Toshiba Rule-Based Machine Translation System at NTCIR-7 PAT MT

Tatsuya Izuha Akira Kumano Yuka Kuroda
Knowledge Media Laboratory, Toshiba Corporate R&D Center
tatsuya.izuha@toshiba.co.jp

Abstract

Toshiba (tsbmt) participated in the Patent Translation Task at NTCIR-7. We submitted two runs for Japanese-English intrinsic evaluation, one run for English-Japanese intrinsic evaluation and one run for extrinsic evaluation. The machine translation system used for those runs is rule-based one developed for translating open-domain written texts. A technical term dictionary for patent domain is used for all the runs as well as the common word dictionary. In addition, one of the two runs for Japanese-English intrinsic evaluation uses a dictionary built semi-automatically from the training data. Although it is not fair to compare translation quality between our system and purely statistical ones since the former uses extra knowledge (hand-crafted dictionary entries and rules) for all the runs, we believe that we have contributed to the research community by providing useful data. This paper describes the overview of our machine translation system.

Keywords: *rule-based machine translation.*

1. Introduction

Toshiba (tsbmt) participated in the Patent Translation Task[1]. We submitted two runs for Japanese-English intrinsic evaluation (tsbmt-je1, tsbmt-je2), one run for English-Japanese intrinsic evaluation (tsbmt-ej1) and one run for extrinsic evaluation (tsbmt-ex1). The machine translation system used for those runs is rule-based one developed for translating open-domain written texts.

Section 2 describes the overview of the machine translation system. Tuning done for the Patent Translation Task is described in Section 3. Section 4 concludes this paper.

2. Overview of Toshiba Rule-Based Machine Translation System

Our system is transfer-based as well as rule-based. While the core framework has been well-established for decades, R&D targets in recent years include introduction of statistical approach[2], technique for specific domains[3], utilization of contextual information[4] and mutilingualization[5]. Since the system used in the Patent Translation Task only uses the core framework, it is described in this section.

2.1. Morphological Analysis

First, input text to MT system is segmented into words and phrases converting each of them into the infinitive form if inflected. The grammatical categories (parts of speech) are attached to all words and idioms after the retrieval of lexical dictionaries. Since Japanese has no spaces between words, there may exist morphological ambiguity at the time of segmentation.

2.2. Syntactic Analysis

Our system parses sentences using the analysis grammar in ATN (Augmented Translation Network Grammar) based on the context-free grammar. This parser is reinforced with a look-ahead function [6]. Some of the rules in the ATN grammar for English sentence parsing are enumerated in the Figure 1.

[1] S → NP, VP;
[2] NP → det, adj, noun, PREP;
[3] VP → vi, PREP;
[4] VP → vt, NP, PREP;
[5] PREP → pre, NP;

Figure.1 sample ATN grammar rules for English

While analyzing a sentence, the parser constructs its internal structure, connecting concept units by arcs with syntactic relation labels. The English sentence (1) is analyzed to have a tree structure in Figure. 2 where each concept unit is illustrated by the oval with a source won inside.

(1) This machine translates an English sentence into Japanese.

Since a prepositional phrase can modify a noun phrase as well as a verb phrase, another parsing tree in Figure. 3 could be drawn from the sentence (1), although it is unnatural. In parsing a sentence, ambiguity of this

kind often occurs and it is resolved using hand-crafted rules.

2.3. Semantic Analysis

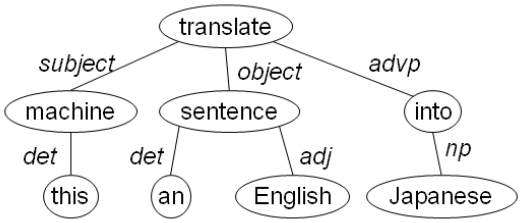

Figure.2 parse tree derived from sentence (1)

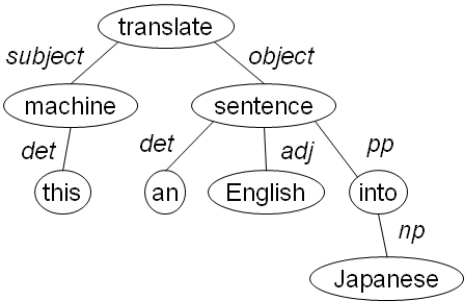

Figure .3 another parse tree which can be derived from sentence (1)

Syntactic analysis is not enough for real MT systems to extract a correct meaning; it is semantic analysis that maps a purely syntactic structure onto a conceptual one. In fact, many practical MT systems use semantic information to translate a natural language sentence accurately. It helps sentence analysis choose the correct parsing tree. It is one way to use semantic information during parsing phase, and it sometimes makes selecting a proper tree easier. But the amalgation of processing of syntactic analysis and semantic analysis results in a complicated system, making the analysis grammar much difficult to maintain.

In our system, the semantic analysis phase is clearly separated from the syntactic analysis. We assume the hypothesis that the meaning is lexical. Thus, the syntactic analysis derives a purely syntactic structure, simplifying the analysis grammar and accelerating the parsing speed. Semantic analysis is invoked by lexical rules on the output tree of syntactic analysis.

(2) He is able to swim fast.
(3) He is happy to swim fast.

Both sentences above have the identical part-of-speech sequence and create syntactically the same structures as illustrated in Figure. 4 (1)a and Figire. 5 (2)a. Yet each tree is converted into different conceptual structures after the semantic analysis.

Figure 4 (1)b shows the descriptive verb "swim" included information for the modality "can", so the whole meaning of the sentence is identical to "He can swim fast." In Figure. 4 (2)b "swim" plays a role of reason for making him happy. Those transfers are based on the lexical rules. Here the lexical rules of the adjectives "able" and "happy" act on the semantic analysis.

The semantic analysis by the lexical rules determined the preference of many interpretation possibilities of the output tree structure built by the syntactic analysis with no semantic grammar. The combination of these two processes, or the "Lexical Translation Network Grammar" (LTNG) as we call it, realizes the accurate interpretation.

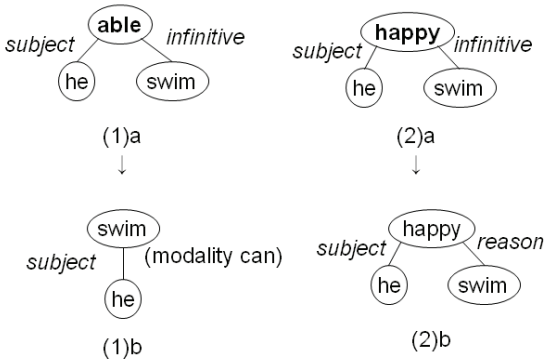

Figure.4 semantic analysis based on lexical rules of adjective "able" and "happy"

2.4. Selection of Translation Words

The quality of translation depends partly on the accuracy of syntactic and semantic analysis, but the selection of translation word (TW) for each concept also reflects the acceptability of the output sentence of MT.

In the majority of language, including Japanese and English, basic verbs generally have a variety of usages, and each usage might need a different TW. Consequently, the selection of the proper TW is one of the important functions of the semantic analysis.

The meaning of a Japanese transitive verb "かける"(kake-ru) ranges wide. Some of its usages and their English translations are listed below:

[i] フックに帽子をかける。
to hang one's cap on the hook
[ii] モーターに電圧をかける。
to apply voltage to a moter
[iii] 椅子に腰をかける。
to sit on a chair
[iv] 部屋に鍵をかける。
to lock the room

Each English TW for "かける" should be properly selected because the representations in Japanese are alike. This kind of TW selection is realized by the tree conversion invoked by the lexical transfer rules. In the above, the Japanese lexical entry "かける" has at least four different transfer rules, each producing the correct English conceptual structure. In Figure.5 where the figures of the four tree conversions are given, SW

denotes the source word in the source language, and TW denotes the translation word in the target language.

In this phase, the semantic markers are ready to play an important role. Consider selecting Japanese TW for "take" in the following examples:

[i] take a bus　バスに乗る
[ii] take a taxi　タクシーに乗る
[iii] take a train　列車に乗る
[iv] take a bath　風呂に入る
[v] take a medicine　薬を飲む

TW selections for [i] through [iii] are enabled by a separate lexical rule to pick the identical TW "乗る", but many similar rules are also needed for the reasonable TW selection for other vehicles like airplanes, boats, bicycles, and so on, because TW of "take" is always "乗る" when the object is a noun representing a vehicle. Such being the case, when a new vehicle word is added to the MT dictionary, another TW selection rule stipulating "take" select "乗る" must be added. This method, however, does not seem efficient. To reduce the rule description, the semantic marker is used in the lexical rules. Vehicle words, such as "bus", "train", "airplane", "boat", are attached with the semantic marker "vehicle", and many rules in "take" to select "乗る" are merged into one rule like Figure. 6. In this figure, SM means the semantic marker.

Therefore, dictionary rules are kept economical, thereby facilitating the maintenance of dictionaries.

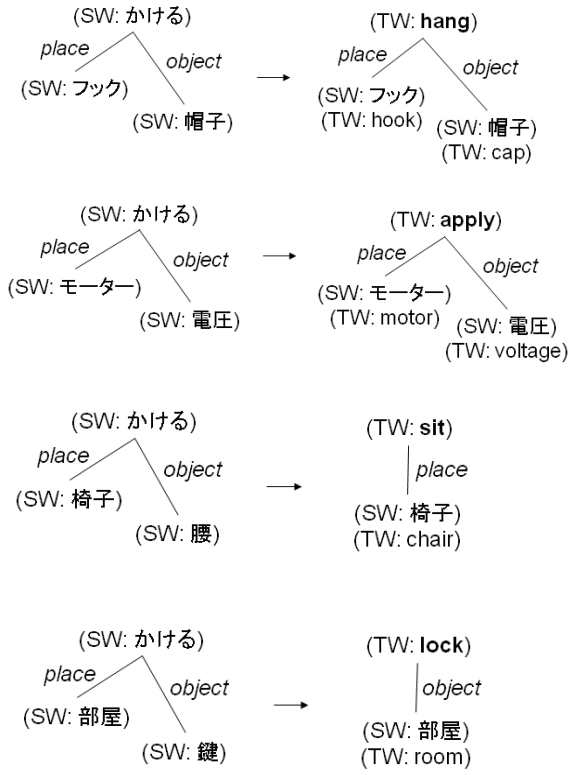

Figure.5 lexical transfer rules corresponding to four different meanings of a Japanese basic verb "かける"

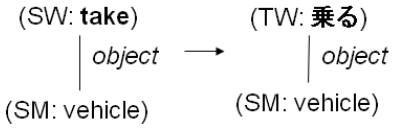

Figure.6 lexical transfer rule referring to a semantic marker "vehicle"

2.5. Structural Transfer

While the lexical transfer phase acts to make the difference small between English and Japanese conceptual structures, it is also true that better translation will be generated independent of lexical information.

To prove this, let us adduce an example. In comparison with Japanese sentences, English sentences can often be in the passive voice. Although passive English sentences are usually mapped to grammatically correct Japanese passive sentences, the native Japanese speaker still feels that some of them are a little awkward. They are not incorrect, but a better representation surely exists. In many cases, representation in the active voice is rather natural. When the condition to convert the passive voice into the active is obvious, the rule for the conversion is encoded in the structural transfer rules, enabling the lexical entry-free conceptual structure transfer.

2.6. Syntactic Generation

In the generation phase, word order is the CD structure is determined to form a string of words in the target language. Since the word order of Japanese greatly differs from that of English, the Japanese syntactic grammar is required to generate a correct sentence. Figure.7 gives sample rules of simple Japanese syntactic generation grammar.

[1] S → SUBJ, OBJ, POSTP, verb;
[2] SUBJ → $NP, "は";
[3] OBJ → $NP, (casemarker|"を");
[4] POSTP → $NP, postp;
[5] $NP → ADJ, noun;
[6] ADJ → adjective;

Figure.7 sample rules of Japanese syntactic generation grammar

[1] shows that, in a simple Japanese sentence, the subject phrase appers first, the object phrase next, the postpositional phrase follows and a descriptive verb is places at the end of a sentence. [2] indicates that the subject phrase is a noun phrase ($NP) followed by the postposition "は". [3] shows that the object phrase is a noun phrase followed by a case marker, or the postposition "を" if the case marker is not specified. [4] shows that the postpositional phrase is a noun phrase followed by some postposition, while [5] shows the

noun phrase consists of a pre-modifying adjective and a head noun. Finally, [6] says an adjective comes below the ADJ arc in the Japanese conceptual structure.

In Japanese, the restriction on the order of phrases in a sentence is loose except that verb must come at the end of a sentence. The subject phrase is usually located near the beginning of a sentence, but could follow the object phrase or could immediately precede the last verb. There may be a case where a sentence with the subject phrase at the beginning is natural. Accordingly, determination of word order often affects the quality of translation. However it is one of the knotty problems in generating a Japanese sentence.

2.7. Morphological Generation

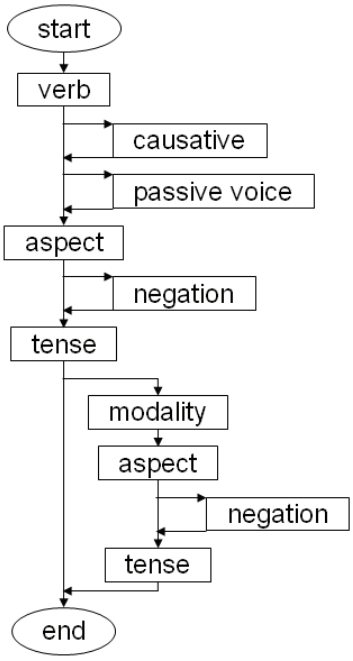

Figure.8 generation flow for Japanese verbs

In the final phase of machine translation, a fixed form must be given to each inflectional word. Unlike English, a Japanese fixed form consists of a stem and agglutinative morphemes, each inflecting in order.

TW information in a concept unit represents an infinite form and an inflection type. Morphological and syntactic features of verbs and adjectives, i.e. negation, tense, aspect, modality, voice and so on, are stored in each concept unit at this point. Such features are generated as Japanese inflective morphemes, which are encoded in a table again as an infinite form and an inflection type. The agglutinating order may change depending on the combination of the features[7]. In a simple case, the generation flow is like Figure.8.

For example, the English verb phrase "could not read" is analyzed to be made up of concept unit [read (verb): "読/む(inflection type = 5-dan) + (modality = "can") (tense = "past") (type = "negative")]. According to the flow in Figure.8, the fixed form is generated like the following:

verb → 読/む
modality = "can" → ことができ/
type = "negative" → な/かっ
tense = "past" → /た

In the example above, the slash divides each Japanese morpheme into the stem and the inflectional part.

Note that the negation morpheme and the tense morpheme are generated after modality, which could be generated before it. The combination of those feature determines the sequence by referring to the characteristics of a Japanese representation. The result string for the verb phrase is "読むことができなかった".

2.8. Dictionaries

Our MT system uses three levels of dictionaries: a common word dictionary, a technical-term dictionary and a user-defined dictionary [8].

The common word dictionary includes about one million words for both English-Japanese and Japanese-English translation, which is generally used independent of document fields. In general, common words have diverse usages and several meanings corresponding to them. TW's might be different when the meaning differs. Since the knowledge for TW selection is encoded as lexical transfer rules as described in 2.4, the total amount of description in the common word dictionary is relatively large.

The technical term dictionary includes domain-specific technical terms. There are 28 domains such as computer, machinery and medicine. Total size is more than 2,600,000 words for each translation direction, and many of those words are compound words. Technical terms tend to have only one meaning within a domain. It results in the relatively simple structure of the dictionary.

Even with the two types of dictionaries above, the MT system cannot always output the proper translation. In the user's documents, there may exist some words unknown to the MT dictionaries. To accomplish the output of machine translation, the user should teach the system the information of unknown words, at least the grammatical features and the TW, and the system should automatically learn to store the knowledge in the user-defined dictionary.

3. Tuning for the Patent Translation Task

3.1. Patent Dictionary

There is a patent domain dictionary among the technical term dictionaries described in 2.8. Its size including both translation directions is 210,000 words. In addition to the common word dictionary, the patent domain dictionary is used for all the runs Toshiba submitted.

3.2. Fine Tuning of Translation of Frequent Expressions

We have modified translation of several expressions which appear frequently in the patent domain according to the training data. For example, the translation of "図 N" ("N" stands for any number) has been modified from "Drawing N" to "Fig. N". Another example is that the translation of "実施例" has been modified from "example" to "embodiment". Such fine tuning has been done for only Japanese-English translation (tsbmt-je1 and tsbmt-je2). Although it will improve the BLEU with single-reference, it will have little effect on other evaluations.

3.3. Semi-automatic Bilingual Term Extraction from Training Data

In one of the two runs of Japanese-English intrinsic evaluation (tsbmt-je2), a bilingual term dictionary built semi-automatically from the training data (PSD-1) is used. The method proposed by Kumano et al.[9][10] is adopted for candidate extraction. Among bilingual term candidates with the confidence score higher than 90%, about 80,000 bilingual terms were registered into the user-defined dictionary described in 2.8 after rough manual checking. But the effect on BLEU is very small. Close analysis should be done in the near future.

4. Conclusion

The overview of Toshiba (tsbmt) rule-based machine translation system used in the Patent Translation Task is described in this paper. Although it is not fair to compare translation quality between our system and purely statistical ones since the former used extra knowledge (hand-crafted dictionary entries and rules), we believe that we have contributed to the research community by providing useful data.

References

[1] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, 2008.

[2] Hirokazu Suzuki, Akira Kumano. Learning Translations from Monolingual Corpora, Proceedings of MT Summit X, 2005

[3] Kenji Ono. Translation of news headlines, Proceedings of MT Summit IX, 2003

[4] 鈴木博和. 文書全体の情報の利用による機械翻訳の高精度化, 第 5 回情報科学技術フォーラム(FIT2006)講演論文集, 2006

[5] Tatsuya Izuha, Kumano Akira. Chinese-to-Japanese / Japanese-to-Chinese Machine Translation System, Toshiba Review Vol.62, No.4, 2007

[6] Hiroyasu Nogami, Yumiko Yoshimura, Sinya Amano. Parsing with Look-ahead in Real-time On-line Machine Translation System. Proceedings of COLING-88, 1988

[7] Sinya Amano, Hideki Hirakawa, Hiroyasu Nogami and Akira Kumano. The Toshiba Machine Translation System. *Future Computing Systems Vol.2, No.3*, 1989

[8] Seiji Miike. AS-TRANSAC Bilingual Dictionaries. *Proceedings of International Symposium on Electric Dictionary*, 1988

[9] Akira Kumano, Hideki Hirakawa. Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information. Proceedings of COLING-94, 1994

[10] 熊野明. カタカナ表記からの英訳推定による専門用語辞書作成, 言語処理学会第 1 回年次大会論文集, 1995