Investigation for Translation Disambiguation of Verbs in Patent Sentences using Word Grouping

Shoichi Yokoyama **Yuichi Takano**
Graduate School of Science and Engineering (Informatics), Yamagata University
4-3-16 Jonan, Yonezawa, Yamagata 992-8510, Japan

yokoyama@yz.yamagata-u.ac.jp thx78502@st.yamagata-u.ac.jp

Abstract

In the automatic translation of complicated patent sentences, one of the issues to improve translation quality is to translate verbs in the source language with various meanings to corresponding different words in the target language correctly. This paper proposes the disambiguation method using the word grouping. Verbs with various meanings usually co-occur with their corresponding nouns, and show the different meanings. Valence and frame structures of verbs were used to resolve such problems. However, the meanings should be dealt with more deeply and appropriately. This paper describes the trial of word grouping based on a thesaurus.

1 Introduction

Most of patent sentences have long, complicated structure in problems, claims, expressions, and details (Yokoyama 2005). If these sentences are input to a machine translation system, their complicated structure causes the insufficient analysis, and then the correct translation cannot be performed. The automatic processing for patent sentences is very important research topic by virtue of several viewpoints such as simultaneous international information exchange and time and cost saving for translation. In addition, the movement of worldwide patent application unification promotes the necessity of machine translation.

However, the quality of the translation is not so good. One of the reasons is the existence of verbs with various meanings. Such verbs should be differently translated into words in the target language. Especially in Japanese, verbs originated in traditional Japanese have many meanings. For example, a Japanese verb “ateru” has many meanings such that a sentence “battoni bo-ruwo ateru” is translated into English sentence “hit a ball with a bat”, “kabeni tewo ateru” into “put hands onto the wall”, and “kuziwo ateru” into “win a lot.”

This paper proposes the grouping of words which co-occur with a verb. In this paper, we investigate how useful the word grouping is for the translation disambiguation of verbs.

2 Related Works

Our group has first classified the modification relation found in patent sentences (Yokoyama 2005), and then constructed the error correcting system prototype for analysis of Japanese patent sentences (Yokoyama 2007). Afterwards we have utilized case frames for disambiguation (Yokoyama 2009, Suzuki 2010). As the result, the case frames of verbs originated from the active nouns are possibly useful to disambiguate the English verbs. However, the case frames of traditional Japanese verbs are problematic. The improvement of the case frame lexicon is necessary for the future work.

Proceedings of the 4th Workshop on Patent Translation, pages 60-63,
MT Summit XIII, Xiamen, China, September 2011

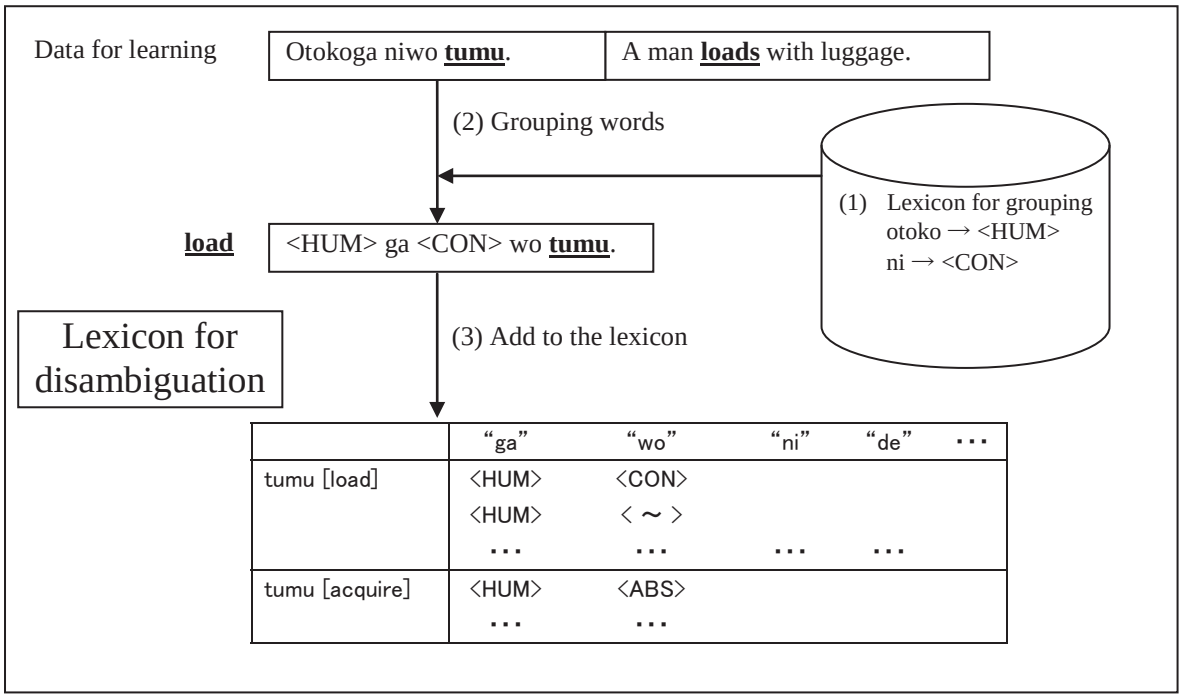

Fig.1 Flowchart of Disambiguation

Tsunakawa et al. proposed the knowledge acquisition for equivalent selection of words using a comparable corpus (Tsunakawa 2011). Unsupervised learning for disambiguation using the data from a lexicon and a corpus is proposed (Ide 1998). In their paper, grammatical information such as part-of-speech, words with syntactic relation, and words relative to the field with coocurrence are used as cues.

Dagan et al. proposed the disambiguation using a mono-lingual corpus in the target language (Dagan 1994). Li constructed the bootstrapping classifier for disambiguation of word translation (Li 2002). Vickrey et al. introduced the selection of translation equivalents considering the context into SMT system (Vickrey 2005). One of the important issues here is to introduce the SMT features into total sentences.

3 Procedure and Investigations

3.1 Procedure

Figure 1 shows the procedure with three steps:

(1) Semantic grouping lexicon is constructed for patent sentences.
(2) Data for learning are divided into predicate argument structures (Iida 2007), and grouping the words are done.
(3) Sentences for grouping and corresponding words are added into the lexicon for disambiguation.

In order to construct the lexicon and confirm its usefulness, we have two investigations. One is the disambiguation of verbs using translation web sites. The other is to confirm the effectiveness of word grouping.

3.2 Investigation 1: Disambiguation in web sites

In order to confirm the efficiency of semantic interpretation to the disambiguation for machine translation, we investigate the efficiency of disambiguation using grouping.

The data are extracted from the claims in Japanese patent application A61B (medicine, veterinary medicine, or hygiene) published in 2004 (Japio 2004). In the database, Japanese sentences and their translation by human are included.

In our investigation, we deal with the Japanese verb "hukumu", which occurs very frequently in

patent sentences. The verb "hukumu" mainly, in one hand, means "include" (as a part of the whole things), while on the other hand, means "contain" (to have things in the something).

We extract randomly 60 sentences which include the verb "hukumu" and in which 30 of the corresponding English verb is "include", the rest 30 is "contain." These 60 sentences are translated using 4 web machine translation sites (MTsites 2011).

	excite	Google	Infoseek	Yahoo!
include	12	16	30	30
contain	18	10	0	0
others	0	4	0	0

Table 1 Translation results for the English corresponding word "include"

	excite	Google	Infoseek	Yahoo!
include	19	14	30	29
contain	9	15	0	0
others	2	1	0	1

Table 2 Translation results for the English corresponding word "contain"

Tables 1 and 2 show the results. Some sites may use the semantic information, while some sites not. Even the site which may consider to disambiguate shows not so high quality. There are some mistakes according to the reference number such as "means11" and "processing circuit32."

3.3 Investigation 2: Grouping

We then investigate the efficiency of disambiguation using grouping the words. 60 sentences mentioned above are selected. Based on one of the famous Japanese thesauri "Nihongo Goi Taikei" (Ikehara 1997), the concepts at the sixth layer are adopted. Nouns which co-occur with the verb "hukumu" are replaced into the concept at the sixth layer by human. In case of compound nouns, only the last noun is left and rewritten. The rewritten sentences are automatically translated on the web sites, and the resulted disambiguation is investigated.

	Japanese	English
(a) Before	<Sihatu musen Syuhasu reiki parusu> wo sorezore hukunda	Each of the first train that includes <a radio frequency excitation pulse>
After	<buturi gen-syo> wo sorezore hukunda	Each contains the <physics>
(b) Before	<yoyaku zyoho> ni hukumareru kensa naiyo	Contents included in the <best inspection>
After	<tisiki > ni hukumareru kensa naiyo	Contents contained in the <knowledge examination>

Table 3 Examples improved by word grouping

	Japanese	English
Before	<seitai kasseina baio seramikkusu huntai> wo hukunda <seitai nai kyusyuseino takositu tai> to,	And <in vivo absorption of porous powder> containing <a bioactive bioceramics>
After	<kozo> wo hukunda <kozo> to,	<Structure> including <the structure>

Table 4 Errored example from correct one

	excite		Google	
	contain	include	contain	include
×→○	1	3	10	9
○→×	2	0	6	4
×→×	20	15	7	6
○→○	7	12	7	11

Table 5 Efficiency of replacement of words

Tables 3-5 show the results. In Table 3(a), the first < > is exchanged into its super-concept, and then, its English translation is replaced from "include" into "contain." Table 3(b) shows an example of passive case. Here, the verb "included" is replaced into "contained" correctly.

However, in Table 4, the correct word "contain" is replaced into the erroneous word "include" because both brackets are replaced into the same super-concepts "structure."

Table 5 shows the total results. "×→○" shows that the erroneous results are improved to the correct results, and so on. Totally, the results are improved.

4 Discussions and Future Works

Based on the results of Experiment 2 mentioned above, we now consider the following procedure:

(1) Nouns occurring in a patent sentence are classified and categorized into the sixth layer of the "Nihongo Goi Taikei" (Ikehara 1997).
(2) Every noun in the sentence is replaced into the sixth-layer category.
(3) Every category is combined to the corresponding verb in the sentence.
(4) The lexicon for disambiguation is made from the results of combination.

In order to avoid the combinatorial explosion, the extent of the combination will be restricted around the corresponding verb.

As mentioned above, we leave only the last noun and replace it in case of compound nouns, the affect must be investigated.

We will be able not only to deal with unknown words and/or terminology, but also to treat phrases and/or idioms.

Acknowledgments

We acknowledge Japio for the patent database, and all members in the Special Interest Committee of Patent Translation supported by AAMT and Japio (Chair: Prof. Jun'ichi Tsujii, Microsoft) for the useful advices and discussions.

References

Ido Dagan and Alon Itai. 1994. Word Sense Disambiguation Using a Second Language Monolingual Corpus. Computational Linguistics, 20(4) pp. 563-594.

Nancy Ide and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation Using Bilingual Comparable Corpora. In *Proc. of the 19th International Conference on Computational Linguistics*, pp. 411-417.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. NAIST Text Corpus: Annotating Predicate Argument and Coreference Relations (in Japanese). *IPSJ SIG Technical Report 2007-NL-177 (10)*.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi (eds.). 1997. Nihongo Goi Taikei (Systematic Thesaurus of Japanese) (in Japanese). Iwanami Shoten, Tokyo.

Japio (Japan Patent Information Organization). 2004. Patent Information Database.

Cong Li and Hang Li. 2002. Word translation disambiguation using bilingual bootstrapping. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 343-351.

MT sites. 2011.
excite translation：(http://www.excite.co.jp/world/)
Google translation
(http://www.google.co.jp/language_tools?hl=ja)
Infoseek multi-translation.
(http://translation.infoseek.co.jp/)
Yahoo! Translation:
(http://honyaku.yahoo.co.jp/)

Kampei Suzuki and Shoichi Yokoyama. 2010. Analysis of Patent Sentences using Case Information of Verbs (in Japanese), In *Proc. of 72nd Annual Meetings of Information Processing Society Japan*, 4W-2.

Ryuji Tsunakawa and Hiroyuki Kaji. 2011. Selection of Translation Equivalents using Syntactic Cooccurrence and Semantic Classes (in Japanese). In *Proc. of 2010 AAMT/Japio Patent Translation Research Group*, pp.25-30.

David Vickrey, Luke Biewald, Marc Teyssier and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proc. of the Conference on HLT/HMNLP*, pp. 771-778

Shoichi Yokoyama and Yuya Kaneda, 2005. Classification of Modified Relationships in Japanese Patent Sentences, *Proceedings of Workshop on Patent Translation*, pp.16-20.

Shoichi Yokoyama and Shigehiro Kennendai, 2007. Error Correcting System for Analysis of Japanese Patent Sentences, *Proceedings of Second Workshop on Patent Translation*, pp.24-27.

Shoichi Yokoyama and Masumi Okuyama, 2009. Translation Disambiguation of Patent Sentence using Case Frames, *Proceedings of Third Workshop on Patent Translation*.