A Pattern-Based Approach to Hyponymy Relation Acquisition for the Agricultural Thesaurus

Makoto Nakamura[1], Ryusei Kobayashi[2], Yasuhiro Ogawa[1,2], and Katsuhiko Toyama[1,2]

[1] Japan Legal Information Institute, Graduate School of Law, Nagoya University
[2] Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
mnakamur@law.nagoya-u.ac.jp

Abstract. This paper aims to increment the vocabulary of an existing thesaurus using hyponymy relations. We focus on the agricultural thesaurus AGROVOC. Our main intent is to acquire AGROVOC-qualified candidates from the hyponymy relations of legal texts. We propose a pattern-based approach to hyponymy relation acquisition. We show that over 1,000 terms were extracted, some of which have not been registered in AGROVOC despite agricultural related terms.

1 Introduction

Legal terms are special, idiomatic expressions that often describe legal matters in legal documents. Since legal terms differ from daily use expression, they are defined by law prior to use. The goal of our study is to construct a legal ontology based on legal terms and the hyponymy relation between them contained in large collections of legal texts consisting of statutory sentences of laws and regulations. A hyponymy relation is a kind of IS-A relation between word senses, where one word sense (the hypernym) represents a more general concept than the other word sense (the hyponym). In particular, this paper aims to increment the vocabulary of an existing thesaurus using the hyponymy relations. We focus on the agricultural thesaurus AGROVOC [1].

AGROVOC [1] is the world's most comprehensive multilingual agricultural vocabulary [2]. It contains more than 40,000 concepts in up to 21 languages covering topics on food, nutrition, agriculture, fisheries, forestry, environment, and other related domains. AGROVOC is expressed in a Simple Knowledge Organization System (SKOS) and published as Linked Data [3]. All the terms or concepts have been added to the thesaurus by the domain experts in different languages. This laborious human work is very time consuming and expensive. At times, these partner organizations are unable to frequently update the terms due to unforeseen circumstances [4]. In order to prevent this problem, we are working toward the practical use of automatic term extraction from legal texts. We assume that terms appearing in laws and regulations are qualified for AGROVOC as long as they are related to the agricultural domain.

Our main idea is to use hyponymy relations of the legal texts to find candidates qualified for AGROVOC. Since terms registered in AGROVOC are definitively related to AGROVOC, their hypernyms or hyponyms are also considered to be registerable candidates. Therefore, our approach is that if one of the terms in a hyponymy relation has been registered, the other may be a candidate term. In addition, new links between a hypernym and a hyponym may be acquired when both of the terms in a relation have been registered, but are not linked to each other in AGROVOC.

Our paper is organized as follows: We first introduce previous studies on legal text processing in Section 2. In Section 3, we propose a method of hyponymy acquisition on the basis of analysis of Japanese legal texts. In Section 4, we examine how the method works in the legal domain, and we conclude with a short summary and outlook toward future research in Section 5.

2 Previous Works on Legal Text Processing

Thus far, researchers of natural language processing have studied legal text processing using surface pattern recognition. Surface pattern rules are typically described in regular expressions, which provide a concise and flexible means to match and extract strings of text. Kimura et al. [5] tried to acquire knowledge from itemized expressions in law texts. The experimental result showed that only a few surface patterns following itemization succeeded in extracting itemized expressions and representing semantics. Höfler et al. [6] have attempted to detect legal definitions in order to support domain-specific style checking in legislative drafts. These studies show that surface pattern recognition is sufficient for legal text processing because legal documents are often written with boilerplate expressions. In this point of view, the target of this study is different from that of probabilistic models for learning ontologies that expand existing ontologies taking into account both corpus-extracted evidence and the structure of the generated ontologies [7, 8].

It is possible for surface pattern rules to extract hyponymy relations as well as legal terms. For example, the expression "y is a (kind of) x," in which both x and y are noun phrases, implies that x is a hypernym of y, as well as "such x as y" [9, 10]. This approach is applicable to Japanese as well, as Ando et al. [11] proposed a set of Japanese surface patterns. These studies suggest that legal ontologies could automatically be constructed from legal texts containing boilerplate expressions.

3 Acquisition of Legal Terms From Legal Texts

3.1 Extracting Terms and Their Explanations in Japanese Legal Texts

We analyzed a set of statutory sentences from laws and regulations. Hereafter, we call this set '*the legal corpus*.' In general, a law consists of a number of

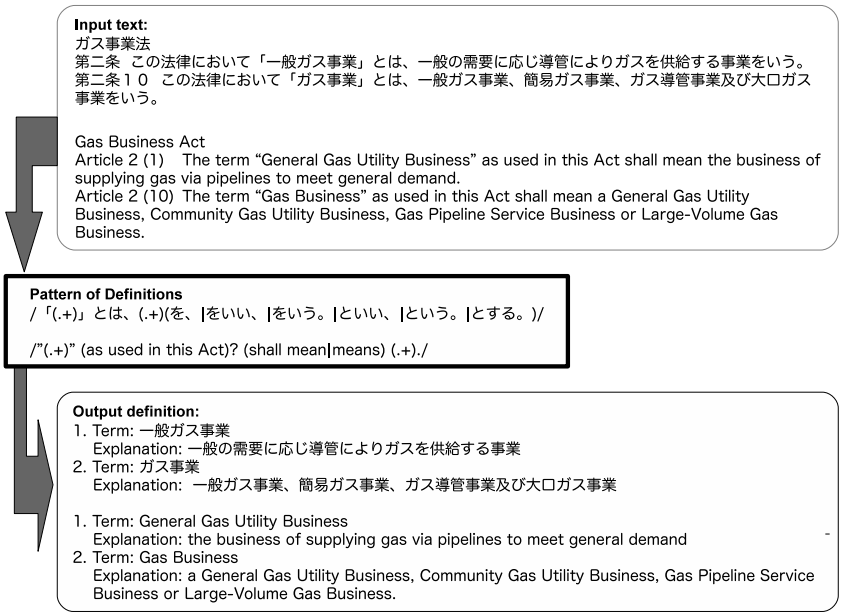

Fig. 1. Analysis of definitions by surface pattern rules

articles, each of which may be further subdivided into a number of paragraphs or items. Both articles, paragraphs, and items have sequential numbers with a typeface different from each other. In this study, we focus on definitions of legal terms for legal term extraction. Legal terms are defined prior to use in a law and are typically placed in Article 2 following the first article for that purpose. The legal term's definition is written in an explanatory sentence. Since explanations include hypernyms and/or hyponyms of the defined legal terms, hyponymy relations are likely to be acquired from these explanations. Therefore, we made a set of patterns for extracting definitions from the legal corpus. There are six rules in total.

Figure 1 illustrates that legal terms and their explanations are well extracted with a surface pattern described in a regular expression. Although we deal with Japanese texts, for ease of understanding we have included an English translation below every sentence. The set of regular expression rules are predefined, as with the middle box in the figure, where the symbol '`.+`' matches any character sequence. When one of the rules accepts the input sentence, a legal term and its explanation are extracted, corresponding to the first and second character sequences, respectively.

For example, Article 2 (1) of the Gas Business Act is a definition of the term 'General Gas Utility Business' explained with its hypernym 'business.' The pattern matches the sentence; that is, the term and the explanation are extracted as 'General Gas Utility Business' and 'the business of supplying gas via pipelines to

meet general demand,' respectively. In the same way, another term-explanation pair is extracted from Article 2 (10) of the Gas Business Act as 'Gas Business' and is defined as an enumeration of hyponyms.

3.2 Text Processing for Hyponymy Relations

From the viewpoint of text processing, Japanese is different from English in that words are not segmented by space. On the process of word segmentation, we often deal with *bunsetsu* instead of words. A bunsetsu is a linguistic unit in a Japanese sentence consisting of one or more content words with some particles or other suffixes. The English equivalent for a bunsetsu would be a small noun phrase, a prepositional phrase, or a verb phrase consisting of auxiliary verbs and a main verb, and so on [12].

For automatic acquisition of a hyponymy relation from a term-explanation pair, we need to find a head bunsetsu corresponding to the hypernym or hyponyms of the legal term from the explanation sentence. Therefore, we use a syntactic parser for extracting legal terms. CaboCha is a Japanese dependency parser using a cascaded chunking model [13] that enables state of the art analysis for Japanese dependency structure. Japanese sentences written in laws and regulations, however, use special terms and syntactic rules peculiar to the legal domain, which cause a decline in accuracy of analysis. In order to eliminate the possibility of error due to this syntactic peculiarity, we complemented the parser with some additional rules for the legal domain. This set of rules follows the same procedure for Ogawa et al. [12].

These explanations are classified into the intensive, extensive, and mixed types using some surface patterns. Figure 2 shows the process of extraction of the hypernym and the hyponym from the explanations shown in Fig. 1. The first explanation is the intensive type, in which the head bunsetsu of an explanation is the hypernym of the legal term. In this case, the term 'business' is the hypernym of 'General Gas Utility Business.' On the other hand, we can determine that an explanation belongs to the extensive type with cue phrases. The second sentence enumerates hyponyms of the term 'Gas Business' separated with commas (,) and 'or.' Even though further steps are required for Japanese text processing, due to limited space, we omit details of the procedure for extracting hyponymy relations from the explanations.

As a result, we can acquire a set of hyponymy relations forming a tuple of two noun phrases and a conceptual relation. Note that a number of tuples may be acquired from a definition.

4 Experiments

We examined a set of 109,380 Japanese legal sentences in 241 laws and regulations[3]. The corpus covers a wide variety of laws and regulations, such as

[3] This set is the one provided by the Japanese Law Translation System [15] (`http://www.japaneselawtranslation.go.jp/?re=02`) as of Dec. 2010.

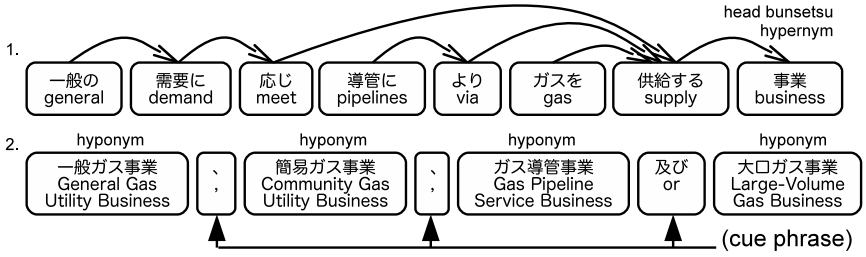

Fig. 2. Extraction of the hypernym and the hyponym

Table 1. The experimental result in finding terms related to AGROVOC

Category of a hyponymy pair	# of types	Precision
Category (i)-(iv)	1,027	†64.0%
Category (ii) & (iii)	222	67.1%
Category (ii)	137	89.1%
Category (iii)	75	21.3%
unknown	10	—
Category (iv)	25	88.0%
Existing relations	9	88.9%
New relations	16	87.5%

† is calculated from 100 samples chosen at random.

Bankruptcy Act, Measurement Act, Act on Promotion of Global Warming Countermeasures, and so on. Namely, we need not restrict the domain to the agricultural related laws.

Although the proposed method is measured by precision and recall, the legal corpus is too large to extract the whole answer set for calculating recall. Therefore, we calculated the precision of the output in terms of hyponymy relations, which are manually discriminated.

Hyponymy relations extracted from the legal corpus are classified into four categories:

Category (i) Neither the hypernym or the hyponym in a hyponymy pair is registered in AGROVOC; that is, neither is related to AGROVOC.
Category (ii) Only the hyponym is not registered in AGROVOC, while the hypernym is registered.
Category (iii) Only the hypernym is not registered in AGROVOC, while the hyponym is registered.
Category (iv) Both are registered; that is, neither is a candidate term.

We classified acquired hyponymy relations into four categories, shown in Table 1. The right most column shows the precision in terms of hyponymy relations between two terms in a pair. Table 2 shows an example of hyponymy relations in each category. Registered terms are shown in parenthesis. An asterisk denotes an example of failure.

Table 2. An example of hyponymy relations in each category

Cat		Hyponym	Hypernym
(i)		破産裁判所 / bankruptcy court	地方裁判所 / district court
	*	漁獲努力可能量 / total allowable effort	最高限度 / maximum limit
(ii)		一般ガス事業 / General Gas Utility Business	(事業 / business)
	*	真菌 / fungus	(有害植物 / injurious plant)
(iii)		(未受精卵 / Unfetilized Egg)	卵母細胞 / oocyte
	*	(装置 / equipment)	計量器 / measuring instrument
(iv)		(二酸化炭素 / Carbon dioxide)	(温室効果ガス / greenhouse gases)
		(土地 / land)	(不動産 / real property)
–	*	共同漁業 / common fishery – (漁業 / fishery)	

植物防疫法
第二条２ この法律で「有害植物」とは、真菌、粘菌、細菌、寄生植物及びウイルスであつて、直接又は間接に有用な植物を害するものをいう。

Plant Protection Act
Article 2 (2) Injurious plant as used in this Act means fungus, slime mold, bacterium, parasitic plant and virus that are injurious to useful plants directly or indirectly.

Fig. 3. Plant Protection Act

Overall, we acquired 1,027 types of hyponymy relations from 2,050 tokens. Some hypernyms are, however, too abstract, such as 'matter' and 'issue,' to be regarded as new terms related to the thesaurus. We eliminate them from Category (ii)-(iv) before calculation.

We acquired terms belonging to Category (ii) with high precision. In most cases, the defined term is not registered, while the hypernym in the explanation has been registered. Taking the case of Article 2 (1) in Figures 1 and 2 as an example, the term 'General Gas Utility Business' can be registered to AGROVOC. However, a problem still remains because it may be too abstract for the candidate term to link to the hypernym 'business' as a broader term. In the current model, we can see no alternative but to manually judge if the hypernym is appropriate as a broader term. The example of failure in Category (ii) in Table 2 was extracted from a sentence of Plant Protection Act shown in Fig. 3. Even though injurious plant may include fungus satisfying some condition in this Act, it is difficult to find a hyponymy relation between them.

Category (iii) has quite low precision. The error analysis revealed two problems. One is for the extraction of hyponyms from the explanation. In case of Category (iii), where a hypernym is mainly defined with hyponyms in the explanation like the second sentence in Fig. 2, explanations are often described in a relatively complicated sentence. These sentences result in a failure of text processing. As a result, 30 out of 75 types of relation, that is 40%, have little relation between terms.

The other problem comes from identification of the direction in the hyponymy relation. Actually, the number of extracted relations which should be categorized into Category (iii) is less than the experimental result in Table 1. This is because some incorrect regular expression rules brought reversed relations which should belong to Category (ii). In fact, 27 out of 75 types of relation, that is 36%, are included in this case. The example of failure in Category (iii) in Table 2 shows this mistake. After getting rid of these 27 types, Category (iii) improved precision to 33.3%. In addition, there are ten unknown cases in which the hyponymy relation answers between two terms are ambiguous in a number of laws. An example of unknown cases is shown at the bottom in Table 2. This relation was extracted from Fishery Act four times. One of them was categorized into Category (ii), while the rest were into Category (iii). These problems may also come from some incorrect procedures.

For Category (iv), the precision seems high enough, regardless of whether there is an existing relation or not. The second example of Category (iv) in Table 2 is the discovery of a new hyponymy relation between terms, although they have been registered in AGROVOC.

5 Conclusion and Future Work

In this paper, we focused on Japanese laws and regulations. Since legal documents are likely to use similar expressions, surface pattern rules work well for term extraction. As a result, we succeeded in finding 222 terms that seem qualified for AGROVOC with high precision. On the other hand, we detected some error-prone rules and a procedural mistake. This will be improved for the next version.

In future work, we plan to expand our method to multilingualism. Providing multilingual terminology could be supportive for AGROVOC. As long as boilerplate expressions are used often, as with Japanese laws, our simple method is applicable to any language. Although the problem of cost still remains for manually making patterns, this may be solved by some previous studies.

One simple solution toward multilingualism is to use bilingual lexicons as a dictionary for translating legal terms as our output to the other language. In this case, we need not define a set of patterns in the target language. Jin et al. [14] proposed a method for extracting bilingual lexicons from a bilingual corpus for English translations of major Japanese statutory laws called the Japanese Law Translation Database System (JLT) [15], which we can freely access via the Internet. The experimental result showed that information hidden in the context in one language is useful for term extraction in the other.

We have omitted the details of the procedure in Section 3 due to limited space. These details will be formalized in another paper.

Acknowledgments

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

1. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering Thesauri for New Applications: The AGROVOC Example. Journal of Digital Information **4** (2004)
2. Niu, M.L.S., Onn, K.W., Sean, L.Y., Lukose, D., Sadanandan, A.A.: Agriculture Linked Open Data. In: Proc of SNLP-AOS 2011. (2012) 7–12
3. AGROVOC: Food and Agriculture Organization of the United Nations (FAO), Rome, Italy (2012) `http://aims.fao.org/standards/agrovoc`.
4. Niu, M.L.S., Johannsen, G., Morshed, A., Rajbhandari, S., Keizer, J., Kiran, L., Thunkijjanukij, A., Selan, N.E.: Multilinguality in AGROVOC Concept Scheme: Challenges and Experiences. In: Proc of SNLP-AOS 2011. (2012) 13–19
5. Kimura, Y., Nakamura, M., Shimazu, A.: Treatment of Legal Sentences Including Itemized and Referential Expressions –Towards Translation into Logical Forms–. In: New Frontiers in Artificial Intelligence. LNAI5447, Springer (2009) 242–253
6. Höfler, S., Sugisaki, K.: From Drafting Guideline to Error Detection: Automating Style Checking for Legislative Texts. In: Proc of CL&W 2012. (2012) 9–18
7. Fallucchi, F., Zanzotto, F.M. In: Exploiting Transitivity in Probabilistic Models for Ontology Learning. IGI Global (2012) 259–293
8. Snow, R., Jurafsky, D., Ng, A.Y.: Semantic taxonomy induction from heterogenous evidence. In: Proc of COLING/ACL2006. ACL-44, Stroudsburg, PA, USA, ACL (2006) 801–808
9. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: An On-line Lexical Database. Journal of Lexicography **3** (1990) 235–244
10. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: COLING. (1992) 539–545
11. Ando, M., Sekine, S., Ishizaki, S.: Automatic Extraction of Hyponyms from Japanese Newspapers Using Lexico-syntactic Patterns. In: Proc of LREC2004. (2004) 387–390
12. Ogawa, Y., Yamada, M., Toyama, K.: Design and Compilation of Syntactically Tagged Corpus of Japanese Statutory Sentences. In: New Frontiers in Artificial Intelligence. LNAI6797, Springer (2011) 141–152
13. Taku Kudo, Y.M.: Japanese Dependency Analysis using Cascaded Chunking. In: CoNLL 2002. (2002) 63–69
14. Jin, R., Ogawa, Y., Rachmatullah, A., Toyama, K.: Bootstrapping-based Extraction of Bilingual Dictionary Terms from Parallel Corpus. In: Proc of SNLP-AOS 2011. (2012) 95–99
15. Toyama, K., Saito, D., Sekine, Y., Ogawa, Y., Kakuta, T., Kimura, T., Matsuura, Y.: Design and Development of Japanese Law Translation Database System. In: Proc of Law via the Internet. (2011) 12 pages