A Study on Abstract Meaning Representation

Information Science
NGUYEN Laboratory

Vu, Trong Sinh
1720003

1 Research Content

Humans are born with the ability to communicate with their natural language. Computing machines, on the other hand, only understand several specific programming languages, with a limit of expressions. To bridge the gap between humans and computers languages, semantic representation is one such a solution, with the ability to convert natural language utterances into machine-understandable forms. Many semantic schemes have been introduced and developed, such as Combinatory Categorial Grammar (CCG), Groningen Meaning Bank (GMB) or Abstract Meaning Representation. Two traditional problems of semantic representations are producing them from natural language (parsing) as well as producing natural language from them (generation). In this thesis, we present our study in Abstract Meaning Representation (AMR) Parsing and Generation, which are showing lots of potential in computational linguistics community recently. We also present our first attempt on the domain adaptation in parsing and generation for legal text.

In the first part of our thesis, we present our AMR-to-text generator incorporating the self-attention mechanism. Motivated by the domination of the Transformer architecture in various Natural Language Processing tasks, e.g. machine translation, text summarization, we adopt its core component - the self-attention - to build our generation models. We conduct experiments on both sequence to sequence and graph to sequence strategies, which are dominating in solving this problem when incorporating the self-attention mechanism. Our proposed method obtains competitive results on a benchmark AMR dataset, with an improvement of 3.2 BLEU score over the baseline sequence-to-sequence model. We also figure out current limitations of the self-attention mechanism when dealing with graph structure inputs.

Despite several developments in AMR parsing and generation for text in general domain, current methods for these tasks still struggle in dealing with the legal domain. The legal text is often structurally complicated, consists of longer sentences and contains specific terminologies that are rarely seen in general-domain text. This also causes lots of difficulties in natural language understanding in

general, and AMR parsing in our study. In the second part of our thesis, we provide a literature survey over different methods in AMR parsing and show their performances on analyzing legal documents. We divide current AMR parsers into three main approaches: *alignment-based, grammar-based* and *neural-based*, then choose 7 typical parsers that obtain high accuracy on a benchmark AMR dataset for our experiments. Since there is no AMR dataset in legal domain, we manually annotate our own dataset for testing. We call this dataset JCivilCode, with the source text extracted from the Japanese Civil Code in English version. We conduct our parsing experiments on both our legal dataset JCivilCode and the general domain dataset in various ranges of sentence length. Our results show the current limitations and also open a room for improvements of current parsing techniques for legal domain adaptation.

For the legal text generation direction, we observe that text generated from AMR using current deep learning models usually become awkward with lots of "out of vocabulary" and repetitive tokens. In the third part of our thesis, we propose our domain adaptation method by some modifications in the training and decoding phase of the encoder-decoder AMR generation model to have a better text realization. Our model is tested using our manually annotated legal dataset JCivilCode, showing an improvement compared to the baseline model.

To summarize, our study in AMR parsing and generation along with the legal domain adaptation contribute to the literature of semantic representation. Despite some improvements and findings, our work still remains specific drawbacks. Since our first results are still preliminary, we figure out several ideas to improve our performance in the future.

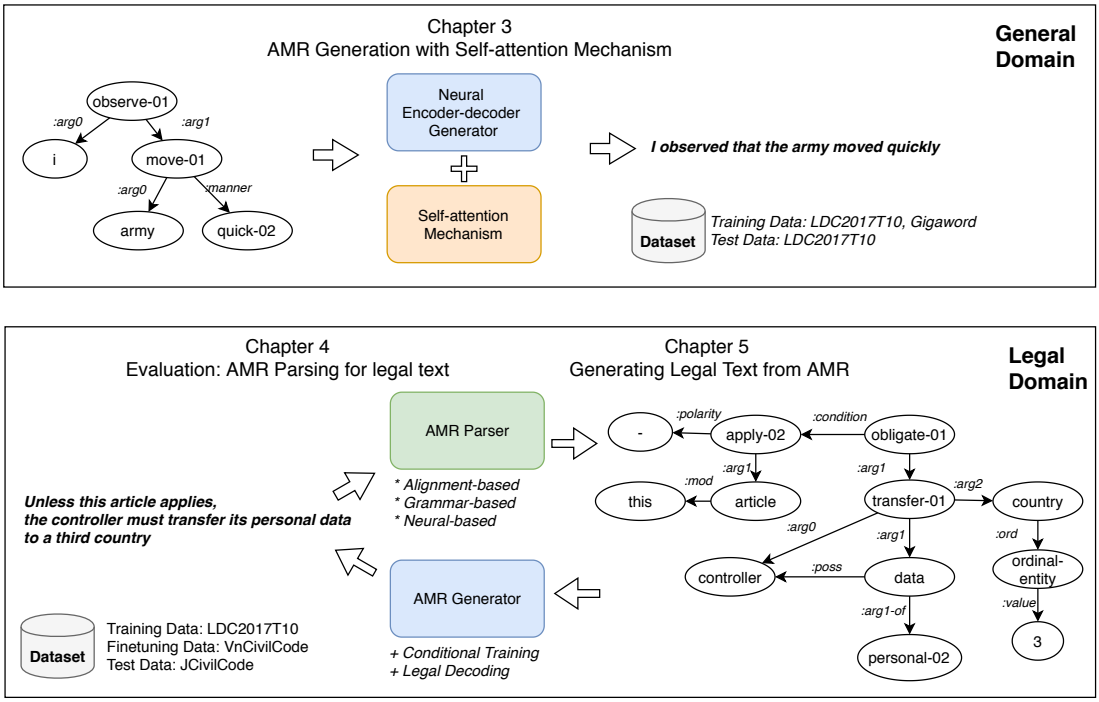

Figure 1: Dissertation Outline

2 Research Purpose

The main contributions of this dissertation can be summarized as follow:

- **Legal Dataset JCivilCode** : We extract the first four chapters in the Japanese civil code (in English version) and manually annotate in the format of AMR. This is the first AMR dataset in legal domain, rather than popular datasets in general domain. Though the number of samples is still small, this dataset helps develop the research in domain adaptation in legal domain. We conduct experiments of different AMR parsers in three main parsing approaches on our annotated dataset to see the quality of legal text parsing. Our results show the difficulties as well as suggest several ideas for future improvement in AMR parsing for legal documents.

- **Self-attention text generation from AMR graphs**: We propose a transformer approach in converting an AMR graph into a natural language sentence. We incorporate the self-attention mechanism into the encoder-decoder model in both sequence to sequence and graph to sequence strategies. Evaluating by a benchmark dataset, our method obtains comparative results comparing to existing neural models in the literature.

- **Legal Style Text Generation**: We propose two modifications in the training and decoding phase of the neural graph to sequence AMR generation model. With these modifications, we provide more legal-related constraints for generating text from an input legal AMR. We then finetune our models using a silver annotated dataset in legal domain. The experimental results prove the effectiveness of our method over the baseline model.

Based on promising results of this thesis, we figure out some future works:

- Data augmentation: Training data play an important role in training deep neural network models for both parsing and generation. To obtain more high-quality legal data, we need to discover some data augmentation techniques, i.e. data recombination strategy that generate new AMRs from current pair of (sentence, AMR) based on heuristic rules.

- Logical complexity: as mentioned in two chapters of this dissertation, this complexity causes lots of errors for both AMR parsing and generation models. Several research have been proposed to generate logical forms from text and entities graph recently. We plan to explore these works to build a logical attention mechanism to capture these information more effetively.

- AMR applications: we plan to apply AMR in several downstream problems such as Legal Question Answering. Despite the capablity of AMR in expressing the "who is doing what to whom" aspects, there are not many works investigate the application of AMR in question answering (QA), especially for legal domain. We expect our legal dataset can be enlarged and contributes to a legal QA system.

3 Research Accomplishment

[1] Vu Trong Sinh and Le Minh Nguyen. A study on self-attention mechanism for amr-to-text generation. In *Natural Language Processing and Information Systems - 24th International Conference on Applications of Natural Language to Information Systems, NLDB 2019, Salford, UK, June 26-28, 2019, Proceedings*, pages 321–328, 2019. doi: 10.1007/978-3-030-23281-8_27. URL `https://doi.org/10.1007/978-3-030-23281-8_27`.

[2] Vu Trong Sinh and Nguyen Le Minh. An empirical evaluation of amr parsing for legal documents. In *New Frontiers in Artificial Intelligence JSAI-isAI 2018 Workshops, JURISIN, AI-Biz, SKL, LENLS, IDAA*, pages 131–145, Yokohama, Japan, 2018.

[3] Lai Dac Viet, Vu Trong Sinh, Nguyen Le Minh, and Ken Satoh. Convamr: Abstract meaning representation parsing for legal document. *arXiv preprint arXiv:1711.06141*, 2017.

[4] Vu Trong Sinh, Nguyen Le Minh, and Satoh Ken. Legal text generation from abstract meaning representation. In *Proceedings of the 32nd International Conference on Legal Knowledge and Information Systems*, Madrid, Spain, 2019.

[5] Vu Trong Sinh, Nguyen Le Minh, and Satoh Ken. Abstract meaning representation for legal documents. *submitted to the Journal of Artificial Intelligence and Law - Springer Nature*, 2020.

Keywords: **Abstract Meaning Representation**, **Deep Learning**, **Semantic Parsing**, **Text Generation**, **Legal Domain**.