Translation Disambiguation of Patent Sentences using Case Frames

Shoichi YOKOYAMA
Yamagata University
4-3-16 Jonan, Yonezawa
Yamagata 992-8510, JAPAN
yokoyama@yz.yamagata-u.ac.jp

Masumi OKUYAMA
Yamagata University
(Now in Alpha Systems Inc.)

Abstract

Patent sentences have long, complicated structures, and correctly analyzing and translating such sentences is difficult. We have classified the modified relationships in Japanese patent sentences and constructed an error correcting system for analysis of such sentences. In the present paper, we investigate manually whether the Japanese case frames, which are automatically derived from a web corpus, correspond to disambiguated English words. The present study reveals that case frames of verbs originating from active nouns may be useful for disambiguating English verbs, whereas the case frames of traditional Japanese verbs are problematic.

Keywords: patent sentence, case frame, translation disambiguation, Japanese-English correspondence

1 Introduction[1]

Most patent sentences have long, complicated structures, including problems, claims, expressions, and details. If these sentences are input to a machine translation system, their complexity results in insufficient analysis, and so correct translation cannot be performed. Considering simultaneous international information exchange and time and cost savings for translation, for example, automatic processing of patent sentences is a very important area of research.

In the present paper, we discuss translation disambiguation using case frames. We use case frames registered in the case frame lexicon (Kawahara, 2005 and 2006), which was automatically produced from a web corpus.

Practically speaking, we have investigated manually whether differences in Japanese case frames reflect differences translation into English. The results revealed that a number of traditional Japanese verbs ("wago-dousi" in Japanese) are misclassified in the case frame lexicon, and so translation is not well disambiguated. However, many verbs originating from active nouns ("sahen" verbs in Japanese) have restricted meanings and so are easily translated by reflecting the classification of case frames.

Improvement of the case frame lexicon might provide an effective syntactic procedure by which to machine-translate patent sentences having long, complicated structures.

2 Research Background

A number of studies have examined the properties of patent sentences (Yokoyama, 2005 and 2007), and patterns of modified relationships in Japanese patent sentences have been classified (Yokoyama, 2005). An error correcting system that automatically corrects erroneous modified relationships in Japanese sentences has been constructed for the analysis of Japanese patent sentences (Yokoyama, 2007). In addition, a number of papers on patent sentences were presented and related issues were discussed at the

[1] This is an English, extended version of (Okuyama, 2009).

MT Summit X Workshop, 2005 and the MT Summit XI Workshop, 2007.

In the present study, we consider the case frame structure, which classifies verbs into a combination of meaning and attributed noun class. If the different combination can be translated into the different English representation, we can use and disambiguate the case frame as a clue in word selection and/or selection restriction in machine translation.

In the present study, we use the patent database (Japio, 2004, JPO), which contains all of the patent applications presented to the Japan Patent Office (JPO) over the course of one year. This database is open to the public. We first analyze and classify the case frames included in patent sentences and investigate the translation correspondence between Japanese and English patent sentences, which are translated by human translators. We then consider the possibility of machine translation using case frames.

2.1 Patent sentences

In the present study, we used the Japio 2004 database, part of which is freely available to the public at the website of the Japan Patent Office (JPO). Our patent translation research committee, which is made up of members from AAMT and the Japan Patent Information Organization (Japio), has constructed a patent database. In this database, the abstract parts were translated by human translators from Japanese to English. We use the original Japanese text and its English translation as the material for the present study.

2.2 Kurohashi-Nagao Parser (KNP)

The Kurohashi-Nagao Parser (KNP) is a freeware parser for Japanese developed at Kyoto University. The input is the result of morphological analysis of Japanese sentences, and several morphemes are concatenated as a "bunsetu." We use the information on case frames that is extracted during the analysis.

Figures 1 and 2 show the analytical results for the Japanese verb "kaihatusuru" (to develop) obtained using KNP. The term "#ID: …" is the number of the patent and its problem. Sentences are listed starting from the next line. Under the dotted line, "<Case Structure Analysis Data>" presents the results of case frame analysis.

```
#ID:2004089106_PROBLEM_2-2
ゲルマイクロドロップ法及びフローサイトメトリー法を組み合わせた自然界から微生物を取得する方法を開発する。
(The method obtaining bacteria is developed in combination of gel micro-drop method and flow cytometry method.)
…….
<Case Structure Analysis Data>
【開発／かいはつ】　動 [1] D 方法を《ヲ》
(to develop)  vt.  [1] (method)
```

Fig.1 Example of KNP analysis (1)

```
#ID:200400800046_PROBLEM_1-1
口腔疾患等に対する有効成分の探索・評価および口腔疾患予防剤等の開発に有効なインビトロバイオフィルムを提供すること。
(Supplying the in vitro biofilm effective to search and estimate active components for oral diseases etc. and to develop oral disease prophylactics etc.
……..
<Case Structure Analysis Data>
【開発／かいはつ】　動 [3] D 剤等の《ヲ／ガ》
(to develop) vt. [3] prevention, prophylactics
```

Fig.2 Example of KNP analysis (2)

In Fig. 1, the verb is marked as type [1], whereas in Fig. 2, the verb is marked as type [3]. This number denotes the classification number of the case frame. In both cases, the English translation of the verb is "develop." In such cases, we investigate whether the classification reflects the difference of the translation.

2.3 Case frames

(1) {従業員、運転手、…}が{車、トラック、…}に{荷物、物資}を　積む
{An employee, A driver, …} loads {the luggage, goods, …} onto {the car, the truck, …}.
(2) {選手、従業員}が{経験、体験}を　積む
{A player, An employee} acquires {experience}.

Fig.3 Examples of case frames

Figure 3 shows examples of case frames. In (1), the case structure of a Japanese verb "tumu" is "N1 ga N2 wo N3 ni tumu." (N1 loads N2 onto N3.) Whereas in (2), the same verb "tumu" has a different case structure, namely, "N1 ga N2 wo tumu." (N1 acquires N2.) This difference in structure represents the different meanings of the verb "tumu." Such differences are directly reflected in the English translation, so that, in (1), the verb "tumu" is translated as "load", and, in (2), the same verb is translated as "acquire." The purpose of our research is to investigate whether differences in case structure reflect differences in translation.

2.4 Verbs in a Japanese-English Dictionary

Traditional Japanese verbs, or "wago-dousi", are different from "sahen" verbs (verbs that originated from nouns), in that "sahen" verbs have relatively fewer meanings than "wago-dousi," and their translations into English are also restricted. As such, the classification of translation would be easier. In contrast, "wago-dousi" can have numerous meanings, making them difficult to translate.

3 Classification of translation of patent sentences

3.1 Procedure

We analyze and classify the following four steps:
(1) Input sentences are analyzed using the KNP.
(2) English sentences are compared with Japanese sentences based on the output of the KNP.
(3) Each case frame is classified from the results of the comparison.
(4) The possibility of classification of the translation is investigated.

3.2 Results of analysis

A total of 6,249 Japanese sentences from the patent database are analyzed using the KNP. The corresponding English sentences are derived from the same database (Japio, 2004). Their case structures are obtained from the KNP results.

3.2.1 Classification of "sahen" verbs

Table 1 Classification of translation of the Japanese verb/noun "bunpitu"

Case frames	Example sentences	No. of sentences	English words
Type [1]	細菌株は SAM を培地中へ分泌する (the bacterium strain secretes the SAM in the culture medium)	13	secrete
	細菌は増殖時にコラゲナーゼを分泌する (Bacteria secrete the collagenase in the medium at the time of their proliferation)	1	secret
Type [2]	神経栄養因子の分泌に関わっている (participates in the secretion of neurotrophic factor)	7	secretion
Type [3]	分泌におけるジスフィルド結合の形成 (Formation of a disulfide bond in secretory)	1	secretory

One of "sahen" verbs, "bunpitu", is exemplified in Table 1. A total of 22 sentences are obtained from 6,249 sentences, among which 14 are classified as type [1], seven as type [2], and one as type [3] for case frames.

In Table 1, the column "English words" are shown as the corresponding English words written by human translators. In this case, the different type of verb clearly corresponds to different English words such that the form of the word is "secrete" in the type [1] sentences, "secretion" in the type [2] sentences, and "secretory" in the type [3] sentence. Although the translation in one type [1] sentence corresponds to "secret," this is obviously due to a spelling error. Practically, in the type [1], there is the conjugation like "secrete", "secretes", "secreted", and "secreting."

3.2.2 Classification of "wago-dousi" (traditional Japanese verbs)

Table 2 Case frame type for traditional Japanese verbs

Verbs	Case frame type	No. of case frames
示す ("simesu", show, point, indicate, …)	[4]	167
持つ ("motu", have, take, grasp, own, …)	[14]	125
異なる ("kotonaru", differ, different, unlike, …)	[1]	100

The classification of case frame was generally unsuccessful for "wago-dousi".

Table 2 shows the results for some traditional Japanese verbs. For example, the verb "simesu" has a total of 1,742 different case frame categories. However, all of the 167 sentences from the patent database are classified as type [4] sentences ("mono" ga … wo simesu.(it shows …)). No sentence is classified to other 1,741 categories. As shown in Table 2, these verbs are usually biased to only one and/or two types. The reason for this is that the original classification is not correct, and most of the examples derived from the web database are classified into a few biased categories.

4 Concluding Remarks

We have investigated the effectiveness of case frames for use in machine translation of patent sentences. The results of the present study reveal the possibility that case frames may be useful in machine translation of patent sentences, although various issues remain to be clarified and further research is required.

First, our investigation was performed manually, and then the work was very ineffective. Syntax analysis was performed using the KNP program, but the classification was not automatic. At present, a human must consult a Japanese-English dictionary and find the correspondence between Japanese and English verbs, in order to classify the case frame. We have previously introduced an alignment program between Japanese and English sentences that will accelerate this task.

Second, the lack and/or bias of information in the case frame lexicon is very problematic. Originally, case frames were generated automatically from a vast web database, but the results were poor, especially for traditional Japanese verbs ("wago-dousi").

In the future, we intend to extend the present research and improve the case frame lexicon for use with patent sentences.

Acknowledgments

The idea for this study came about during a discussion in the Special Interest Committee of Patent Translation supported by AAMT and Japio (Chair: Prof. Jun'ichi TSUJII, University of Tokyo). We would like to thank all of the members of the Committee for the useful advice and discussion. In addition, we would like to thank Japio for providing the patent database. The preset study supported in part by a Grant-in-Aid for Scientific Research (18500102) from KAKENHI.

References

Japio (Japan Patent Information Organization). 2004. *Patent database for AAMT/Japio Special Interest Committee for Patent Translation.*

JPO (Japan Patent Office): *Electronic brary*, http://www.ipdl.inpit.go.jp/homepg.ipdl.

D. Kawahara and S. Kurohashi. 2005. Gradual Fertilization of Case Frames (in Japanese), *Journal of Natural Language Processing*, Vol. 12, No. 2: 109-131.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case Frame Compilation from the Web using High-Performance Computing (in Japanese), *IPSG SIC Technical Report*, 2006-NL-171.

KNP. http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html Kyoto University.

MT Summit X, 2005. *Proceedings of Workshop on Patent Translation*, Phuket, Thailand.

MT Summit XI, 2007, *Proceedings of Second Workshop on Patent Translation*, Copenhagen, Denmark.

Masumi Okuyama and Shoichi Yokoyama, 2009. Translation Disambiguation of Patent Sentences using Case Frames (in Japanese), *Tohoku Branch Meeting of Information Processing Society Japan*, B-1-3.

Shoichi Yokoyama and Yuya Kaneda, 2005. Classification of Modified Relationships in Japanese Patent Sentences, in *MT Summit X, 2005*: 16-20.

Shoichi Yokoyama and Shigehiro Kennendai, 2007. Error Correcting System for Analysis of Japanese Patent Sentences, in *MT Summit XI, 2007*: 24-27.