondly, we introduced an AMR parsing model based on our proposed framework. We proposed a

graph linearization algorithm with the reverse path as the core of graph conversion. Three sequence to sequence models is investigated in this research. They are bidirectional long short-term memory(LSTM) encoder-decoder, fully convolutional model and a combination of convolutional encoder and LSTM decoder. Finally, we put our effort to build an AMR test set for legal documents which we retrieve from the English version of Japan Civil Code.

Our evaluations were conducted with standard SMATCH score and various task-specific metrics which are widely used in other works. Our first experiment of upper bound proved that reverse path linearization method efficiently rewrites a graph as a sequence. Hence, we can totally confident to approach parsing task with neural machine translation method. Our parsers achieved state-of-the-art performance on two golden standard datasets LDC2014T12 and LDC2017T10 in the second evaluation. Our additional analysis indicates that graph linearization method dominates the performance of the parser. The benchmark of throughput confirms that convolutional sequence to sequence model delivers extremely higher throughput than LSTM-based models. In the evaluation of task-specific semantic parsing, the models with reverse path linearization outperformed existing linearization method with parenthesis.

Keywords: abstract meaning representation, convolutional sequence to sequence, graph linearization.