Treatment of Legal Sentences Including Itemization Written in Japanese, English and Vietnamese – Towards Translation into Logical Forms –

Makoto Nakamura, Yusuke Kimura, Minh Quang Nhat Pham, Minh Le Nguyen, and Akira Shimazu

School of Information Science
Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi, Ishikawa, 923-1292, Japan
{mnakamur, minhpqn, nguyenml, shimazu}@jaist.ac.jp

Abstract. This paper investigates a difference of legal sentences including itemized expressions among three languages. Thus far, we have developed a system for translating legal documents into logical formulae. Although our system basically converts words and phrases in a target sentence into predicates in a logical formula, it generates some useless predicates for itemized and referential expressions. In the previous study, focusing on Japanese Law, we have made a front end system which substitutes corresponding referent phrases for these expressions. In this paper, we examine our approach to the Vietnamese law and the English version of it. Our linguistic analysis shows the difference in notation among languages or nations, and we extracted conventional expressions denoting itemization for each language. The experimental result shows high accuracy in terms of generating independent, plain sentences from the ones including itemization. The proposed system generates a meaningful text with high readability, which can be input into our translation system.

1 Introduction

Acquisition of knowledge bases by automatically reading natural language texts has widely been studied. While the definition of semantic representation differs depending on what the language processing systems deal with, some systems try to generate logical formulae based on first order predicate logic [1–3]. Legal documents are suited for knowledge acquisition, since it is different from daily-use texts in that they are described with characteristic expressions in order to avoid ambiguous description. Taking into account linguistic analysis of the expressions, we can extract logical structure of the legal documents.

This paper reports how to treat sentences including itemization, restricting texts to legal documents written in some languages. Thus far, we have developed a system for automatically converting Japanese legal documents into logical forms. The system analyzes law sentences, determines logical structures,

and then generates logical expressions. We have shown our system provides high accuracy in terms of generating logical predicates corresponding to words and their semantic relations [4]. However, some predicates generated concerned with itemization and reference were meaningless, because predicates converted from words and phrases, such as "the items below," "Article 5," and so on are not intrinsic to a logical representation of the sentence. These words should be replaced with appropriate phrases before the process of translation. Since Japanese legal documents have strict rules concerning its description and modification, we succeeded to extract conventional expressions in the documents by some regular expressions [5]. In order to investigate whether the proper method depends on the language or the nation establishing laws, we try to apply our approach to the Vietnamese law and the English version of it. Therefore, our purpose in this paper is to investigate a difference of the method to generate independent, plain sentences from legal texts including itemization. This study is regarded as a derivation of the series of our main study to translate legal documents into logical forms [4]. We expect that this fruitful results are able to be applied not only to the Vietnamese version of the translation system into logical forms, but also to a support system for reading legal documents and a text-to-speech system.

In this paper, we introduce our current system and its problems in Section 2. In Section 3 we show analysis of law sentences including itemization or reference, and we propose a method to rewrite the law sentences into plain sentences in Section 4. We also examine our new method and report its results in Section 5. Finally, we conclude and describe our future work in Section 6.

2 The Current System and Problems

In this section, we describe our current system for translating legal documents into logical forms, and its problems. We call our system WILDCATS[1].

2.1 Legal Engineering and Wildcats

A new research field called *Legal Engineering* was proposed in the 21st Century COE Program, Verifiable and Evolvable e-Society [6, 7]. Legal Engineering serves for computer-aided examination and verification of whether a law has been established appropriately according to its purpose, whether there are logical contradictions or problems in the document per se, whether the law is consistent with related laws, and whether its revisions have been modified, added, and deleted consistently. One approach to verifying law sentences is to convert law sentences into logical or formal expressions and to verify them based on inference [8]. Our system, Wildcats, takes charge of text processing.

[1] WILDCATS is an abbreviation of " 'Wildcats' Is a Legal Domain Controller As a Translation System."

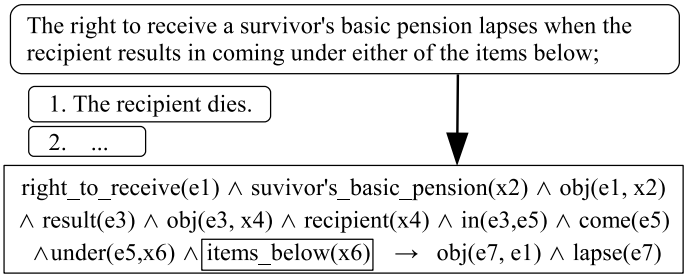

Fig. 1. Converting a law sentence including a reference phrase

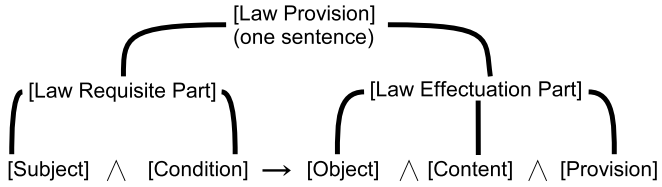

Fig. 2. Structure of requisition and effectuation [9]

2.2 Wildcats

Here, we explain an outline of our current system, which derives logical forms from law sentences. We show an example of input and output in Fig. 1.

In most cases, a law sentence in Japanese Law consists of a law requisite part and a law effectuation part, which designate its legal logical structure [9, 10]. Structure of a sentence in terms of these parts is shown in Fig. 2. The law requisite part is further divided into a subject part and a condition part, and the law effectuation part is divided into an object, content, and provision part.

Dividing a sentence into these two parts in the pre-processing stage makes the main procedure more efficient and accurate. Nagai et al. [10] proposed an acquisition model for this structure from Japanese law sentences. Dealing with strict linguistic constraints of law sentences, their model succeeded in acquiring the structures at fairly high accuracy using a simple method, which specifies the surface forms of law sentences. Our approach is different from theirs in that we consider some semantic analyses in order to represent logical formulae.

The following list is the procedure for one sentence. We repeat it when we process a set of sentences.

1. Analyzing morphology by JUMAN and parsing a target sentence by KNP.
2. Splitting the sentence based on the characteristic structure of a law sentence.
3. Assignment of modal operators with the cue of auxiliary verbs.
4. Making one paraphrase of multiple similar expressions for unified expression.
5. Analyzing clauses and noun phrases using a case frame dictionary.

6. Assigning variables and logical predicates. We assign verb phrases and *sahen*-nouns[2] to a logical predicate and an event variable, e_i, and other content words to x_j, which represents an argument of a logical predicate.
7. Building a logical formula based on fragments of logical connectives, modal operators, and predicates.

The procedure is roughly divided into two parts. One is to make the outside frame of the logical form (Step 1 to 3 and 7), which corresponds to the legal logical structure shown in Fig. 2. The other (Step 4 to 6) is for the inside frame. We assign noun phrases to bound variables and predicates using a case frame dictionary.

2.3 Problems of Wildcats

When our system converts a law sentence including referential phrases, they are not interpreted correctly. For example, in Fig. 1, the enclosed predicate "items_below(x6)" is useless. This is because the generated predicates lack information which must be referred. These phrases should be replaced with appropriate phrases in the items before the process of translation into logical forms. Therefore, substituting corresponding referent phrases for these expressions appropriately, our proposed system in this paper generates a meaningful text with high readability, and then the generated text can be input to the translation system. For example, the system should process the following instead of the input sentences in Fig. 1; "The right to receive a survivor's basic pension lapses when the recipient dies."

2.4 Scope of Our Study

The scope of the study in this paper is restricted to sentences including itemized expressions written in Japanese, Vietnamese, and English. In the preceding study [5], focusing on Japanese legal sentences, we showed that the system successfully worked well using some simple regular expressions. In this paper, we apply it to Vietnamese Law on Enterprises and the English version of it. In general, a notation of law sentences is strictly affected not only by the written language, but also by the nation establishing the law. In other words, it depends on the legislative process whether or not our simple approach is useful for other countries' law. Therefore, in the next section we explain the background of the legislative process, and show linguistic analysis of law sentences.

[2] A *sahen*-noun is a noun which can become a verb with the suffix *-suru*.

3 Analysis of Law Sentences with Itemization

3.1 Characteristics of Law Documents among Nations[3]

The legislation system of Japanese law is rational to keep the notation of expressions of law. A bill is basically proposed by the proper authority of the law. Once the authority has made a draft of the bill, it negotiates with other authorities. After that, the cabinet strictly examines the draft in terms of inconsistency with other laws, expressions, formats and so on, using the database of legislation. Therefore, this system keeps even the usage of comma and period.

Vietnamese laws are also strictly examined by a number of organizations concerning the laws before passing the National Diet. There is a standard of writing notations in administrative documents in Vietnam too. However, it is unknown whether the notation of expressions is surveyed by some authorities as strict as Japan.

Not all other countries have the system similar to Japan. In the United Kingdom, the administrative structure for the description check is not strict as Japan, since in most of cases the draft of a bill is prepared by an outsider of the ministry. In the United States of America, there is no organization or system for consolidating expressions of laws. In Asian countries except Japan and Korea, each ministry independently makes out a draft of a bill without coordinating various opinions from other ministries. As a result, the notation of bills becomes different among ministries. Moreover, in some countries bills are often modified during deliberation in the national assembly, while bills mostly pass the National Diet in Japan as drafted. This political process causes inconsistencies in notation.

3.2 Definition of Itemization

In general, a law consists of a number of articles, each of which is further subdivided into a number of paragraphs or items. Both articles, paragraphs, and items have sequential numbers with a typeface different from each other. For example, in the English version of Vietnamese law, articles start with "Article 1," "Article 2" and so on, paragraphs with "$1., 2., \ldots,$" and items with "$(a), (b), \ldots.$" Although there are a few differences of notation between English and Vietnamese, it can be dealt with easily by preprocessing.

In the case of Japanese law, we basically recognize an itemized expression as a noun phrase or a subordinate clause following the upper paragraph or article, which consists of sentences. An example of itemized expression is shown in Fig. 1. On the contrary, in the case of Vietnamese law, even some articles are expressed as a phrase which lacks the subject or the main verb. We show an example in Fig. 3. Therefore, we define itemization, with which we deal in this study, as a phrase or a sentence following an article or a paragraph being a sentence. The article shown in Fig. 3 is not recognized as itemization, because it does not have a sentence.

[3] This section is written based on the experience by Prof. Matsuura in Nagoya University. There is no reference directly supporting the description. For more detail about the administrative structure of legislation of Japanese law, see Nagano [11].

Article 21 Contents of requests for business registration

1. Name of the enterprise.
2. Address of the head office of the enterprise; telephone number, facsimile number, email transaction address (if any).
3. ....

Fig. 3. Article 21 in Vietnamese Law on Enterprises

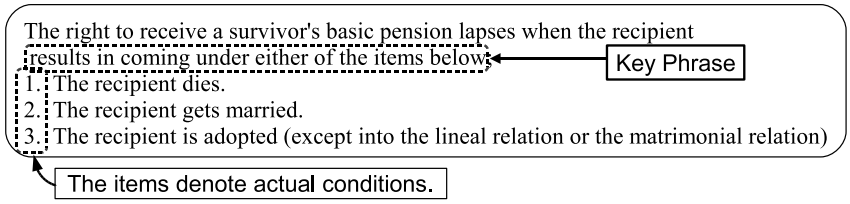

Fig. 4. Itemization of conditions in the law requisite part

3.3 Analysis of Itemization in Japanese Law Sentences

Some law sentences include itemization of conditions in the law requisite part, an example of which is shown in Fig. 4. The enclosed phrase should be replaced with one of the items denoting actual conditions. When one or more conditions are satisfied, the description in the law effectuation part becomes effective. We found 34 sentences of such a style in National Pension Law. Therefore, we considered a method to embed itemized conditions instead of cue phrases of itemization.

We defined *Key Phrases*, which always appear in sentences before itemization[4]. As we analyzed sentences from all 215 articles of the National Pension Law, the set of Key Phrases can be expressed as a regular expression, the diagram of which is shown in Fig. 5. For example, the phrase *"Tsugi no kaku gou ni gaitou suru ni itatta,"* meaning "to result in coming under either of the items below[5]," which is derived from the generative rule in Fig. 5, is regarded as a Key Phrase.

Itemized condition sentences appear next to sentences which contain Key Phrases. The last words of these sentences are "*Toki* (time)," "*Mono* (person)," and so on. In this paper, we call these sentences excluding the last words *Condition Items*. Key Phrases and Condition Items appearing in National Pension Law are shown in Table 1 and Table 2, respectively.

We will describe a method to remove itemization using Key Phrases and Condition Items in Section 4.

[4] There may be a proviso between the sentence and itemization.

[5] If we do not care about word-to-word translation for the Japanese law sentence, the following phrase is more appropriate; "to be included in one of the following cases."

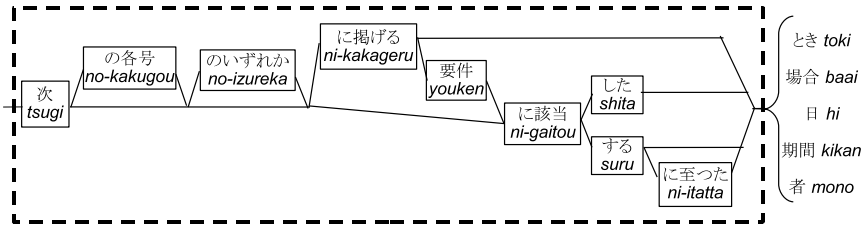

Fig. 5. Key phrases for itemization

Table 1. Frequency of Key Phrases

(**KP**: Key Phrase)

Format of KPs / Frequency			
KP + *toki*	(とき)	when	9
KP + *baai*	(場合)	case	9
KP + *mono*	(者)	person	6
KP + *hi*	(日)	day	3
KP + *kikan*	(期間)	period	1
KP + *youken*	(要件)	requirement	1
KP + a noun			5
Total			34

Table 2. Frequency of Condition Items

CI: Condition Items

Format of CIs / Frequency			
CI + *toki*	(とき)	when	106
CI + *koto*	(こと)	matter	4
CI + *mono*	(もの)	thing	3
CI + *mono*	(者)	person	2
CI + a noun			9
Total			124

3.4 Analyses of Itemization in Vietnamese Law Sentences Written in Vietnamese and English

In order to find *Key Phrases* and *Condition Items*, we analyzed 100 out of 172 articles in Vietnamese Law on Enterprises. Although all of Key Phrases identify one regular expression in Japanese law, we defined 15 and 14 rules of regular expression for Vietnamese and English, respectively. This means there are a variety of expressions denoting itemization in Vietnamese law. We show the set of rules for the English and Vietnamese versions of Vietnamese Law in Fig. 6 and Fig. 7, respectively. Since we manually made the sets of rules in English and Vietnamese separately, each rule in the English version does not correspond to that of Vietnamese with the same label, and vice versa.

In the English version of the law, we need to deal with inflection of words. The number of rules would be reduced, if we did not consider an irregular conjugation for some particular nouns. For example, Rule 9 accepts the phrase "following rights," "following obligations," or "following undertakings" and generates an appropriate phrase with the condition item, omitting the word "following" and the suffix '-s.' Rule 10 works the same with Rule 9 except replacing the suffix '-ies' to '-y.' Each regular expression accepts a number of Key Phrases corresponding to a Condition Item. In other words, some rules which could be merged with other rules are separated due to the different Condition Items.

```
1  ^(.*)\s+(the following terms shall be construed as follows:)$
2  as follows:
3  (at least the two following elements|in one of the following cases):
4  (enterprise|except) in the following cases:
5  ((in) (one|any) of|(with)|(in)) the following (manner|case|provision)s[:;]
6  (all of )?(t|T)he following (condition|case)s(.*[^:;\.])?([:;\.])$
7  following rights and duties:
8  following criteria and conditions:
9  following (right|obligation|undertaking)s:
10 following duties:
11 (one of |either of )?the (two )?following (act|manner)s(.*):
12 following attached (.+):
13 following ((main [a-zA-Z]+)|([a-zA-Z]+particulars)|
             ([a-zA-Z]+((,|\s+or|\s+and)\s+[a-zA-Z]+)*)(\s+.+)?):$
14 (in which|by way of|the right)s?:$
```

Fig. 6. Key Phrases for the English version of Vietnamese Law

```
1  (khi|nếu) (có|thuộc) (một trong|đủ) (những|các) (trường hợp|điều kiện) sau đây
2  trong (những|các) (trường hợp) sau đây
3  (phải) (.+) các tiêu chuẩn và điều kiện sau đây:
4  các (báo cáo|báo cáo và tài liệu|hoạt động|tài liệu) sau đây:
5  các (nội dung|vấn đề) sau đây
6  các nội dung.*:
7  có các (.+) sau đây:
8  (gồm)(.*):
9  Tổ chức, cá nhân sau đây
10 (có ít nhất) (.+) thành tố sau đây:
11 (theo|bằng) (các|một trong các|một trong hai) (.+) sau đây:
12 theo (quy định|nguyên tắc) sau đây:
13 (trong đó):
14 để thực hiện (một trong các|các) hành vi sau đây:
15 (Các|các|Những|những) (.+) sau đây
```

Fig. 7. Key Phrases for Vietnamese Law

Similar to Japanese, Vietnamese does not distinguish singular or plural nouns by inflection. Vietnamese distinguish singular and plural nouns by quantifiers which precede corresponding nouns, such as 'các (all),' 'những (some),' 'tất cả (every),' 'một (one),' 'hai (two),' and so on. The rules of Vietnamese are simpler than that of English, being not necessary for the process of inflection. Some rules similar to each other are distinguished depending on the Condition Items corresponding to the rule.

4 Method for Removing Itemization

In Section 3.3, we defined Key Phrases as cue phrases that always appear with itemization, like "*tsugi-no kaku gou no izureka ni gaitou-suru* ((something) to which either of the following items is applicable)," and we search for itemization with it. If a Key Phrase is found, we regard the following items as Condition Items, and replace the Key Phrase with one of the Condition Items for each. Then

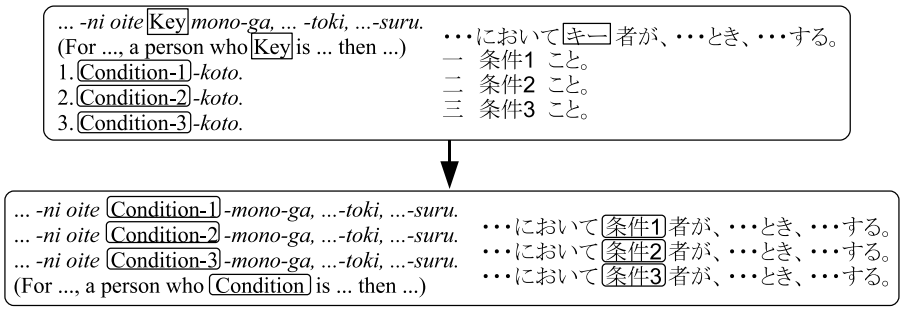

Fig. 8. Removing itemization

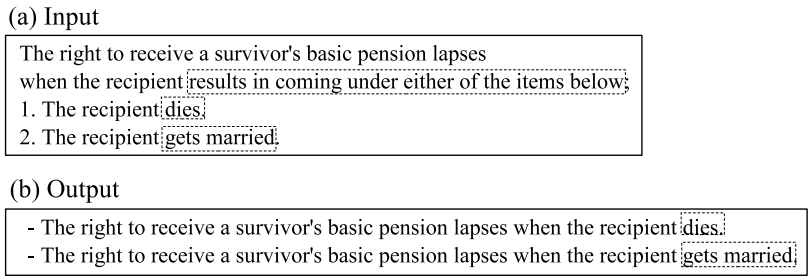

Fig. 9. An example of removing itemization

we have sentences which are understandable separately[6], as shown in Fig. 8. We show an example of the pair of input and output in Fig. 9.

The process of Vietnamese Law is different from that of Japanese in that there are a number of rules of regular expression. Since some rules conflict with other rules, priority is established in the order of the rule number. Each rule has a corresponding Condition Items, which are defined as regular expression.

5 Experiments and Results

We tested our system on itemization. The test set for each language is shown in Table 3. Since we extract Key Phrases of Japanese from National Pension Law, we used it for a closed test. For an open test we used Income Tax Law as the test set.

In these experiments, it is too difficult to establish a baseline due to the distinctiveness of our model and its target. Some studies which extract web contents from HTML or XML documents [12] may be able to deal with itemization in HTML or XML documents. However, our method is different from them in

[6] Even though the converted logical formulae are repetitive, there is no problem as long as the system gives the same logical predicates and variables to the repetitive phrases.

Table 3. Input texts for open and closed tests

	Test	#item	Test Set
Japanese	closed	124	National Pension Law
	open	548	Income Tax Law
Vietnamese	closed	357	Article 1-100 in Law on Enterprises
	open	275	Article 101-172 in Law on Enterprises
English	closed	354	Article 1-100 in Law on Enterprises
	open	273	Article 101-172 in Law on Enterprises

Table 4. Experimental results for removing itemization

	Test	#item	Succ	Over	Err	P	R
Japanese	closed	124	87	5	32	73.1%	70.2%
	open	548	219	122	207	51.4%	40.0%
Vietnamese	closed	357	309	24	24	92.7%	86.5%
	open	275	218	37	20	91.6%	79.3%
English	closed	354	336	0	18	94.9%	94.9%
	open	273	191	12	70	73.2%	70.0%

#item: the number of itemization to be processed, **Succ**: Succeeded, **Over**: Oversight, **Err**: Error, **P**: Precision, **R**: Recall

that it includes process to find itemized phrases without a tag, and to make plain sentences. We examine whether or not our model works well to the law documents in some languages, regardless of the linguistic characteristics, or of its nation which established the law.

We extract Key Phrases of both Vietnamese and English from 100 out of 172 articles in Vietnamese Law on Enterprises. Therefore, we use sentences from Article 1 to 100 for a closed test and from Article 101 to 172 for an open test.

The experimental results are shown in Table 4. In the experiment of Japanese, we found that 11 of the whole errors were items which denote a combination of a Condition Item and an object part in the law effectuation part. In other words, the objects of these sentences change depending on the Condition Items. An example is shown in Fig. 10. This article determines the revision of the rate after the base year about the national pension. An important thing here is that each item consists of a condition part and its result. That is, the first Key Phrase denoting "In the case of the following items," which is emphasized corresponds to the first phrases of each item, while the second Key Phrase denoting "on the basis of the rate on the item" which is underlined corresponds to the second phrases of each item. Our system did not deal with this type of itemization.

For the result of open test with Income tax Law, a little more than half of the sentences were processed well, there seems to be some difference in notation between National Pension Law and Income Tax Law. Particularly, we found the increase of itemization consisting of a combination of a Condition Item and an object part to 84. Results will be improved after an analysis of the mistakes.

Paragraph 2, Article 27-3, National Pension Law *In the case of the following items,* the revision of the rate after the base year is fixed on the basis of the rate on the item, regardless of the provisions stipulated in the preceding paragraph.

1. The price rate exceeds the nominal net wage rate, and the nominal net wage rate exceeds 1 — the nominal net wage rate
2. The price rate exceeds 1, and the nominal net wage rate falls below 1 — 1

Fig. 10. Example of failure in Japanese National Pension Law

The results of both English and Vietnamese show high accuracy in terms of removing itemized expressions. This is because the number of rules of regular expression is increased to 14 and 15, while there is only one rule for Japanese. In the English test, we found that some Key Phrases were followed by a number of types of Condition Items different from each other, so that the set of rules did not cover all the Key Phrases even in the closed test. In the case of Vietnamese law text, we found some errors that the meaning of sentences generated are different from original meaning, and ungrammatical sentences may be generated.

Overall accuracy would be improved depending on the rule set of regular expression. Therefore, we can conclude that our method is quite suitable not only for Japanese legal texts but also for other languages.

6 Conclusion

In this paper we proposed a method to rewrite legal sentences including itemization into independent, plain sentences, focusing on laws written in three languages. From the linguistic analyses, we showed the difference of the number of regular expressions for extracting Key Phrases between Japanese and Vietnamese/English. It implies that fixed expressions are often used in Japanese Law. In Vietnamese law documents, there are some common words and phrases which appear with high frequency at the Key Phrases. If we have the survey of words and phrases more, it may be helpful in building regular expression rules.

In the experiments, we showed that the system successfully extracted itemized expressions with some exceptions. We consider that the system is useful not only for the front end of our main system, Wildcats, but also for assistance in reading legal documents. We can improve this system by introducing a method for measuring readability of the output sentences. In terms of Vietnamese version of translation system (Wildcats), we need to wait for the development of dependency parser for Vietnamese.

Acknowledgment We would like to give special thanks to Prof. Yoshiharu Matsuura in Nagoya University for discussion about the differences among nations in notation of law documents. This research was partly supported by the 21st Century COE Program 'Verifiable and Evolvable e-Society' and Grant-in-Aid for Scientific Research (19650028).

References

1. Hobbs, J.R., Stickel, M., Martin, P., Edwards, D.: Interpretation as abduction. In: Proceedings of the 26th annual meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1988) 95–103
2. Mulkar, R., Hobbs, J.R., Hovy, E.: Learning from reading syntactically complex biology texts. In: Proceedings of the 8th International Symposium on Logical Formalizations of Commonsense Reasoning, part of the AAAI Spring Symposium Series. (2007)
3. Mulkar, R., Hobbs, J.R., Hovy, E., Chalupsky, H., Lin, C.Y.: Learning by reading: Two experiments. In: Proceedings of IJCAI 2007 workshop on Knowledge and Reasoning for Answering Questions. (2007)
4. Nakamura, M., Nobuoka, S., Shimazu, A.: Towards translation of legal sentences into logical forms. In Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T., eds.: New Frontiers in Artificial Intelligence: JSAI 2007 Conference and Workshops, Miyazaki, Japan, June 18-22, 2007, Revised Selected Papers. Volume 4914 of Lecture Notes in Computer Science., Springer (2008) 349–362
5. Kimura, Y., Nakamura, M., Shimazu, A.: Treatment of legal sentences including itemized and referential expressions –towards translation into logical forms–. In: Proc. of the 2nd Intl. Workshop on JURISIN. (2008) 73–82
6. Katayama, T.: The current status of the art of the 21st COE programs in the information sciences field (2): Verifiable and evolvable e-society - realization of trustworthy e-society by computer science - (in Japanese). IPSJ (Information Processing Society of Japan) Journal **46** (2005) 515–521
7. Katayama, T.: Legal engineering – an engineering approach to laws in e-society age –. In: Proc. of the 1st Intl. Workshop on JURISIN. (2007)
8. Hagiwara, S., Tojo, S.: Stable legal knowledge with regard to contradictory arguments. In: AIA'06: Proceedings of the 24th IASTED international conference on Artificial intelligence and applications, Anaheim, CA, USA, ACTA Press (2006) 323–328
9. Tanaka, K., Kawazoe, I., Narita, H.: Standard structure of legal provisions - for the legal knowledge processing by natural language - (in Japanese). In: IPSJ Research Report on Natural Language Processing. (1993) 79–86
10. Nagai, H., Nakamura, T., Nomura, H.: Skeleton structure acquisition of Japanese law sentences based on linguistic characteristics. In: Proc. of NLPRS'95, Vol 1. (1995) 143–148
11. Nagano, H.: Foundation and Common sence for legislation (in Japanese). Jichitai Houmu Kenkyuu (2005)
12. Liu, B., Grossman, R., Zhai, Y.: Mining data records in web pages. In: Proc. of the ninth ACM SIGKDD. (2003) 601–606