Extraction of Legal Definitions from a Japanese Statutory Corpus – Toward Construction of a Legal Term Ontology

Makoto Nakamura*, Yasuhiro Ogawa°, Katsuhiko Toyama°
*Japan Legal Information Institute, Graduate School of Law, Nagoya University
°Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
Email: mnakamur@law.nagoya-u.ac.jp

Abstract. We have been faced with several problems in the process of constructing the Japanese Law Translation Database, one of which is disunity for word selection. A legal terminology is useful for the unification of translation. Our purpose in this paper is to provide legal definitions and their explanations, which include semantic relations with other legal terms. We propose an automatic method for extracting them from a Japanese statutory corpus. Our experimental result shows that over 14,000 legal terms and their explanations in total were extracted with high precision, and the recall rate became better than the previous work. Since some definitions are explained with other legal terms, these relations would help to construct a legal term ontology.

Keywords: terminology, Japanese statutes, automatic extraction

1. Introduction

The Japanese law translation database was released under the Japanese government's leadership in 2009 (Toyama et al., 2011). The number of laws translated into English for publication has increased little by little; as of Sept. 1, 2012, only 264 out of over 7,700 laws and regulations have been translated. One of the most important problems to be solved in the process of translation is disunity for word selection. Since a number of human translators are involved, many legal terms in Japanese correspond to a variety of English translations. We have found even law titles are translated in different ways for citation (Sekine et al., 2012). An identical expression should have an identical translation for the sake of the conservation of meaning. Although the government has compiled a standard translation dictionary, the number of entries, which is 4,482 in the latest version, is not sufficient for the unification of translation.

Our goal in this research is to construct a legal ontology consisting of a sufficient number of entries and semantic relations between them for translation, which would help not only systematic translations but also appropriate word selections depending on context. At the beginning, we focus on collecting legal terms defined in statutes. Therefore, our purpose in this paper is to provide legal definitions and their explanations, which include semantic relations with other legal terms. We propose an automatic method for extracting them from a Japanese statutory corpus. This task is an application of natural language processing to Japanese legal texts.

This paper is organized as follows. In Section 2, we explain the difficulty of Japanese text processing in the legal domain following linguistic characteristics of Japanese and Japanese statutes. In Section 3, we propose a method for extracting legal definitions from Japanese statutes. Section 4 shows how it works in experiments, and we conclude and discuss our future work in Section 5.

2. Japanese Legal Text Processing

We have three different problems with respect to this study:

1. Difficulty of Japanese text processing in the legal domain
2. Definition of legal terms
3. Compilation of a statutory corpus.

The following subsections clarify our position in terms of these problems.

2.1. Characteristics of Japanese in Legal Domain

In this subsection, we describe linguistic characteristics of Japanese in the legal domain, considering three aspects, that is, characters, words and syntax. Due to limited space, we focus on characteristics related to laws and text processing, following a general explanation.

2.1.1. *Characters*

Japanese language is written with three types of characters: Hiragana, Katakana, and Kanji (Chinese characters), which have 83, 86, and 2,136 character types, respectively. Alphabetic letters, Arabic numbers, and Roman numbers are also used for a paragraph or an item number. Both Hiragana and Katakana are phonograms, which we can pronounce as written. Kanji characters are ideograms, and can be replaced with a number of Hiragana or Katakana characters, which can lead to ambiguous expressions. Figure 1 shows examples of Hiragana, Katakana and Kanji, all of which refer to the word “law” and are pronounced as *“houritsu.”*

It is difficult to show the exact number of Chinese characters, which is estimated to be over 100,000. Although Japanese employs a part of

English	Hiragana	Katakana	Kanji	Alphabetic letters
law	ほうりつ	ホーリツ	法律	*houritsu*

Figure 1. Japanese characters

them, it is still over 40,000 character types. The government defined a set of regular-use Kanji consisting of 2,136 characters, as of 2010, and describes legislation with them as possible.

We are faced with a problem regarding characters from the viewpoint of text processing. Since we are collecting all the statutes that have been enacted for over 60 years, old statutes are typically written with outdated characters that are not used any more. Therefore, even if we try to search for a word from a database, a simple string match causes failure in finding one with old characters.

2.1.2. *Words*

In general, nouns are often expressed by Kanji, while other content words such as verbs, adjectives and adverbs typically use Kanji with Hiragana as a suffix. Katakana is mainly used for foreign words. Functional words such as case particles and auxiliary verbs are mainly written with Hiragana. Japanese is different from English in that words are not segmented by space.

In legal documents, there are many functional words that are strictly defined on use. For example, there are several words for coordinate conjunction words corresponding to ‘and’ or ‘or’, each of which is used in order of priority in nested conjunctions.

Since Japanese is not an inflectional language, words can be extracted without a morphological process. Therefore, morphological analysis of Japanese is to separate words and to attach a part of speech tag to each morpheme.

2.1.3. *Syntax*

Japanese is typologically classified as an SOV language, whereas English is classified as SVO. Since basic grammatical relations are marked by special particles as a suffix, Japanese word order is not as strict as English order. If a noun phrase consists of a number of words, its head noun is located at the end.

Japanese sentences written in statutes use special terms and syntactic rules peculiar to the legal domain. In addition, the frequent use of coordinate conjunctions makes the syntactic structure of a sentence complicated. Complementary clauses embedded in the sentence in parenthesis also make the whole sentence hard to read. Legal sentences adhere closely to these syntactic rules, which result in heavy use of boilerplate expressions.

As mentioned above, special syntactic rules in the legal documents cause a decline in accuracy of analysis with a syntactic parser. It turns out, as long as boilerplate expressions are used often, a simple method for surface pattern recognition is sufficient for legal text processing.

2.2. Legal Terms and Their Explanations in Definition

What are recognized as legal terms to be collected depends on the purpose (Lame, 2005; Höfler et al., 2011; Winkels and Hoekstra, 2012). In this paper, we define legal terms as those explicitly defined prior to use in a law, each of which consists of a tuple of a legal term and its explanation. They are typically placed as the following forms:

- An independent provision
- An insert statement in parenthesis.

Figure 2[1] shows examples of definitions both in provision and in parenthesis, where Article 2 is an independent provision that defines the term "Gas Business," and a definition in parenthesis appears in Article 42. A defined term is put in quotations (「*term*」 / "*term*") in Japanese, and the underlined phrase denotes its explanation.

While the definition of a legal term is written in an explanatory sentence, the second item is further divided into two types:

- A defined term appears in parenthesis following a phrase as its explanation in the main text. Abbreviations of terms are often defined with the style.
- A sentence in parenthesis explains a legal term just before the parenthesis, as shown in Article 42 of Act on the Treatment of Prisoners of War and Other Detainees in Armed Attack Situations in Figure 3.

Legal terms or explanations in parenthesis are easily extracted by analysis of the character string, while the analysis of content outside parenthesis is difficult. For example, in Article 42 in Figure 2, the term "Gas Utility, etc." in parenthesis is easy to extract, while it is possible for its explanation to be "Supplier," "Business Supplier," …, and "A Gas Utility Service Provider or a Wholesale Gas Business Supplier;" that

[1] Hereafter, an English sentence following the slash is a translation of the Japanese one before it.

ガス事業法 / Gas Business Act

（昭和二十九年法律第五十一号） / (Act No.51 of 1954)

第二条　この法律において、「ガス事業」とは、一般の需用に応じ導管によりガスを供給する事業をいう。/
Article 2 The term “*Gas Business*” as used in this Act shall mean the business of supplying gas via pipelines to meet general demand.

第四十二条　ガス事業者又は卸供給事業者（以下「ガス事業者等」という。）は、その事業の用に供するため、道路、橋、みぞ、河川、堤防その他公共の用に供せられる土地の地上又は地中に導管を設賀する必要があるときは、その効用を妨げない限度において、その官理者の許可を受けて、これを使用することができる。/
Article 42 A Gas Utility Service Provider or a Wholesale Gas Business Supplier (these persons shall hereinafter be referred to as a “*Gas Utility, etc.*”) may, when it is necessary to install pipelines on or under a road, bridge, ditch, river, embankment or other public land in order to use such pipelines for the businesses, use them with permission from the administrator thereof to the extent that such use does not impair their usability.

Figure 2. Example of definitions in provision and in parenthesis

is, we need to determine which word the explanation phrase starts from in the main text.

2.3. Japanese Statutory Corpus

There are at least two public websites where we can read Japanese statutes; one is run by the Ministry of Internal Affairs and Communications[2], and the other is linked from the website of the House of Representatives[3]. In addition, newly enacted laws go into the official gazette, which is also accessible online[4]. These three databases have good and bad points from the viewpoint of readability, that is, typos by OCR error, error correction and consolidation by amendment laws. If they are compiled with digitally scanned data, they may include typos by OCR error. For error correction, some typos have been included at the time of release, for which an errata list is published in the

[2] `http://law.e-gov.go.jp/cgi-bin/idxsearch.cgi`

[3] `http://www.shugiin.go.jp/index.nsf/html/index_housei.htm`

[4] `https://search.npb.go.jp/kanpou/`

武力攻撃事態における捕虜等の取扱いに関する法律 /

Act on the Treatment of Prisoners of War and Other Detainees in Armed Attack Situations
（平成十六年法律第百十七号） / (Act No. 117 of 2004)

第四十二条　宗教要員等（宗教要員及び第六十九条の規定により第六十四条第四号に掲げる業務に従事することを許された捕虜 をいう。第八十四条第三項において同じ。）は、捕虜収容所内において、被収容者の行う第四十条に規定する宗教上の行為を補助し、又は前条第一項に規定する宗教上の儀式行事を行うことができる。/
Article 42 In the prisoner of war camp, *chaplains, etc.* (i.e. chaplains, and prisoners of war who are permitted to engage in works listed in item (iv) of Article 64 pursuant to the provision of Article 69. The same shall apply in paragraph (3) of Article 84) may assist the detainees in performing religious acts prescribed in Article 40 and may perform the religious ceremonies prescribed in paragraph (1) of the preceding Article.

第百八条 ２　審査会は、資格認定審査請求をした者（以下「資格認定審査請求人」という。）が前項の期間内に補正をしないときは、裁決をもって、資格認定審査請求を却下することができる。ただし、その不適法が軽微なものであるときは、この限りでない。/
Article 108 (2) In the case that a person who has appealed for a review on the recognition of internment status (hereinafter referred to as "*the applicant of the appeal for review on the recognition of internment status*") fails to correct the defect within the period set forth in the preceding paragraph, the Review Board may dismiss the appeal for review on the recognition of internment status by determination; provided, however, that this shall not apply when such defect is minor.

Figure 3. Example of definitions in provision and in parenthesis

following official gazette. Finally, the databases are also different if laws are consolidated by amendment laws. Regardless, as is common among the governmental websites, these are far from a satisfactory level of quality as a database.

We compiled a corpus of all of the Japanese statutes consisting of 9,915 acts that have been enacted up to 2012 since promulgation of the new constitution of Japan in 1947. The size of corpus is 252MB. The statutory corpus is based on articles of legislation in the official gazettes. Since most of them are digitally scanned, there are many typos that are not included in the published version. We need to deal with them, developing a preprocessor.

Since amendment laws are to describe how to revise the pre-existing laws with amendment sentences, it would be difficult to properly extract legal terms unless consolidation is properly processed. Therefore, we eliminate in advance all laws concerning amendment and repeal of pre-existing laws, which are inferable from the title of the laws.

3. Our Approach to Extraction of Legal Definitions

Despite the presence of high-quality dependency parsers for Japanese, we cannot count upon their performance with legal texts. As mentioned in Section 2.1, since legal sentences are to avoid ambiguity of expression, they are likely to be long and syntactically complicated, which often lead to failure of parsing. Therefore, we employ a simple method without a parser.

We have proposed methods to extract legal terms in different forms; one is for an independent sentence for definition (Nakamura et al., 2012), and the other is for an expression parenthesized in a sentence (Nakamura et al., 2013). For the former, the legal term's definition is written in an explanatory sentence. Therefore, we made a set of patterns for extracting definitions from the legal corpus.

For the latter, although legal terms defined in parenthesis are easily extracted using a simple pattern match, it is difficult to extract their explanations. We need to find which word in the sentence an explanation starts from, while it ends just before the parenthesis. Since explanations are likely to start from the beginning of the sentence, we restrict our target to the ones satisfying this condition.

In addition, we deal with another condition that the subject with comma appearing at the beginning of the sentence is ignored, as shown in Article 108 (2) in Figure 3, where the subject with comma located at the beginning is out of the explanation. This additional rule makes it possible to extract a part of definitions in parenthesis located in the object.

4. Experimental Results

The number of definitions and their explanations collected by our method is shown Table I. The scores of precision are calculated from 100 samples chosen at random. The scores of recall are calculated based on

Table I. Analysis of collected definitions and their explanations

definitions	# of tokens	# of types	precision	recall
in provisions	5,250	3,799	0.980	0.980
in parenthesis	9,624	6,030	0.850	0.396

the assumption that at least all the legal terms in quotations are perfectly obtained by our method. Our experimental results show that over 14,000 terms were extracted in total with high precision. Since some terms are defined in multiple laws, the numbers of types are different from that of tokens. Although the recall rate for definitions in parenthesis is still low, it exceeded the previous rate that was 0.262 (Nakamura et al., 2013). Additional extraction rules are further expected to salvage the rest. On the contrary, the score of precision for definitions in parenthesis is worse than the previous rate that was 0.875[5]. This is because our method does not work well for finding a defined term in terms of the second definition type: a sentence in parenthesis explains a legal term just before the parenthesis, as shown in Article 42 in Figure 3.

Some legal terms are explained using other legal terms defined in advance. In this case, these legal terms are recognized to have an inclusive relation, and a set of these relations between legal terms can form a network. Let us take the legal term “Gas Business” as an example. This term is defined in Article 2 in the Gas Business Act, and is referred to not only from the explanation of the term “Gas Facilities” in the same Act, but also from that of five legal terms in other acts. Table II shows a list of legal terms defined in provision whose explanation includes the term “Gas Business.” Even though a simple pattern match does not guarantee all the legal terms refer to this term, it can make a semantically connected network, which would be useful for translation. For instance, although the Japanese term corresponding to “(gas) rate” is defined a total of four times, it has several candidates for translation such as *charge, fee, rate, toll, fare,* and so on. The semantic network may suggest an appropriate translation; that is, the one defined in the Local Tax Act includes the term “Gas Business” in the explanation, which can lead to a proper translation as “rate.”

Since the most frequent terms may play an important role in the network, we searched all relations in the statutory corpus. Table III shows a list of frequently referred to legal terms. We expect important legal terms to appear in the list, but contrary to our expectation, the

[5] The previous rate (Nakamura et al., 2013) was calculated as $\frac{1,501}{1,652} + \frac{440}{566} \simeq 0.875$.

Table II. List of legal terms defined in provision referring to the term “Gas Business”

Legal term		The title of Act
(English)	(Japanese)	
Gas Business	ガス事業	Gas Business Act (Act No.51 of 1954)
Gas Facilities	ガス工作物	Gas Business Act (Act No.51 of 1954)
(gas) rate	料金	Local Tax Act (Act No.226 of 1950)
Specified Gas Appliance	特定ガス消費機器	Act concerning Supervising Installation Work of Specified Gas Appliance (Act No.33 of 1979)
public benefit service operator	公益事業者	Special Measures Act on Preparation, etc. for Common-Use Tunnel (Act No.81 of 1963)
local public enterprise	地方公営企業	Act on Labor Relations of Local Public Enterprises (Act No.289 of 1952)
business operator	事業者	Enterprise Rationalization Promotion Act (Act No.5 of 1952)

list is occupied by ordinary words that are defined as an abbreviation of some particular terms. Most of the terms seem to be used just as a common noun in the explanation because frequently used common nouns are defined as an abbreviation by chance. As shown in Table IV, legal terms defined in parenthesis seem to have a stronger tendency to be an abbreviation, which leads the network to become dense. Note that the number of appearances of the term “business” exceeds that of “business operator” in principle because the simple pattern match finds the character string “business” in “business operator.” This may be a shortcoming of our method.

5. Conclusion and Future Work

In this paper, we focused on Japanese statutory sentences. First of all, we compiled a statutory corpus consisting of all acts enacted up to 2012 since promulgation of the new constitution of Japan in 1947.

Based on our linguistic analysis, we found that legal documents are likely to use similar expressions. We reached a decision that surface pattern rules are sufficient for term extraction. As a result, we succeeded in finding over 14,000 terms with high precision. On the other hand,

Table III. List of frequently referred to terms (Provision)

#	English translation of legal terms	Japanese
1,028	business	事業
508	facility	施設
400	use	使用
364	area	区
267	business operator	事業者
246	utilization	利用
244	enterprise	企業
236	company	会社
197	organization	団体
169	development	開発

Table IV. List of frequently referred to terms (Parenthesis)

#	English translation of legal terms	Japanese
3,576	Act	法
1,185	business	事業
625	corporation	法人
605	Cabinet Order	令
523	plan	計画
472	partnership	組合
404	designation	指定
368	facility	施設
349	company	会社
315	incorporated administrative agency	独立行政法人

we detected some error-prone rules and a procedural mistake. This will be improved for the next version.

Since some legal terms are explained with other legal terms, we listed a set of inclusive relations between legal terms, which is regarded as a conceptual network and would help to construct a legal term ontology.

Acknowledgment

This research was partly supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B) No.23300094 and for Young Scientists (B) No.23700310.

References

Höfler, S., Bünzli, A., and Sugisaki, K. (2011), *Detecting Legal Definitions for Automated Style Checking in Draft Laws.* Technical Report CL-2011.01, University of Zurich, Institute of Computational Linguistics.

Lame, G. (2005), *Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations.* In Law and the Semantic Web, LNAI3369, pp. 169–184. Springer.

Nakamura, M., Kobayashi, R., Ogawa, Y., and Toyama, K. (2012), *A Pattern-Based Approach to Hyponymy Relation Acquisition for the Agricultural Thesaurus.* In Proceedings of AOS2012, pp. 2–9.

Nakamura, M., Ogawa, Y., and Toyama, K. (2013), *Extraction of Legal Terms and Their Explanations Defined in Parenthesis in Statutes.* In Proceedings of 19th Annual Meeting on Natural Language Processing (NLP2013), pp. 670–673. (in Japanese)

Sekine, Y., Toyama, K., Ogawa, Y., and Matsuura, Y. (2012), *The development of translation memory database system for law translation.* In Proceedings of 2012 Law via the Internet Conference, 21 pages.

Toyama, K., Saito, D., Sekine, Y., Ogawa, Y., Kakuta, T., Kimura, T., and Matsuura, Y. (2011), *Design and Development of Japanese Law Translation Database System.* In Proceedings of Law via the Internet, 12 pages.

Winkels, R. and Hoekstra, R. (2012), *Automatic extraction of legal concepts and definitions.* In Schäfer, B., editor, JURIX, volume 250 of Frontiers in Artificial Intelligence and Applications, pp. 157–166. IOS Press.