Disambiguation of Lexical Translations Based on Bilingual Comparable Corpora

Hiroshi Nakagawa
Information Technology Center,
The University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo, 113-0033, JAPAN
nakagawa@r.dl.itc.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 a 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 disambiguate 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 Mckeown1990), (Smadja1993), (Kageura Umino1996), (Frantzi Ananiadou1996), (Hisamitsu Nitta1996), (Shimohata et al.1997), (Nakagawa1997). 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 (Fung1995a) 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 $Rank(Tx)$. Moreover, the word Ta of language A is supposed to have more than one translated words $Tb_1(Ta), Tb_2(Ta), ...$ in language B. Then the basic idea is this.

Basic Idea 1 *If $Tb_i(Ta)$ is more relevant to Ta than $Tb_j(Ta)$ is, then*

$$| Rank(Ta) - Rank(Tb_i(Ta)) |$$

$$<| Rank(Ta) - Rank(Tb_j(Ta)) |.$$

The opposite direction also holds.

Here $Tb_1(Ta), Tb_2(Ta), ...$ are sorted in ascending order of
$| Rank(Ta) - Rank(Tb_i(Ta)) |$, and result in $Tb1(Ta), Tb2(Ta), ...$ Namely ,

$$| Rank(Ta) - Rank(Tb1(Ta)) |$$
$$<| Rank(Ta) - Rank(Tb2(Ta)) |$$
$$< ....$$

Then, owing to the basic idea described above, the word to be selected as the most relevant translation of Ta is $Tb1(Ta)$. The second most relevant translation of Ta is $Tb2(Ta)$, 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 Umino1996), which roughly means the degree that a linguistic unit is related to domain-specific concepts. As written in (Kageura Umino1996) , 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. Pre(simple word) and $Post$(simple word) are introduced for this purpose, and defined as follows.

Definition 1 *In the given corpus, $Pre(N)$, 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 $Post(N)$ is the number of distinct nouns that adjoin N and make compound nouns with N.*

The key point of this definition is that $Pre(N)$ and $Post(N)$ 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 $Pre(N)$ and $Post(N)$ 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, Pre and $Post$ very directly measure termhood. Figure 1 shows an example of Pre and $Post$.

Next, we extend this scoring method to cover compound nouns. For the given compound noun

$$\begin{array}{cc} 1 & dictionary \\ \vdots & \vdots \\ m & user \end{array} \Bigg\} file \Bigg\{ \begin{array}{cc} manager & 1 \\ \vdots & \vdots \\ system & n \end{array}$$

$Pre(\text{“}file\text{”}) = m$ and $Post(\text{“}file\text{”}) = n$

Figure 1: An example of Pre and $Post$

$N_1N_2\cdots N_k$ where N_is are simple nouns, the scores of importance of $N_1N_2\cdots N_k$, which is called $Imp(N_1N_2\cdots N_k)$, is defined as follows.

$$Imp(N_1N_2\cdots N_k) = (\prod_{i=1}^{k}((Pre(N_i)+1)\cdot(Post(N_i)+1)))^{\frac{1}{2k}}$$

$Imp(N)$ is normalized by the length of compound noun N, 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 $Te(Tj)$ 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(Kando1999). 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.

Society name	No. of J	No. of E
JAS	55715	50236
IEE	18008	399
IECE	86364	33076
IPSJ	26815	11860

“ No. of J” means the number of abstracts in Japanese and “ No. of E” means the number of abstracts in English.

Table 1: Corpora used in our experiment

Society Name	Number of one-to-one nouns
JAS	813
IEE	377
IECE	1092
IPSJ	720

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 $Te(Tj)$, say $Te1(Tj)$, $Te2(Tj)$, ... for Tj. However, if a set of words extracted from English corpora includes only one of $Te1(Tj)$, $Te2(Tj)$, ..., say $Te(Tj)$, then Tj has one-to-one correspondence to $Te(Tj)$. 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:$Te1(Tj), Te2(Tj), ...$ for Tj.

Disambiguation
We show one example of Tj and $Te1(Tj)$, $Te2(Tj)$, ... translated with Edict for information science area corpora in the following, where $distance(Tj,Te)$ is defined as follows.

$$distance(Tj,Te) = \mid Rank(Tj) - Rank(Te(Tj)) \mid$$

and the ranking method is Imp based one.

Tj = 処理
$Te1(Tj)$ = processing *distance*(処理, processing)=1.29
$Te2(Tj)$ = teatment *distance*(処理, treatment)=44.7
$Te3(Tj)$ = disposition *distance*(処理, disposition)=88.6
$Te4(Tj)$ = disposal *distance*(処理, disposal)=96.4

As you expect from this example, $Te1(Tj)$, which has the smallest *distance*, would be the best translation, and $Te2(Tj)$ of the second smallest *distance* would be the second best translation, and so on. In real applications, the important problem is how many translations we select as Tj's translations. However, we have already ranked translations according to *distance*. Thus, we could use $distance(Tj, Te(Tj))$ as the weight of $Te(Tj)$ in actual applications. Anyway, at this moment, it is important to know how accurate disambiguated translations based on $distance(Tj, Te(Tj))$ are. In table 3, we show the recalls and the precisions for three cases. The first row shows the results where $Te1(Tj)$ is selected. The second row shows the results where $Te1(Tj)$ and $Te2(Tj)$ are selected and the third row shows the results where $Te1(Tj)$, $Te2(Tj)$ and $Te3(Tj)$ 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.

parallel text ratio = $(ParaJ \times ParaE)^{1/2}$

where
ParaJ = (Number of parallel abstracts)/
(Number of the whole Japanese abstracts),
and

ParaE = (Number of parallel abstracts)/
(Number of the whole English abstracts).

Also in table 3, *Imp* means *Imp* based ranking, and FB means frequency based ranking. R and P mean Recall and Precision, respectively.

As expected, $Te1$ cases show high precisions and low reall. In $Te1$ and $Te2$ cases, disambiguation using *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 CLIR. As for ranking, *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

JAS parallel text ratio = 0.95

	Imp		FB	
	R	P	R	P
$Te1$	0.66	0.95	0.63	0.89
$Te1$ and $Te2$	0.96	0.68	1.0	0.71
$Te1, -2$ and -3	1.0	0.64	1.0	0.64

IEE parallel text ratio = 0.12

	Imp		FB	
	R	P	R	P
$Te1$	0.54	0.65	0.46	0.55
$Te1$ and $Te2$	0.83	0.50	0.83	0.50
$Te1, -2$ and -3	1.0	0.49	1.0	0.49

IECE
parallel text ratio = 0.59

	Imp		FB	
	R	P	R	P
$Te1$	0.73	0.76	0.50	0.52
$Te1$ and $Te2$	0.91	0.48	0.91	0.48
$Te1, -2$ and -3	1.0	0.38	1.0	0.38

IPSJ parallel text ratio = 0.65

	Imp		FB	
	R	P	R	P
$Te1$	0.59	0.68	0.59	0.68
$Te1$ and $Te2$	0.91	0.52	0.81	0.46
$Te1, -2$ and -3	1.0	0.42	0.97	0.41

Table 3: Recall and precision for the corpora

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)

Imp based ranking method

R(0) P(0)	R(50) P(50)	R(100) P(100)
$Te1$		
0.57 0.67	0.58 0.62	0.58 0.68
$Te1$ and $Te2$		
0.88 0.52	0.85 0.51	0.89 0.52
$Te1, -2$ and -3		
0.95 0.45	0.92 0.44	0.97 0.44

Frequency based ranking method

R(0) P(0)	R(50) P(50)	R(100) P(100)
$Te1$		
0.57 0.67	0.55 0.65	0.58 0.68
$Te1$ and $Te2$		
0.86 0.51	0.83 0.50	0.86 0.50
$Te1, -2$ and -3		
0.94 0.44	0.92 0.44	0.97 0.44

Table 4: Recall and precision for the corpora of Institute of Electronics and Communication Engineering

対象 → target,object,subject
運動 → motion,exercise
強度 → strength, intensity
操作 → operation, management
音響 → sound, noise, echo, acoustic
集合 → set
線形 → linear
仕様 → specification
評価 → evaluation

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 Academics of Japan.

References

Breen, J. W. 1995. Edict, Freeware Japanese/English Dictionary. Monash University, Australia.

Brill, E. 1994. Supervised part of speech tagger http://www.cs.jhu.edu/ brill/.

Daille, B., Gaussier, E., Lange, J. M. 1994. Towards automatic extraction of monolingual and bilingual terminology In Proceedings of COLING'94, 515 – 521.

Aoki, S.(editor) 1993. Encyclopedia of Architecture and Building. Shoukokusha, Tokyo.

Hirayama, H.(editor) 1995. Electronics, Information and Communication English-Japanese & Japanese English Dictionary. Kyouritu Shuppan, Tokyo.

Nagao, M.(editor) 1990. Encyclopedic Dictionary of Computer Science. Iwanami Shoten, Tokyo.

Frantzi, K. T. Ananiadou, S. 1996. Extracting nested collocations In COLING'96, 41 – 46.

Fung, P. 1995a. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus In Third Annual Workshop on Very Large Corpora, 173 – 183.

Fung, P. 1995b. A pattern matching method for finding noun and proper noun translation from noisy parallel corpora In Proceedings of ACL'95, 236 – 243.

Haruno, M., Ikehara, S., Yamazaki, T. 1996. Learning bilingual collocations by word-level sorting In Proceedings of COLING'96, 525 – 530.

Hisamitsu, T. Nitta, Y. 1996. Analysis of japanese compound nouns by direct text scanning In Proceedings of COLING'96, 550 – 555.

Kageura, K. Umino, B. 1996. Methods of automatic term recognition:a review Terminology, 3(2), 259 – 289.

Kando, N. 1999. Overview of ir tasks at the first ntcir workshop In Proceedings of the First NTCIR Workshop, 11 – 44.

Kumano, A. Hirakawa, H. 1994. Building an mt dictionary from parallel texts based on linguistic and statistical information In Proceedings of COLING'94, 76 –81.

Kupiec, J. 1993. An algorithm for finding noun phrase corespondences in bilingual corpora In Proceedings of ACL'93, 22 –28.

Matsumoto, Y. 1997. Japanese Morphological Analyzer "ChaSen". Matsumoto Lab. at NAIST.

Nakagawa, H. 1997. Extraction of index words from manuals In Proceedings of RIAO'97, 598 – 611.

Shimohata, S., Sugio, T., Nagata, J. 1997. Retrieving collocations by co-occurrences and word order constraints In Proceedings of 35th ACL, 476 – 481.

Smadja, F. 1993. Retrieving collocations from text: Xtract Computational Linguistics, 19(1), 143 – 177.

Smadja, F., McKewon, K., Hatzivassiloglou, V. 1996. Translating collocations for bilingual lexicons: A statistical approach Computational Linguistics, 22(1), 1 – 38.

Smadja, F. A. Mckeown, K. R. 1990. Automatically extracting and representing collocations for language generation In Proceedings of the 28th ACL, 252 – 259.

Tanaka, K. Iwasaki, H. 1996. Extraction of lexical translations from non-aligned corpora In Proceedings of COLING'96, 580 – 585.