Processing of Proper Nouns and Use of Estimated Subject Area for Web Page Translation

Yumiko Yoshimura, Satoshi Kinoshita and Miwako Shimazu

Research & Development Center, Toshiba Corp.
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 210 Japan
{yumiko,kino,miwako}@eel.rdc.toshiba.co.jp

Abstract. In Japan, in response to the recent increase of the general public's interest in the Internet, English-to-Japanese machine translation systems are attracting considerable attention as convenient tools to read English Web pages, which are characterized by wide-ranging contents and high frequency of proper nouns. This paper describes a program for finding proper nouns and the subject area of a text as it reads on and uses the information acquired for correct parsing and target word selection, in order to improve translations of various Web pages.

1 Introduction

In Japan, in response to the recent increase of the general public's interest in the Internet, English-to-Japanese machine translation (MT) systems, originally developed for technical translation to produce written documents, are attracting considerable attention as a convenient tool to read English Web pages. As such, there has been a change in MT users; many ordinary people without language expertise increasingly use the system rather than professional translators. Last year, we here at Toshiba introduced an MT system "ASTRANSAC for Internet," onto the market, one of the developments of our original MT system.

Compared with technical documents, general texts on the Web have the following characteristics :

- Textual characteristics
  - A wide range of topics which cannot be covered by technical dictionaries alone.
  - High frequency of proper nouns and newly created words.
- Characteristics related to user operations
  - Users rarely add knowledge to MT systems by registering new words into their user dictionaries or correcting target words.
  - Most users do not make adjustments to the translation environment (e.g. setup of technical dictionaries and preediting of source texts) each time they netsurf different sites because of the burden, although adjustments are necessary to get the best translation.

These characteristics suggest that the current MT systems should be more robust. First, proper nouns need special treatment because often they serve as key terms in conveying the message expressed in the source language. Second, we need to devise an alternative to the conventional approach where knowledge acquisition is insufficient in view of syntactic and semantic flexibility, and some of the knowledge cannot be represented in a simple form.

According to one discourse model, on their first reading of new texts, people generally grasp the message from a set of words used and their relations as they read on, without having specific senses of words in mind. Following this model, we developed a program for finding proper nouns and the subject area of a text while reading it so that the information thus acquired can be used for correct parsing and target word selection. We then conducted experiments to evaluate its feasibility. This paper first describes the proposed program; then it presents the experimental results and the observations made as well as possible applications as part of the future work, followed by some concluding remarks.

2 Processing of Proper Nouns

2.1 Translation of proper nouns

In the development of natural language processing technology, in particular, information retrieval and message understanding, correct recognition of proper nouns has received most attention as the past literature reveals ([Wakao et al., 1996], [Strzalkowski & Wang, 1996], and [Mani, 1996]). However, problems inherent in English-to-Japanese translation are relatively unexplored.

Web pages are characterized by high frequency of proper nouns of many different types, including names of persons, organizations, products, and places. Because Web pages are expected to keep the public posted with up-to-date news, the set of proper nouns on these pages is not closed and new names are constantly being added to the public's lexicon. By contrast, the capacity of MT dictionaries is limited primarily because it is not possible to predict the formation of new unseen proper nouns.

The major problem with new proper nouns lies in the difficulty of distinguishing true proper nouns from those non-proper nouns which are capitalized for emphasis. In the case of non-proper nouns, we need to convert source language words into lower case and then look them up in dictionaries to give the corresponding target language words. The first case is more complex. Here, we are allowed to output the original source language words as they are. But for the benefit of readers it is preferable to translate those words into target language, which indicate the meaning of proper nouns, as in (Ex.1). Compare this with (Exs.2-4) which should not be translated. In these examples, we find the phonetic equivalents instead. That is, "Apple" does not mean apples and the same is true of "Gene" and "Glazer." This is typical of nouns indicating names of persons, companies, and places. "Gene Glazer" can be identified as a person's name from the preceding word. On the second appearance of the last name alone, humans can recognize it is a person's name. Likewise, if the term "Apple Computer" is found in previous sentences, the word "Apple" in the sentences that follow is not taken as a name of fruit, while with the appearance of the "Prudential Insurance Co." humans would not judge the word "Prudential" in (Ex.4) as an adjective. Note that in machine translation, the analysis of the word "Prudential" as an adjective phrase would result in the parsing failure of the whole sentence.

(Ex.1) E : This is Beijing Automation Technology Research Institute.
J : これは北京オートメーション技術研究所である。
(Ex.2) E : "Apple has to do something," said Gene Glazer.
J : 「りんごは何かを行わなければならない。」と遺伝子つや付け工が言った。
(りんご= a fruit name, 遺伝子 = "gene", つや付け工 = "glazer")
(Ex.3) E : In the joint briefing Tuesday, Glazer unveiled . . .
(Ex.4) E : Prudential says it will . . .

Furthermore, if "Apple Computer" and "Gene Glazer" are entered in the dictionary, but not "Apple" and "Glazer", the translations of "Apple" and "Glazer" appearing in subsequent sentences should be generated by referring to the dictionary translations of "Apple Computer" and "Gene Glazer" and automatically extracting the corresponding part from them. Giving different translations to the word "Apple" in "Apple Computer" and "Apple" alone would result in general users' misinterpretation of the original English.

In conclusion, without correct analysis and controlled translation of proper nouns, MT systems are bound to fail in parsing and generate translations which have lost much of the meaning expressed in the source language text. Hence the need arises to develop a program which estimates from context whether a capitalized word, or a string of capitalized words, is a proper noun, and then inherits the result to the translations of the sentences that follow. Additionally, correct analysis of proper nouns improves the translation of co-occurring words.

2.2 Estimating proper nouns and processing part of proper nouns

Current estimation of proper nouns is straightforward. For each sequence of more than two words beginning with a capitalized letter, the system determines whether the capitalized letter should be converted into a lower letter for dictionary look-up and then judges whether it denotes a person, company, or place. The key items for the judgement are:
(1) approximately 70 types of key elements of proper nouns (eg. "Corp.", "Dr.", "Island")
(2) 8 types of a string of neighboring words which trigger proper nouns (eg. ", who/whose")

The next step is to process part of proper nouns. Normally, names of a person and a company first appear in texts in their official names, that is, fully spelled and unabbreviated; then, on their second appearance and thereafter, only part of the name is used. Based on this observation, from text we extracted key elements, including the estimated ones, of proper nouns which are made up of multiple words, in addition to this dictionary translation, so that they may be referred again in the processing of subsequent sentences. Thus, even if the word "Apple" itself does not have a company name entry, the interpretation as a company name is given priority on the ground that the term "Apple Computer," which is entered in the dictionary as a company name, has appeared in previous context. Next, the dictionary translation of "Apple Computer" is morphologically analyzed. If the dictionary translation of "Apple Computer" is "アップル・コンピュータ," then the part corresponding to "Apple," namely "アップル" will be generated as the correct translation of "Apple" in this particular context. Likewise, suppose "Prudential Insurance Co." is found in previous sentences but is not entered in the dictionary as a company name. In this case, the morpheme "Co." triggers the estimation of proper nouns and "Prudential" gets the interpretation as a company name as the first priority. Then, both "Prudential Insurance Co." and "Prudential" will be the output with no changes made. The same applies to "Glazer." On the contrary, if in preceding context words in lower case, i.e. "apples" and "prudential," are found but not in capital, they are also saved as non-proper words so that they may be referred back in the processing of subsequent sentences to generate translations from uncapitalized words.

3 Selection of Target Words Using the Estimated Subject Area

3.1 Translation of field-specific terms

In a specific field or subject area, a source language word which has several different target language words could be translated unambiguously. For example, in case of the two polysemous words "base" and "conviction," contexts are useful in selecting the appropriate sense from many possibilities. On the other hand, the non-polysemous English word "administrator" has different Japanese equivalents depending upon subject areas like politics, business, law and computer.

- "base"
  - in military: a fortified center of operations; a supply center for a large force of military personnel. (J: "基地")
  - in baseball: any one of the four corners; infield, marked by a bay or plate. (J: "ベース")
- "conviction"
  - in law: the judgement that a person is guilty of a crime as charged (J: "有罪判決")
  - in general: the state of being convinced; the act or process of convincing. (J: "確信")

This means that even if syntactic relations with other words or selectional restrictions are not sufficient to determine the specific sense of words, the correct target language word can be identified by its subject area. With this in mind, we developed a new framework using information about subject area to select the appropriate target language word.

On word sense disambiguation, previous research has used several different types of knowledge. This includes neural networks ([Collier, 1996]), although the scaleability of such systems may be limited and the knowledge-base is hard to manage and expand. Other paradigms make use of (semi-)automatic knowledge acquisition from sources such as dictionary definitions ([Guthrie et al., 1991] and [Cowie, 1992]), encyclopedia descriptions ([Yarowsky, 1992]), and bi/monolingual corpora ([Dagan & Itai, 1994] and [Gale et al., 1992]). The use of corpora we think merits particular attention due to the wide availability of their knowledge sources). What we are particularly concerned with, however, is not the method of acquisition itself, but the application of resulting knowledge, such as incorporating newly acquired knowledge into the existing knowledge. This becomes very important when one is trying to further improve the quality of commercial MT systems which already have a store of knowledge required to produce fairly satisfactory translation. Accordingly, we placed an emphasis on the ease of building and managing knowledge and practicality, while keeping the process of determining a target word as simple as possible.

The following section describes our method of target word selection, which consists of the following two processes: (1) estimating subject area and (2) determining a target language word based on the estimated subject area.

3.2 Estimating subject area

Definition of subject area First of all, one subject area may be further divided into smaller groups, as there are many types of sports, like tennis and baseball. Second, subject areas are hierarchically structured to represent hyponymy by connecting hyponyms (tennis, baseball, etc.) with their superordinates (sports).

Sources of information for estimating subject area For information sources for estimating subject area, we refer to the following three types of information.

(1) Subject labels assigned to entry words in translation dictionaries: If a word is judged to be a unique term in a special subject area, the corresponding subject label is attached to the word as part of its lexical features. One word may have more than one subject labels.

`(Ex.5) immune cell; noun;(TW 免疫細胞)(fld medicine biology)`
(where TW is a target language word, and (fld xxxx) indicates a subject label.)

(2) Dictionaries for estimating subject area: They contain a set of words which individually cannot define the subject area (like *(1)* in the above) but give a good estimate of a specific subject area when they appear together within one sentence or in local places. To illustrate, such words as "pitcher," "catcher," and "dugout" are closely associated with the subject baseball.

(3) Transfer rules: After parsing a sentence, information about subject area can be attached in the form of transfer rules based on the syntactic features of a sentence and lexical collocation. For example, the words "start" and "proceeding" *per se* do not specify a specific subject area, but the phrase "to start proceedings against him" can pinpoint to one meaning, namely "to start to take legal action against him," which suggests that it relates to law topics. The example below illustrates this type of rule description. With this rule, the system will start translating subsequent sentences using the fact that the text deals with law, once the phrase "start proceedings against him" appears in the text.

```
(Ex.6) MP:0(object_1(prep_2)) / TP:#
       [0.sw="start|take":1.number="plural":2.sw="against|for":]
       [append($fld;"legal");]
```

(where MP = matching pattern, TP = target pattern, "TP:#" means no structural transfer)

We designed our subject hierarchy after analyzing uncategorized general news articles from two news agencies for three months. Simultaneously we manually compiled information for estimating subject areas and attached it to each target word, so that the correct target word be generated by the strategies to be described in the next section. This method proves extremely easy to expand the hierarchy of knowledge since the hyponymous or superordinate relationship was the only requirement to be taken into consideration.

Estimation procedures Subject area is determined after morphological analysis, using the sources of information for estimation given in *(1)* and *(2)* above. Here, correspondence is made between those words in the sentence and those words in *(1)* and *(2)*. Then, a maximum of three subject labels are stored under the following conditions.

- **Rank:** The deeper the level of subject area the higher the priority.
- **Number of appearances:** For labels with the same rank, those with a greater number of appearance will have a higher priority.

3.3 Determining a target word using information about subject area

To determine a target word according to the subject area, two strategies are available:

1. Formulate transfer rules in which the estimated subject area is given as one of the conditions for rule application.
2. Attach information about subject area to translations specified by transfer rules, or default translations, which are used when no transfer rule is applied; adopt the corresponding translations when subject area has been estimated.

1. in the above is in the form of conventional transfer rules and enables more minute, reinforced distinction by combining conditions on subject area with conditions on syntax and grammatical attributes. (Ex.7) below gives an example of a transfer rule the system needs to translate the phrase "miss the call" as "判定を誤る," roughly meaning "to misjudge," after estimating the subject area of the whole text, which in this case is sports.

```
(Ex.7) MP:0(object_1) / TP:0(object_1)
       [$fld>"sports":0.sw="miss":]
       [set(1;(TW;"判定"));set(0;(TW;"誤る"));]
```

For 2., see (Exs.8-9). (Ex.8) illustrates a word which contains default translations attached with subject labels. In the conventional method, the translation for the noun "round" would always be "丸" unless the system is provided with information to select "ラウンド" over "丸" by users' selection of target words or other possible user operations. By contrast, in our proposed method of target word selection, the word "ラウンド" is selected when the subject area has been determined to be sport. This in turn implies that even when such hyponyms as baseball and boxing are selected as subject area, the word "round" would still be translated as "ラウンド," since the information about subject area is hierarchically defined.

Compare (Ex.8) with the word "steal" in (Ex.9), where the word "盗塁" is selected only when the subject area is judged to be "baseball." If other types of sports like tennis or its superordinate "sports" are estimated to be the subject area, then the output target word would be the same as before, since the subject label of "盗塁"is incompatible with the estimated subject area.

```
(Ex.8) round; noun;(TW 丸 ラウンド (fld=sports));
(Ex.9) steal; noun;(TW 盗み 盗品 盗塁(fld=baseball));
```

3.4 Translation using co-occurring words within one sentence

There are cases where we can determine the topic on the basis of word co-occurrence within one subject area. Consider the word "plane" in (Ex.10), a sentence reporting a plane crash accident. Among the many different meanings of the word, including an airplane, a flat surface, and a type or level, we can determine that in this sentence the word means an airplane from the co-occurring word "takeoff."

If we assign a subject area for each topic, we would end up with an increasing number of subject areas. To incorporate topic-specific characteristics while avoiding this problem, we introduced new transfer rules which give the most suitable translation by referring to words which collocate within one sentence. As (Ex.11), which presents a typical rule description of the noun "plane", shows, the rule does not set intermediate semantic codes. In our system, this type of non-syntactic knowledge is written in the same framework as the existing lexical transfer rules. This enables us to adjust the preference of knowledge flexibly and at the same time to specify the syntactic constraints on knowledge application.

```
(Ex.10) The plane, which had trouble gaining altitude after takeoff, ...
(Ex.11) MP:1 / TP:#
        [1:1.co-occur>"airplane|crash|flight|fly|hijack|jet|pilot|takeoff":]
        [set(1;(TW;"飛行機 航空機"));]
```

In the current system, the scope of search for collocational words listed next to the notation "co-occur" is intrasentential. For the maximum effect we should widen the scope to incorporate a paragraph, an article, or a page.

4 Evaluation

To measure the effectiveness of the methods we have presented, we machine-translated on-line news articles which appeared in Web texts and studied the difference the application of our proposed processing would make to the output translations. Note that to ensure the reliability of our results the test data here is taken from entirely different news articles from the ones we used for our analysis of subject-dependent knowledge.

Below are the news types of our test text, its size, and the experimental environment:

- Text type
  - types of news: top news, business, entertainment, and sports
  - total number of words: 4,410 words (630 proper nouns, 1,959 other content words)
- Dictionary size
  - common word dictionary: 194,099 entry words
  - proper noun dictionary: 33,323 entry words
- Number of subject areas
  - number of top-level subject areas: 14
  - total number of subject area of all levels: 41
- Size of vocabulary attached with information related to subject area
  - number of words attached with key information in determining subject area: 7,196 words
  - number of words in dictionaries for estimating subject area: 733 words
  - number of words which have subject- or topic-specific knowledge of translation rules for specifying translations : 897 words

The important characteristic of headlines is that only part of a proper noun is generally used before giving it in full in the following text body; therefore, in our experiment, we switched the translation procedure so that the information acquired in the body can be used for better analysis of headlines.

Under the above environment, we checked the resulting differences in translation quality between the previous and new systems and obtained the following data shown in Table 1. The number given under "Incorrect" shows the number of incorrect Japanese words which were generated because the given knowledge of the previous system failed to resolve lexical ambiguities. There are two cases where estimating proper nouns contributed to improved translation of co-occurring words. That is, out of 26 counted as improved, the number of improved proper nouns alone is 24.

Number of translation changes			Rate of improvement			
Kind of method	Improved	Worse	Kind of words	Incorrect(total)	[Improved]-[Worse]	Rate
Subject areas	47	3	Proper nouns	96	20	20.8%
Proper nouns	26	4	Other content words	383	46	12.0%
Total	73	7	Total	479	66	13.8%

Table 1. Changes in translations and the rate of improvement

Finally we will present two examples which showed an improvement in terms of word selection. (Ex.12) is a news headline. Here the system correctly recognizes the word "Shields" is part of a person's name and translates as such, from the fact that "Brooke Shields" appeares in the main text. In (Ex.13) the system correctly estimates that it is a news article on baseball and reflects this on the translation, giving a baseball term "リリーフ投手," meaning a relief pitcher. This contrasts with the previous translation, where the word "救済者" does not have a baseball meaning.

(Ex.12) E : Shields: Maintaining Mystery
J (old) : 保護物：謎を保つ
J (new) : シールズ：謎を保つ

(Ex.13) E: . . . the Chicago Cubs have released former All-Star reliever Doug Jones.
J (old) : . . . シカゴ・カブスは前のオールスターの救済者ダグ・ジョーンズを解放した。
J (new) : . . . シカゴ・カブスは前のオールスターのリリーフ投手ダグ・ジョーンズを解放した。

```
(Ex.14) relievet; noun;(TW 救済者 救済物 リリーフ投手(fld=baseball))
```

5 Discussion

In our experiment, where 0.4% of all the vocabulary in the dictionary totaling 227,422 words has knowledge for determining translations on the basis of subject area, we succeeded in eliminating the errors in translations for content words, excluding proper nouns, by 12.0%, after introducing the proposed method. Considering that the above experiment was conducted with an extremely limited knowledge, we can expect that, by adding more target language words attached with this kind of information as well as knowledge for specifying target words, this method would prove more effective in reflecting the diversity of text on the output translation. With this method, managing and expanding knowledge would be much easier due to its simplified knowledge structure. This in turn enables us to automatically extract information useful for translation from corpora, such as words and expressions frequently appearing in each subject area and co-occurring words frequently used within one news article (which normally sticks to one topic).

In the remaining section, let us examine the cases counted as worse as a result of using information about subject area. One case is where the application of information about subject area was not strict enough. This can be resolved by setting the necessary constraints, usually syntactic ones. The remaining two cases involve two or more different subjects, which makes it difficult to decide precisely which subject should be given priority for each word. Here the system selected the target word "役員" ("company manager"), which is strongly associated with economics, because the article was about economics. But as it turned out, its content had to do with government policy and therefore the translation "高官" ("government officer"), which has stronger associations with politics, was actually preferable.

(Ex.15) E : The dollar rose against the mark Tuesday but lost ground against the yen after surging earlier in the day on a call by a senior Japanese official for a weaker Japanese currency.
J : (old) . . . 日本の高官による. . . / (new) . . . 日本の役員による. . .

As for processing of proper nouns, we succeeded in eliminating 20.8% of the errors for all proper nouns after introducing the proposed processing. Most typically, this has contributed to eliminating misanalysis of part of speech as well as syntax, both of which often occur when names of persons or organizations given in full appear later in the text. In two of four cases where the translation quality fell, translations of the corresponding word in lower case are preferable even though they are proper nouns. This is the same problem seen in (Ex.1).

6 Conclusion

We demonstrated a new approach to improving translation of Web text, in particular the accuracy of translations, and numerically proved its feasibility by evaluation of experiments. Since target word selection using information about subject area has shown a significant effect, more improvements will be made on the present specifications. The next task would be to expand the range of subjects and refine them, and at the same time elaborate knowledge for target word selection using subject area information. For this purpose, we are planning to develop a method of automatic extraction from corpora that enables us to accumulate a large amount of knowledge effectively and efficiently. With regards to the processing of proper nouns, we will revise the present specifications by making use of large amounts of text while collecting the necessary knowledge.

References

[Collier, 1996] Collier, N. 1996. English-Japanese Lexical Transfer Using a Hopfield Neural Network, PhD. Thesis, Department of Language Engineering, UMIST, Manchester, UK.

[Cowie, 1992] Cowie, J., J. Guthrie and L. Guthrie. 1992. Lexical Disambiguation using Simulated Annealing, In *Proceedings of COLING-92: The 14th International Conference on Computational Linguistics*, Nantes, pp.359-365.

[Dagan & Itai, 1994] Dagan, I. and A. Itai. 1994. Word Sense Disambiguation Using a Second Language Monolingual Corpus, In *Computational Linguistics*, No.4, 1994, pp.563-596.

[Gale et al., 1992] Gale, W. A., K. W. Church and D. Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods, In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI '92, pp.101-112.*

[Guthrie et al., 1991] Guthrie, J. A., L. Guthrie, Y. Wilks, and H. Aidinejad. 1991, Subject-Dependent Co-occurrence and Word Sense Disambiguation, In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, Berkeley, CA, 1991, pp.146-152.

[Mani, 1996] Mani, I. and T. R. MacMillan. 1996, Identifying Unknown Proper Names in Newswire Text, In Branimir Boguraev and James Pustejovsky(eds) *Corpus Processing for Lexical Acquisition*, MIT Press, Cambridge MA, pp.41-59.

[Strzalkowski & Wang, 1996] Strzalkowski, T. and J. Wang. 1996. A Self-Learning Universal Concept Spotter, In *Proceedings of COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, 1996, pp.931-936.

[Wakao et al., 1996] Wakao, Takahiro, Robert Gaizauskas and Yorick Wilks. 1996. Evaluation of an Algorithm for the Recognition and Classification of Proper Names, In *Proceedings of COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, 1996, pp.418-423.

[Yarowsky, 1992] Yarowsky, D. 1992 Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora, In *Proceedings of COLING-92: The 14th International Conference on Computational Linguistics*, Nantes, pp.454-460.