Abtract

In recent years, thanks to breakthroughs in neural network techniques especially attentive deep learning models, natural language processing has made many impressive achievements. However, automated legal word processing is still a difficult branch of natural language processing. Legal sentences are often long and contain complicated legal terminologies. Hence, models that work well on general documents still face challenges in dealing with legal documents. We have verified the existence of this problem with our experiments in this work. In this dissertation, we selectively present the main achievements in improving attentive neural networks in automatic legal document processing. Language models tend to grow larger and larger, though, without expert knowledge, these models can still fail in domain adaptation, especially for specialized fields like law.

This dissertation has three main tasks to achieve the goal of improving attentive models in legal document processing. First, we survey and verify the factors affecting the performance of the models when operating on a specific domain such as law. This investigation is to provide clearer insights to improve models in this domain. Second, as pretrained language models are recently the most well-known attentive approaches in natural language processing, we provide methods to create language models specific to the legal domain, producing state-of-the-art results on reliable datasets. These models are built on features from the data of legal documents, with the goal of overcoming the challenges found in our previous survey. Third, besides the approach to let the model learn completely from raw data, we propose and prove the effectiveness of using different knowledge sources to inject into the model in different ways to adjust their output. This approach not only increases explainability but also allows humans to control pretrained language models and take advantage of the knowledge resources available during the development of the field such as vocabulary, grammar, logic and law.

Keywords: Legal Text Processing, Attentive Neural Network, Deep Legal, Pretrained Language Model, Knowledge Injection.