A Study of Open Information Extraction from Legal Texts

Chien-Xuan Tran[1], Minh-Le Nguyen[1] and Ken Satoh[2]

[1] Japan Advanced Institute of Science and Technology (JAIST)
[2] National Institute of Informatics (NII)
chien-tran@jaist.ac.jp, nguyenml@jaist.ac.jp, ksatoh@nii.ac.jp

Abstract. This paper presents an on-going work in a research aiming to extract structured information from text data in the legal domain. We first introduce the problem of open information extraction and some challenges needed to be tackled in the legal domain. We describe two systems in this research and show our evaluation results when running these systems on the text corpus of Japan Civil Code, including our analysis of the results. Lastly, we give a conclusion and our future research directions.

Keywords: document analysis, information extraction, legal texts.

1 Introduction

In natural language processing, information extraction (IE) is a task which aims to extract structured data from the raw text. Structured data here refer to events, entities, facts or relationship between entities presented in the text. This structured information allows computers to perform logic inference or computation on the data, which is challenging if we only use raw text representation.

Traditional approaches on IE focused on extracting a specific type of relations or events from texts. These approaches require a pre-specified vocabulary and relations to instruct how to extract the needed information. On the other hand, there is also a need to massively extract all possible relations from plain text without any pre-specified relations. This is called open information extraction (Open IE). Usually, we want to extract binary relations in the form *(head, label, tail)*, in which *head* and *tail* are entities and *label* is their relationship.

Because of the characteristics of Open IE, we can apply it to different domains without much work. In this research, we are interested in applying Open IE to texts in the legal domain. Our goal is to extract useful information from the legal text and use this information to enhance the performance of other systems such as legal information retrieval or legal question answering system. Building such systems in the legal domain is not a new task, but it possesses many challenges compared to general systems. Some challenges are due to the nature of the legal text, such as:

– Legal texts often contain words or phrases which have a specific interpretation and might be different from the common usage [3].

– Legal sentences can be very long and complicated to understand. Following is a sentence copied from Article 556(1) of the Japan Civil Code:
If no period is provided in relation to the manifestation of intention set forth in the preceding paragraph, the other party to the pre-contact may issue a notice of demand to the other party, specifying a reasonable period, to the effect that the other party is to give a definite answer as to whether or not he/she will complete the sale within that period.
– Legal sentences often have clauses in forms like conditional, cause-effect or comparison.

Common approaches for building information retrieval system are query-based. These approaches, however, are not performing well in legal domain due to synonymy and ambivalence of words[1]. Hence, many researchers proposed using legal ontology for improving the performance of legal information retrieval system and achieved promising results[1, 2].

The main challenge with ontology-based approaches in legal domain is how to build the ontology. This process is not trivial because it requires knowledge of the law experts and it is time-consuming. By using Open IE, we are able to massively extract entities and relationship from a huge amount of legal texts and map them into a structured knowledge-base. This knowledge base can directly be used as an ontology or be indirectly used to support the ontology construction process, thus saving both time and cost of ontology construction. To the best of our knowledge, there are no Open IE systems built specifically for the legal domain. Our interest is to find out the limitation of some generic Open IE systems in the legal domain and propose to build a better system.

2 Experiment on two Open IE systems

There have been several studies about Open IE and many systems were developed for this purpose. In the scope of our research, we experimented with two systems, each follows a different approach to extract target relations. This comparison serves as a foundation for building our own system in legal domain in the future.

The first system is ReVerb [4], one of well-known systems in Open IE. It follows a rule-based approach to quickly extract binary relations from texts. In ReVerb, a sentence is first POS tagged and chunked using OpenNLP tool[3]. Then, it uses rules similar to regular expression to find target relations in the parsed text. These rules are carefully designed by the authors and they are fixed in its implementation. In addition, ReVerb also has a regression classifier for assigning a confidence score to a relation to say how good a relation is. This score acts as a trade-off between precision and recall.

Unlike ReVerb which applies rules directly on the chunked text to extract relations, ArgOE [5] is another tool which applies rules on dependency relations

[3] https://opennlp.apache.org/

between words in a sentence. It uses DepPattern[4] to perform partial parsing before extracting relations. Similar to ReVerb, its rules are also carefully designed by the authors. This method was shown to achieve a better performance than ReVerb while its running time is not marginally different.

Evaluation on Japan Civil Code

In order to evaluate the performance of ReVerb and ArgOE in the legal domain, we conducted an experiment using the English content of Japan Civil Code[5], which was provided in the COLIEE shared task in 2015[6]. We first split it into individual sentences, then we ran ReVerb and ArgOE on these sentences to extract all possible relations. We manually assigned *GOOD* or *BAD* label to each relation based on two criteria similar to the criteria used for evaluating ReVerb: (i) *informative*: whether the relation contains critical information, and (ii) *coherent*: whether the relation is meaningful.

Due to our limited time and resources, we only evaluated first 200 sentences from the corpus (in the total of 1191 sentences). Our evaluation result is presented in Table 1.

Table 1. Evaluation result of ReVerb and ArgOE on Japan Civil Code.

	ReVerb	**ArgOE**
Number of sentences with no relations	19	14
Total number of extracted relations	431	828
Number of correct relations	269	381
Number of *incoherent* relations	114	357
Number of *uninformative* relations	48	90

As we can see from the results, the number of relations extracted by ReVerb is only about a half of ArgOE. One reason is because ArgOE has access to the dependency tree so it has more information and able to extract more relations, while Reverb can only rely on POS and chunking information. In addition, its set of rules is different to Reverb. For example, considering the sentence: *The minor is unable to perform the relevant business for any reason.* As shown in Table 2, Reverb is not able to extract any relation from it meanwhile ArgOE is able to extract 3 relations.

Even though ArgOE extracts more relations, their quality is a question. According to our evaluation, more than half of relations extracted by ArgOE are incoherent or uninformative. This makes ArgOE perform poorly in term of accuracy. In the example shown in Table 2, the first relation extracted is considered as *uninformative* in our evaluation because it does not contain critical information

[4] http://gramatica.usc.es/pln/tools/deppattern.html
[5] http://www.moj.go.jp/content/000056024.pdf
[6] http://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2015/

Table 2. Relations extracted by ReVerb and ArgOE on a sentence.

Sentence	The minor is unable to perform the relevant business for any reason.
ReVerb	*No relation*
ArgOE	*(The minor, is unable to perform for, any reason)*
	(The minor, is unable to perform, the relevant business)
	(The minor, is unable to perform the relevant business for, any reason)

to express the relation. In our understanding, this can be due to the tight dependency of ArgOE with its dependency parser. DepPattern is a general-purpose partial dependency parser, which might be weak to correctly parse long and complex sentences in the legal domain. This consequently leads to more errors when extracting relations. In addition, the number of *incoherent* relations in both tools is much greater than *uninformative* relations. This is mainly because they both have difficulty identifying correct arguments for long sentence. It is especially severe in case of ArgOE for the same reason we have just mentioned. In Table 3, we show an example of incorrect relations extracted by ArgOE from a long legal sentence.

Table 3. An example of incoherents and uninformative relations extracted by ArgOE

A liquidator who has assumed his/her office during the course of the liquidation must register his/her name and domicile within two weeks from the assumption of his/her office at the location of the principal office , and within three weeks from the assumption of his/her office at the location of its other office , and file such matter with the competent government agency .	
Incorrect extracted relations	**Reason**
(A liquidator, must register within, two weeks)	uninformative
(A liquidator, must register from, the assumption of his/her office)	incoherent
(A liquidator, must register his/her name and domicile from, the assumption of his/her office)	incoherent
(A liquidator, must register his/her name and domicile at, the location of its other office)	incoherent
(A liquidator, has assumed during, the course of the liquidation)	incoherent
(A liquidator, has assumed, his/her office)	uninformative

An obvious advantage of ArgOE is its ability to extract relations from sentences containing relative clause. ReVerb's rules are designed to extract relations only from a sequence of continuous words, therefore it is unable to capture relations containing distant words (which is commonly in relative clause). ArgOE, on the other hand, makes use of the dependency between words thus has no issue with this type of clause as long as the parser correctly identifies the dependent words. We present one example in Table 4 to demonstrate ArgOE's capability to extract relations from sentences containing relative clause.

Additionally, ArgOE is superior to ReVerb in terms of the number of supported languages. ArgOE supports 5 languages: English, Spanish, Portuguese, French and Galician. Meanwhile ReVerb currently supports extracting relations from English texts only.

Table 4. Comparision of ReVerb and ArgOE on a sentence containing relative clause.

Sentence	Neither party to a juristic act **which** is subject to any condition may infringe the interests of the counterparty **which** should arise from such juristic act upon fulfillment of the condition while it is uncertain whether or not such condition has been fulfilled.
ReVerb	*(a juristic act, is subject to, any condition)* *(any condition, may infringe, the interests of the counterparty)*
ArgOE	*(a juristic act, is subject to, any condition)* *(Neither party to a juristic act, may infringe, the interests of the counterparty)*

3 Conclusion and Future work

We introduced the problem of Open IE and challenges when applying it in the legal domain. We then described briefly two Open IE systems, ReVerb and ArgOE, and presented our evaluation when running them on the English content of Japan Civil Code. Finally, we gave our analysis on the experimental results and pointed out weaknesses in each system.

In our future research, we would like to adopt a data-driven approach for extracting target relations from sentences. This system will use machine learning to automatically learn features from the text. Our target is to tackle the challenges of Open IE from a different point of view and achieve a competitive performance compared to other rule-based systems.

Acknowledgment

This work was supported by JAIST CREST, Japan.

References

1. Saravanan, M., Ravindran, B., Raman, S.: Improving legal information retrieval using an ontological framework. Artificial Intelligence and Law, 17(2), 101-124. (2009)
2. Schweighofer, E., Geist, A.: Legal Query Expansion using Ontologies and Relevance Feedback. In LOAIT (pp. 149-160) (2007)
3. Francesconi, E., Montemagni, S., Peters, W., Tiscornia, D. (Eds.): Semantic processing of legal texts: Where the language of law meets the law of language (Vol. 6036). Springer (2010)

4. Fader, A., Soderland, S., Etzioni, O.: Identifying relations for open information extraction. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1535-1545). Association for Computational Linguistics (2011)
5. Gamallo, P., Garcia, M.: Multilingual Open Information Extraction. In Portuguese Conference on Artificial Intelligence (pp. 711-722). Springer International Publishing (2015)