A Study on Recognition of Requisite Part and Effectuation Part in Law Sentences

by

Ngo Xuan Bach (0910021)

School of Information Science

Japan Advanced Institute of Science and Technology

February 08, 2011

Keywords: Law Sentences, Requisite Part, Effectuation Part, Sequence Learning, Semi-supervised Learning

Abstract

In recent years, a new research field called Legal Engineering has been proposed in the 21st Century COE Program, Verifiable and Evolvable e-Society [4,5,6]. Legal Engineering serves to exam and verify whether a law has been established appropriately according to its purpose, whether the law is consistent with related laws, and whether the law has been modified, added, and deleted consistently. There are two important goals of Legal Engineering. The first goal is to help experts make complete and consistent laws, and the other is to design an information system which works based on laws.

Legal Engineering regards laws as a kind of software for our society. Specifically, laws such as pension law are specifications for information systems such as pension systems. To achieve a trustworthy society, laws need to be verified about their consistency and contradiction.

Legal texts have some specific characteristics that make them different from other daily-use documents. Legal texts are usually long and complicated. They are composed by experts who spent a lot of time to write and check carefully. One of the most important characteristics of legal texts is that legal texts usually have some specific structures.

In most cases, a law sentence can roughly be divided into two logical parts: *requisite part* and *effectuation part*. The requisite part and the effectuation part of a law sentence are composed from three parts: a *topic part*, an *antecedent part*, and a *consequent part*. In a law sentence, the consequent part usually describes a law provision, and the antecedent part describes cases in which the law provision can be applied. The topic part describes subjects which are related to the law provision.

Analyzing the logical structure of legal texts is a key problem in Legal Engineering. The results of this process are helpful to not only lawyers but also people who want to

understand the legal texts. This is a preliminary step to support other tasks in legal text processing (such as translating legal articles into logical and formal representations, legal article retrieval, legal text summarization, question answering in legal domains, etc) and serve to verify legal documents [9].

In this thesis, we focus on two tasks which analyze the logical structure of legal texts at the sentence level and the paragraph level, respectively: *Recognition of Requisite Part and Effectuation Part in Law Sentences* (or RRE task) and *Recognition of Requisite Parts and Effectuation Parts in Paragraphs Consisting of Multiple Sentences* (or RREP task). The goal of the RRE task is to recognize logical parts given a law sentence. The goal of the RREP task is to recognize logical parts and logical structures (a set of some related logical parts) given a legal paragraph.

For the RRE task, our approach is modeling the task as a sequence learning problem and using Conditional random fields [7,11] as learning method. We present several supervised learning models for the RRE task: word-based model (consider a law sentence as a sequence of words), Bunsetsu-based model (consider a law sentence as a sequence of Bunsetsus), and reranking model (use a linear score function to rerank N-best outputs of the Bunsetsu-based model with a variant of the perceptron algorithm [2,3]). Our experimental results show that word features are very important to the RRE task. Features other than word and part-of-speech features are not effective. In the problem modeling aspect, modeling based on Bunsetsu is better than modeling based on words. An other interesting result is that the model using only head words and functional words gives better performance than the model using all words.

We describe an investigation on contributions of words to the RRE task. To investigate contributions of a word w, we remove all features related to w, and compare the performance of the system before and after removing features. A decrease in the performance means that word w is important to the task. Our experimental results show that words that are important to human in recognizing logical structures of law sentences are also important to our statistical machine learning models.

We also present a simple semi-supervised learning method for the RRE task. The main idea of this method is to use *unsupervised* word representation as extra word features of a *supervised* model. Extra word features derived from Brown word clusters [1,8,12] are integrated into our Bunsetsu-based model and reranking model. Experimental results show that semi-supervised learning method outperforms supervised models and the improvement is more clearly when the amount of training data is small. In the RRE task, our best model achieves 88.84% in $F_{\beta=1}$ score on the Japanese National Pension Law corpus.

For the RREP task, we present a two-phase framework in which we recognize logical parts in the first phase and logical structures in the second phase. We divide logical parts in a law sentence into some layers and provide a multi-layer sequence learning model to recognize them. We consider the sub-task of recognizing logical structures as an optimization problem on a weighted graph, where each node corresponds to a logical part and a sub-graph corresponds to a logical structure. The weight on an edge expresses

the degree that two nodes belonging the same logical structures. We also give a heuristic algorithm to solve the optimization problem. The main idea of our algorithm is to choose as many positive edges as possible and as few negative edges as possible. Our models achieve 74.37% in recognizing logical parts, 75.89% in recognizing logical structures, and 51.12% in the whole task on the Japanese National Pension Law corpus. Our results provide a baseline for further researches on this interesting task.

In the future, we will continue to investigate these two tasks. We will compare Markov and semi-Markov models (semi-CRFs [10]) on the RRE task. Some studies show that sometimes semi-Markov models can improve performance over Markov models [8]. For the RREP task, we will try to integrate two phases into a unified process, where we recognize both logical parts and logical structures at the same time. From the results of these two tasks, we will also investigate the task of *Translating Legal Articles into Logical and Formal Representations*, where the input is a set of documents and the outputs are their formal representations.

References

1. P.F. Brown, P.V. deSouza, R.L. Mercer, V.J.D. Pietra, J.C. Lai. Class-Based n-gram Models of Natural Language. In *Computational Linguistics*, Volume 18, Issue 4, pp.467-479, 1992.

2. M. Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of EMNLP*, pp.1-8, 2002.

3. M. Collins, T. Koo. Discriminative Reranking for Natural Language Parsing. In *Computational Linguistics*, Volume 31, Issue 1, pp.25-70, 2005.

4. T. Katayama. The current status of the art of the 21st COE programs in the information sciences field. Verifiable and evolvable e-society - realization of trustworthy e-society by computer science - (in Japanese). In *IPSJ (Information Processing Society of Japan) Journal*, 46(5), pp.515-521, 2005.

5. T. Katayama. Legal engineering - an engineering approach to laws in e-society age. In *Proceedings of the 1st International Workshop on JURISIN*, 2007.

6. T. Katayama, A. Shimazu, S. Tojo, K. Futatsugi, K. Ochimizu. e-Society and Legal Engineering (in Japanese). In *Journal of the Japanese Society for Artificial Intelligence*, 23(4), pp.529-536, 2008.

7. J. Lafferty, A. McCallum, F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pp.282-289, 2001.

8. P. Liang. Semi-Supervised Learning for Natural Language. *Master's thesis, Massachusetts Institute of Technology*, 2005.

9. M. Nakamura, S. Nobuoka, A. Shimazu. Towards Translation of Legal Sentences into Logical Forms. In *Proceedings of the 1st International Workshop on JURISIN*, 2007.

10. S. Sarawagi, W. Cohen. Semi-Markov Conditional Random Fields for Information Extraction. In *Proceedings of NIPS*, pp.1185–1192, 2004.

11. C. Sutton, A. McCallum. An Introduction to Conditional Random Fields for Relational Learning. In *Introduction to Statistical Relational Learning*, Chapter 4, MIT Press, 2006.

12. J. Turian, L. Ratinov, Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*, pp.384-394, 2010.