Disambiguation of Lexical Translations Based on Bilingual Corpora

Hiroshi Nakagawa
Information Technology Center,
The University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo, 113-0033, JAPAN
nakagawa@r.d1.ite.u-tokyo.ac.jp

Abstract

Bilingual dictionaries of machine readable form are important and indispensable information resources for cross-language information retrieval (CLIR), machine translation (MT), and so on. Specific academic areas or technology fields become focused on in these cross language informational activities. In this paper, we describe bilingual dictionary acquisition system which extracts translations from non-parallel but comparable corpora of specific academic fields and disambiguates the extracted translations. The proposed method is two fold. At the first stage, candidates of terms are extracted and ranked from Japanese and English corpus, respectively. At the second stage, ambiguous translations are resolved by selecting a translation of target language which is the nearest ranked to the source language term. Finally, we experimentally evaluate the proposed method.

1 Introduction

Bilingual dictionaries of machine readable form, which we call "MRD" henceforth, are important and indispensable information resources for cross-language information retrieval (CLIR), machine translation (MT), and so on. Specific academic areas or technology fields become focused on in these cross language informational activities. The major difficulty is that developing MRD manually costs too much and also consumes too much time to catch up large number of new terminologies created day by day. To solve this situation, we have to develop an automatic bilingual dictionary acquisition system which uses bilingual corpora as information resources. For this purpose, much research has been done to extract lexical translations including translations of collocations from aligned bilingual parallel corpora (Daille et al.1994), (Smadja et al.1996), (Fung1995b), (Kupiec1993), (Kumano HiraKawa1994), (Haruno et al.1996). However, bilingual parallel corpora are rarely found in the above mentioned academic and technology areas because these areas are growing rapidly. Then, we need lexical translation acquisition system which extracts lexical translations from non-parallel bilingual corpora that are not parallel but cover the same academic or technological area. We call this type of corpora as bilingual comparable corpora henceforth. Very few research results, i.e. (Tanaka1996; Fung1995a) have been published, but they have not yet been satisfactory results. Actually, the method we propose here is similar to (Fung1995a) in its basic idea, but different in several aspects. We describe these differences component by component in the rest of the paper.

It is almost impossible to acquire lexical translation from bilingual comparable corpora from scratch. We usually use bilingual dictionary like Edict(Breen1995) for Japanese-English translation to get the first approximation of lexical translations. Since the resultant translations got directly from the dictionary are often ambiguous, it is essential to disambiguates lexical translations extracted directly from the dictionary.

In this paper, we propose a disambiguation method for Japanese word to English word translations. The proposed method is two fold. At the first stage, simple words and compound words are extracted from Japanese and English corpora respectively. These extracted words are ranked by the method described in section 3. At the second stage, among English words that are the lexical translations found in Edict for the given Japanese word, only the highly relevant words are selected. In this draft, we limit our focus only on translations of simple nouns. The principle of our disambiguation is described in section 2 and 3, and the experimental evaluation is described in section 4.

2 Parallelism of Bilingual Terminologies

As already described, we deal with not parallel but comparable corpora. That means that we cannot use the information of sentences alignment between bilingual corpora. Thus, we need another type of information for disambiguation of lexical translations. For this purpose, we adopt a rank of each word
which is usually used for automatic term recognition (ATR henceforth) task, such as term frequency, tf-idf, etc. Actually much work has been done for ATR (Smadja Mckown 1990), (Smadja 1993), (Kageura Umino 1996), (Franzi Ananiadou 1996), (Hisamitsu Nitta 1996), (Shimohata et al. 1997), (Nakagawa 1997). We extract two sets of words from Japanese and English corpora respectively by applying one of these ATR methods. Extracted words are ranked according to the evaluation measures of individual ATR method. Since we use comparable corpora of the same academic or technology area, extracted words of one language probably find their translations in extracted word candidates of the other language. In this situation, we pose the basic idea as follows.

Suppose that the rank of word in language X is normalized by the number of words in the set of words extracted from the corpora in language X, where X is either A or B. This normalization is extremely different from (Fung 1995a) which normalizes with the number of occurrences of the word. Apparently her normalization depends on the size of corpus. On the contrary, our normalization depends not on the corpus size but on the corpus’ coverage of academic field. Obviously our normalization is more relevant to the academic contents the corpus deals with. Then, the normalized rank is written as \( \text{Rank}(T_x) \). Moreover, the word \( T_a \) of language A is supposed to have more than one translated words \( T_b(T_a), T_b(T_a), \ldots \) in language B. Then the basic idea is this.

**Basic Idea 1** If \( T_b(T_a) \) is more relevant to \( T_a \) than \( T_b(T_a) \), then

\[
\left\| \text{Rank}(T_a) - \text{Rank}(T_b(T_a)) \right\| < \left\| \text{Rank}(T_a) - \text{Rank}(T_b(T_a)) \right\|.
\]

The opposite direction also holds.

Here \( T_b(T_a), T_b(T_a), \ldots \) are sorted in ascending order of

\[
\left\| \text{Rank}(T_a) - \text{Rank}(T_b(T_a)) \right\|,
\]

and result in \( T_b(T_a), T_b(T_a), \ldots \). Namely,

\[
\left\| \text{Rank}(T_a) - \text{Rank}(T_b(T_a)) \right\| < \left\| \text{Rank}(T_a) - \text{Rank}(T_b(T_a)) \right\| < \ldots
\]

Then, owing to the basic idea described above, the word to be selected as the most relevant translation of \( T_a \) is \( T_b(T_a) \). The second most relevant translation of \( T_a \) is \( T_b(T_a) \), and so on.

Now, we have two problems. The first problem is to evaluate how accurate this selecting mechanism is, in other words, to what extent the basic idea 1 holds. We describe experimental evaluation for this problem in section 4.

The second problem is what ATR ranking method fits well for the basic idea 1. In the following section, we introduce the ranking method which would be promising for this purpose. In section 4, we experimentally evaluate the proposed disambiguation methods for lexical translations based on two ranking methods.

### 3 Ranking

In order to extract domain specific words from the given corpora, we have to rank them according to their termhood (Kageura Umino 1996), which roughly means the degree that a linguistic unit is related to domain-specific concepts. As written in (Kageura Umino 1996), the frequency information about a word, like tf-idf, is an approximation of termhood. Obviously the relation between the simple word and complex words which include the simple word is very important. To my knowledge, this relation has not been paid enough attention so far. Nakagawa97 focuses on the method to use this relation. In technical documents, the majority of domain specific words are complex words, more precisely compound nouns. In spite of huge number of technical words being compound nouns, not so many number of simple nouns contribute to make these compound nouns. Considering this fact, we propose a new scoring method which measures the importance of each simple noun. This scoring method measures how many distinct compound nouns contain the simple noun as their parts in a given document or a set of documents. \( \text{Pre} \) (simple word ) and \( \text{Post} \) (simple word ) are introduced for this purpose, and defined as follows.

**Definition 1** In the given corpus, \( \text{Pre}(N) \), where \( N \) is a noun appeared in the document, is the number of distinct nouns that \( N \) adjoins and make compound nouns with \( N \), and \( \text{Post}(N) \) is the number of distinct nouns that adjoin \( N \) and make compound nouns with \( N \).

The key point of this definition is that \( \text{Pre}(N) \) and \( \text{Post}(N) \) count not the number of total occurrences of word which is adjacent to \( N \), but the number of distinct words that adjoin \( N \) or \( N \) adjoins. That means that \( \text{Pre}(N) \) and \( \text{Post}(N) \) do not measure surface statistics of compound nouns containing \( N \), but do measure how the writer of the technical document interprets \( N \) and uses it in the document. If a certain word, say \( W \), expresses the key concept of the system that the document describes, the writer of the document must use \( W \) not only many times but also in various ways that include forming and using many compound nouns that contain \( W \). This kind of usage really reflects the termhood of that word.

In this sense, \( \text{Pre} \) and \( \text{Post} \) very directly measure termhood. Figure 1 shows an example of \( \text{Pre} \) and \( \text{Post} \).

Next, we extend this scoring method to cover compound nouns. For the given compound noun
Figure 1: An example of Pre and Post

\[ \text{Pre}(\text{"file"}) = m \text{ and } \text{Post}(\text{"file"}) = n \]

\[ \text{Imp}(N_1N_2\cdots N_k) = \left( \prod_{i=1}^{k} \left( (\text{Pre}(N_i) + 1) \cdot (\text{Post}(N_i) + 1) \right) \right)^{\frac{1}{k}} \]

\text{Imp}(N) \text{ is normalized by the length of compound noun } N, \text{ and doesn’t depend on the length of } N.

Although Pre and Post are very similar to Context Heterogeneity proposed in (Fung1995a), in our term, she uses Pre and Post separately. On the contrary, we combine them as one single score Imp. In fact, our preliminary experiments of term extraction showed that biasing either Pre or Post over the other did not improve term extraction accuracies. Then, we adopt Imp defined here.

4 Experimental Evaluations

In this section, we experimentally evaluate the method to disambiguate lexical translations which we outlined in section 2. In the actual implementation, we use the difference between the normalized rank of Japanese word Tj and the normalized rank of English word Tc(Tj) which is a translation we find in Edict (Breen1995).

Corpora

The corpora we use for this experimentation are Japanese test collection and English test collection that are used at NTCIR Workshop 1 (Kuno1999). The test collection is the sets of Japanese and English abstracts of papers of four academic societies, namely Japan Architecture Society (JAS), Institute of Electric Engineering (IEE), Institute of Electronics and Communication Engineering (IECE), and Information Processing Society of Japan (IPSJ), published in Japan. A portion of these bilingual corpora are parallel. The percentages of parallel text against the whole corpus of the four corpora will be shown later.

Morphological Analysis and POS Tagging

We use morphological analyzer Chasen (Matsumoto1997) for Japanese corpora, and Brill’s tagger (Brill1994) for English corpora to extract simple and complex nouns.

Ranking

We compare two ranking methods. The first one is the ranking based on Imp described in section 3.

As shown in table 1, the number of J is 55,119 and the number of E is 50,236. The second one is the ranking based only on the word frequencies. The latter ranking is used as a baseline.

Table 1: Corpora used in our experiment

<table>
<thead>
<tr>
<th>Society Name</th>
<th>Number of one-to-one nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAS</td>
<td>815</td>
</tr>
<tr>
<td>IEE</td>
<td>377</td>
</tr>
<tr>
<td>IECE</td>
<td>1092</td>
</tr>
<tr>
<td>IPSJ</td>
<td>720</td>
</tr>
</tbody>
</table>

Table 2: Number of One-to-one corresponding nouns

In this ranking method, not only simple nouns but also complex nouns are equally treated. The second one is the ranking based only on the word frequencies. The latter ranking is used as a baseline.

One-to-one corresponding words

Many Japanese words have just one English translations. More formally, it is stated as follows. Using Edict, usually are there plural Tc(Tj), say Tc1(Tj), Tc2(Tj), ..., for Tj. However, if a set of words extracted from English corpora includes only one of Tc1(Tj), Tc2(Tj), ..., say Tc(Tj), then Tj has one-to-one correspondence to Tc(Tj). These are the ideal cases, where the disambiguation of translations of Tj has been already accomplished. In other words, this is the first fruitful result we obtained by comparing two word sets extracted from Japanese and English corpora, respectively. In table 2, the number of these one-to-one translations in top 10,000 ranked extracted complex and simple nouns are shown for four kinds of corpora described in table 1.

Then, our target is to disambiguate non-one-to-one translations: Tc1(Tj), Tc2(Tj), ..., for Tj.

Disambiguation

We show one example of Tj and Tc1(Tj), Tc2(Tj), ..., translated with Edict for information science area corpora in the following, where distance(Tj, Tc) is defined as follows.

\[ \text{distance}(T_j, T_c) = \left| \text{Rank}(T_j) - \text{Rank}(T_c(T_j)) \right| \]

and the ranking method is Imp based one.
As you expect from this example, \textit{Tc1}(\textit{Tj}) which has the smallest \textit{distance}, would be the best translation, and \textit{Tc2}(\textit{Tj}) of the second smallest \textit{distance} would be the second best translation, and so on. In real applications, the important problem is how many translations we select as \textit{Tj}'s translations. However, we have already ranked translations according to \textit{distance}. Thus, we could use \textit{distance}(\textit{Tj}, \textit{Tc}(\textit{Tj})) as the weight of \textit{Tc}(\textit{Tj}) in actual applications. Anyway, at this moment, it is important to know how accurate disambiguated translations based on \textit{distance}(\textit{Tj}, \textit{Tc}(\textit{Tj})) are. In table 3, we show the recalls and the precisions for three cases. The first row shows the results where \textit{Tc1}(\textit{Tj}) is selected. The second row shows the results where \textit{Tc1}(\textit{Tj}) and \textit{Tc2}(\textit{Tj}) are selected and the third row shows the results where \textit{Tc1}(\textit{Tj}), \textit{Tc2}(\textit{Tj}) and \textit{Tc3}(\textit{Tj}) are selected. To calculate recall and precision, we need the correct translations. In this experiment, we use terminology dictionaries(Aoki1993; Hirayama1995; Nagao1990) to extract correct translations between Japanese terminologies and English terminologies. In table 3, “parallel text ratio” is defined as follows.

\[
\text{parallel text ratio} = (\frac{\text{Para}J \times \text{Para}E}{1})^{1/2}
\]

where

\[
\text{Para}J = (\text{Number of parallel abstracts}) / (\text{Number of the whole Japanese abstracts}),
\]

and

\[
\text{Para}E = (\text{Number of parallel abstracts}) / (\text{Number of the whole English abstracts}).
\]

Also in table 3, \textit{Imp} means \textit{Imp} based ranking, and \textit{FB} means frequency based ranking. \textit{R} and \textit{P} mean Recall and Precision, respectively.

As expected, \textit{Tc1} cases show high precisions and low recall. In \textit{Tc1} and \textit{Tc2} cases, disambiguation using \textit{Imp} based ranking method results in almost 90% of recall. By this fact, if we use these results as translations, we expect higher recall in CLR. As for ranking, \textit{Imp} based method is slightly superior to simple frequency based ranking method.

Moreover, since our method aims at disambiguation of translations for non parallel corpora, we evaluate three cases where the parallel text ratio that is defined previously is 0\%, 50\% and 100\%, respectively. Actually in table 4, R(0), R(50) and R(100) mean recalls for parallel text ratio = 0\%, 50\% and 100\%, respectively, and P(0), P(50) and P(100) are precisions for parallel text ratio = 0\%, 50\% and 100\%, respectively. At this moment we have calculated recall and precision only for Institute of Electronics and Communication Engineering corpora. As shown in table 4, parallel text ratio has no effect on recalls and precisions. That means that our method is proven to be quite robust for extracting translations from non parallel bilingual corpora and disambiguating them.

In these results, we use translations appeared only in terminology dictionaries(Aoki1993; Hirayama1995; Nagao1990) as the correct translations. However, the dictionaries apparently fail to extract quite a few correct translations as follows. In fact, the following translations are examples of automatically extracted translations by our method. This fact implies that the recall and precision of the results of our method would be virtually much higher. Then, our method has already show better quality in many translations than manually made dictionaries. This fact really encourages the promising future of our method.

**Examples of correct translations extracted our method but not appeared in dictionary(Hirayama1995)**
Table 4: Recall and precision for the corpora of Institute of Electronics and Communication Engineering

5 Conclusion

We proposed an automatic translation acquisition system, and experimentally evaluated it. The results are very promising. The next problem to be solved is to extract and disambiguate collocation to collocation translations based on the word to word translations extracted by the method proposed here. This research is financially supported by the grant in aid of Ministry of Education and Academies of Japan.

References


