FINDING TERMINOLOGY TRANSLATIONS FROM NON-PARALLEL CORPORA

Pascale Fung
Dept. of Electrical and Electronic Engineering
University of Science & Technology (HKUST)
Clear Water Bay, Hong Kong
pascale@ee.ust.hk
Kathleen McKeown
Computer Science Department
Columbia University
New York, NY 10027

Summary

We present a statistical word feature, the Word Relation Matrix, which can be used to find translated pairs of words and terms from non-parallel corpora, across language groups. Online dictionary entries are used as seed words to generate Word Relation Matrices for the unknown words according to correlation measures. Word Relation Matrices are then mapped across the corpora to find translation pairs. Translation accuracies are around 30% when only the top candidate is counted. Nevertheless, top 20 candidate output give a 50.9% average increase in accuracy on human translator performance.

Subject Areas: Statistical Language Processing

Word Count: 7936

FINDING TERMINOLOGY TRANSLATIONS FROM NON-PARALLEL CORPORA

Summary

We present a statistical word feature, the Word Relation Matrix, which can be used to find translated pairs of words and terms from non-parallel corpora, across language groups. Online dictionary entries are used as seed words to generate Word Relation Matrices for the unknown words according to correlation measures. Word Relation Matrices are then mapped across the corpora to find translation pairs. Translation accuracies are around 30% when only the top candidate is counted. Nevertheless, top 20 candidate output give a 50.9% average increase in accuracy on human translator performance.

Subject Areas: Statistical Language Processing

Word Count: 7936

1. INTRODUCTION

Despite a surge in research using parallel corpora for various machine translation tasks (Brown et al., 1993),(Brown et al., 1991; Gale and Church, 1993; Church, 1993; Dagan and Church, 1994; Simard et al., 1992; Chen, 1993; Melamed, 1995; Wu and Xia, 1994; Wu, 1994; Smadja et al., 1996), the amount of available bilingual parallel corpora is still relatively small in comparison to the large amount of available monolingual text. It is unlikely that one can find parallel corpora in any given domain in electronic form. This is a particularly acute problem in language pairs such as Chinese/English or Japanese/English where there are fewer translated texts than in European language pairs. While we should make use of any existing parallel corpora as lexical translation resources, we should not ignore the even larger amount of monolingual text. However, using non-parallel corpora for lexical translation has been a daunting task, considered much more difficult than that with parallel corpora.

In this paper, we present an initial algorithm for translating technical terms using a pair of non-parallel corpora. Evalution results show translation precisions at around 30% when only the top candidate is considered. While this precision is lower than that achieved with parallel corpora, we show that top 20 candidate output from our algorithm allows translators to increase their accuracy by 50.9%. In the following sections, we first describe a pair of non-parallel corpora we use for experiments, and then we introduce the Word Relation Matrix (WoRM), a statistical word feature representation for technical term translation from non-parallel corpora. We evaluate the effectiveness of this feature with two sets of experiments, using English/English, and English/Japanese non-parallel corpora.

2. BACKGROUND

Although there is a large amount of work in alignment and translation using parallel bilingual corpora, very little attempts have been made to explore another, even larger quantity of resources, namely non-parallel corpora of monolingual texts in the same domain. The reasons are not difficult to find: starting with the IBM statistical machine translation model, the foundation of all statistical bilingual lexicon compilation work has been the correlation between bilingual word or term pairs in terms of their *parallel occurrence pattern*, either in matching sentences, paragraphs, or segments. These occurrence patterns are the discriminatory features for words. Decision functions are then applied using similarity scores between feature vectors to find bilingual word or phrase pairs. In a pair of monolingual texts which are of the same domain, there is no such parallelness in occurrence patterns for lexical units. One cannot speak of aligning sentences, or even aligning segments between, say, the Wall Street Journal and the Nikkei Financial News. Is there *any* discriminatory feature associating a pair of terms in non-parallel texts of the same domain? (Sager, 1990) gave one of the definitions of a domain-specific term as its consistent relationship with other words and terms. Our goal is to first represent this relationship in a vector form, and then use distance measures to find the most similar vector pairs. This is a *discriminant analysis* process and is one of the basic concepts of pattern recognition (Tou and Gonzalez, 1974).

Discriminant analysis has been applied to author characterization from documents (Mosteller and Wallace, 1968), document categorization from queries (Salton and McGill, 1983; Croft, 1984; Turtle and Croft, 1992; Boostein, 1983; Korfhage, 1995), and sense disambiguation between multiple usages of the same word (Dagan and Itai, 1994; Gale et al., 1992a; Gale et al., 1992b; Gale et al., 1992c; Shütze, 1992; Gale et al., 1993; Yarowsky, 1995; Hearst, 1991). All these works are based on using content or context information as discriminatory features. In this section, we focus on the discussion of discriminant analysis using non-parallel corpora.

(Dagan, 1990) was the first to use a pair of non-parallel texts for the task of lexical disambiguation in one of the two texts. Their algorithm is based on the premise that a polysemous word in one language maps to different words in the other language corresponding to its various senses. In his work for sense classification, (Shütze, 1992) formed large vectors containing context words for each word he tries to classify. He then used Singular Value Decomposition to obtain the most discriminative context words for further classification of other new words. Large vectors containing context or collocational words are also used in (Gale et al., 1992a; Gale et al., 1992b; Gale et al., 1992c; Gale et al., 1993; Yarowsky, 1995), to disambiguate multiple senses of a word.

The basic idea in (Dagan, 1990) extends to choosing a translation among multiple candidates (Dagan and Itai, 1994) given contextual information. Given a small segment containing a few words, they represent a feature for a word in terms of its co-occurrence with other words in that segment. A similar idea is later applied by (Rapp, 1995) to show the plausibility of correlations between words in non-parallel text. His paper reported a preliminary study showing that words which co-occurr in a text are likely to co-occur in another text as well. He proposed a matrix permutation method matching co-occurrence patterns in two non-parellel texts, but noted that there are computational limitations to this method. No further results have been reported from this work.

Using the same idea, (Tanaka and Iwasaki, 1996) demonstrated how to eliminate candidate words in a bilingual dictionary. The possibility of using non-parallel corpora for choosing the best translation among a small set of candidates.

All the above works point to a certain discriminatory feature in monolingual texts —context and word relations. However, these works remain in the realm of solving ambiguities or choosing the best candidate among a small set of possibilities. It is argued in (Gale and Church, 1994) that feature vectors of 100,000 dimensions are likely to be needed for high resolution discriminant analysis. It is so far questionable whether feature vectors of lower dimensions are discriminating enough for extracting bilingual lexical pairs from non-parallel corpora with a large number of candidates. Is it possible to achieve bilingual lexicon translation by looking at words in relation to other words? In this paper, we hope to shed some light on this question.

3. NON-PARALLEL CORPORA

There is a large amount of same-domain monolingual material in multiple languages. Unlike parallel corpora, which are clearly defined as translated texts, there is a wide variation of *non-parallel-ness* in monolingual data. Non-parallel-ness are manifested in the following four dimensions:

- The **authors** of the texts can be different. Since the texts are not translated, they are written independently by different people. The authors' writing styles can be very different.
- The **domains** of the texts can be different.
- The **topics** of the texts in the same domain can range from exactly the same, to approximately overlapping.
- The **time period** of texts such as newspaper articles can vary, leading to variations in topics.

The most common text corpora have non-parallel-ness in all the above dimensions. The higher the degree of non-parallel-ness, the more challenging is the extraction of bilingual information. Thus, it is desirable to reduce the dimensionality of non-parallel-ness in the corpora we use. Parallel corpora represent the extreme example where all dimensions of non-parallel-ness except the language are reduced to zero. At the other extreme, newspapers from different time periods such as the New York Times, and the Chinese People's Daily have different authors, sometimes cover different domains, and even have very different perspective on the same events leading to topical differences. Such a corpus would still be considered non-parallel and it would still be a desirable source of bilingual information.

Note that although both are considered non-parallel corpora, different topic texts are easier than different domain texts for bilingual information extraction. Newspapers in different countries, for example, can range from covering different events, to covering the same event but presenting them in different perspectives. We can choose newspaper articles in different newspapers from the same time period so that the corpus would have a lower degree of non-parallel-ness in the topic dimension. We use one such corpora for evaluation of our algorithm. We consider them *same-domain* texts because they focus on news events during the same time period. In addition, they are both written in non-fiction, journalistic styles. Nevertheless, the degree of non-parallel-ness of this type of texts in the topic dimension is not zero because (1) there are many sub-domains in newspaper articles, and (2) newspapers in different countries tend to have different local news focuses, leading to greater non-parallel-ness between the two texts. However, if newspapers of the same *type* are chosen, such as financial newspapers, then the sub-domains can be more focused.

To further reduce the degree of non-parallel-ness in order to have a control experiment on our algorithms, we use two parts of the same newspaper, the Wall Street Journal, in the same language, but from different time periods as a pilot corpus. This corpus also facilitates fast evaluation of the output. We use this type of non-parallel text primarily for testing and evaluation purposes.

As the non-parallel-ness of the texts increases, it is more difficult to find statistical usage patterns in the terms. Non-parallel-ness leads to the following characteristics:

1. No parallel sentences except for very few "boilerplate", such as *"The Dow Jones raised to X points"*, or *"The meeting is adjourned to X date"*.

2. No parallel paragraphs, except for rare cases involving quotes.

3. Fewer overlapping terms and words. This is a result of the above two points,

4. Many words are polysemous. This is often a problem for terminology translation. Words like *bank* can have two translations in French— *banque (of Paris)* and *bord (of a river)*. However, language pairs sharing a common root sometimes share the same degree of polysemy between word pairs. For example, *interest/intérêt* both have the senses *interest rate* and *in one's interest*. This latter property ensures that polysemy is a lesser problem in parallel corpora where texts are translations of each other. However, when two texts are non-parallel, this becomes a more serious problem. In addition, it is an even bigger problem between language pairs such as English and Chinese since these language pairs are developed completely independently of each other. A higher degree of polysemy leads to more many-to-many mappings in translation.

Because of the above characteristics, lexicon translation from non-parallel corpora is a far more difficult task than that from parallel corpora. In this paper, we describe our findings of a statistical signature feature relating a technical term in a text of one language to its counterpart in a same-domain text in another language. We hope that the discovery of such a signature feature reveals certain usage pattern in technical term usage in multilingual texts, and will boost the interest in further research on using non-parallel corpora for lexicon compilation.

4. TWO PILOT NON-PARALLEL CORPORA

In our experiments, we use two sets of non-parallel corpora: (1) Wall Street Journal (WSJ) from 1993 and 1994, divided into two non-overlapping parts. Each resulting English corpus has 10.36M bytes of data. (2) Wall Street Journal in English and Nikkei Financial News in Japanese, from the same time period. The WSJ text contains 49M bytes of data, and the Nikkei 127M bytes. Since the Nikkei is encoded in two-byte Japanese character sets, the latter is equivalent to about

60M bytes of data in English.

The English Wall Street Journal non-parallel corpus gives us an easier test set on which to start. The output of this corpus should consist of words matching to themselves as translations. It is useful as a baseline evaluation test set providing an estimate on performance.

The WSJ/Nikkei corpus is the most non-parallel type of corpus. In addition to being written in languages across linguistic families by different journalists, WSJ/Nikkei also share only a limited amount of common topic. The Wall Street Journal tends to focus on U.S. domestic economic and political news, whereas the Nikkei Financial News focuses on economic and political events in Japan and in Asia. Due to the large difference in content, language, writing style, we consider this corpus more difficult than others. However, the result we obtain from this corpus gives us a lower-bound on the performance of our algorithm.

5. THE WORD RELATION MATRIX FEATURE

We pointed out earlier that the most important characteristics for technical and domain terms are *standardization* and *consistency* (Pinchuck, 1977). From these, we derived that there is a fixed usage pattern of such terms in large texts.

Statistical domain term extraction algorithms (Smadja, 1993; Fung and Wu, 1994) make use of the consistency characteristic, reflected by frequently appearing token groups, to find closely associated token groups as terms. Bilingual lexicon translation algorithms for parallel corpora again make use of the fixed association between a pair of bilingual terms, reflected in their frequent co-occurrences in translated texts, to find lexicon translations.

We propose to take another look at the associations between monolingual lexical units, and between bilingual or multilingual lexical units, to find a consistent pattern. This pattern will be represented as statistical word features for translation. Based on previous work in monolingual and bilingual lexical unit associations, we postulate the following:

1. If a domain-specific word or term A is closely correlated with another word B in text T, then its counterpart in the other language A' is also closely associated with B', the counterpart of word B, in T'.

2. If A is less related with C, then its translation A' is less associated with C', the translation of C.

3. Given a large set of words $B = (B_1, B_2, \ldots, B_m)$, a word A is closely associated with only some of the words, in a subset $b \in B$.

4. If A is closely associated with a set of words $B_1, B_2, \ldots, B_n$ to *varying degrees*, then A' is closely associated with a set of words $B_1, B_2, \ldots, B_n$ to *similar* varying degrees.

We illustrate the above postulations with the word *debentures* in the WSJ corpus. Let A be the word *debentures.* T is the first part of WSJ from 1987, and T' the second part. Figure 1 shows the segments from both texts containing the word *debentures.*

The first three postulations are illustrated in Figure 1 as follows:

1. *debentures* is most closely correlated with *million* and *due*, in both T and T'.

2. *debentures* is less related to *engineering*, which does not appear in any segments containing *debentures*.

3. Given all words in T, *debentures* is closely correlated with a subset of words. In Figure 1, this subset consists of *million, due, convertible, subordinated*, etc.

Figure 2 illustrates postulation #4: T consists of all the segments in one part of the Wall Street Journal containing *debentures*. T' consists of all the paragraphs in another (non-parallel) part of WSJ containing *debentures*. Among the 541 unique words in T and the 585 words in T', there are 272 common words. Part of the set of words $B_1, B_2, \ldots, B_n$ correspond to the words in Figure 2, in the left column. Part of $B'_1, B'_2, \ldots, B'_n$ are in the right column.

Plots of frequency of these context words in T and T' are shown in Figure 3. The relative frequency of each of the context words are calculated by taking the above frequencies and dividing them by the frequencies of *debentures* in context T and context T'. Since the frequencies of *debentures* are very similar in the two contexts, we can assume that two the plots of the relative frequencies of context words would be almost the same as Figure 3. These plots show that there is a consistency between the *ranking* of the relative frequencies of words in the contexts T and T' for *debentures*— if a word occurs often in context T, it also occurs often in context T'. This illustrates the postulation that *A and A' are associated with the set of words in similar patterns*, that context segments containing the same domain-specific word share a similarity in its lexical profile.

In actual translation task, where context T and context T' are made up of words of different languages, we need to align along the horizontal axis context words in T which correspond to those in T', in order to give plots like those in Figure 3. We also need to do so for every term or word we are interested in translating, and for those in the other langauge we consider as likely candidates. After that, we would have to match the plots and find the most similar pairs.

In the subsequent sections, we demonstrate how we elaborate the above postulations into a statistical word signature feature representation.

Figure 1: Part of the concordances of the word *debenture* in WSJ_1 and WSJ_2 showing a similar set of closely-related words.

Universal said its 15 3/4% **debentures** *due* Dec
$75 *million* of *convertible* **debentures** *due* 2012
sold $75 *million* of 6% **debentures** priced at par and *due* Sept
of **debentures** for each common share
sold $40 *million* of 6 1/4% *convertible* **debentures** priced at par and *due* March 15
Lifestyle will pay $575 plus accrued interest for each of its 13% *convertible* *subordinated* **debentures** inste
GTE offered a $250 *million* issue of 8 1/2% **debentures** *due* in 30 years
Domtar said it launched an offering in Canada of a $100 *million* Canadian offering of 24-year **debentures**
$250 *million* of notes *due* 1997 and $250 *million* of **debentures** *due* 2017
$60 *million* of *convertible* senior *subordinated* **debentures** *due* 2012
sold $300 *million* of 7 1/2% *convertible* **debentures** *due* 2012 at par
of **debentures** for each common share
said it agreed to issue $125 *million* Canadian in *convertible* **debentures**
a $150 *million* issue of FPL Group Capital **debentures** *due* in 30 years was priced for offering today
senior *subordinated* **debentures** was offered through Drexel Burnham Lambert Inc
9 *million* provision to cover a proposed purchase of the company's 10 7/8% senior *subordinated* **debenture**
said it completed the redemption of all $16 *million* of its 9% *subordinated* **debentures** *due* 2003
Alberta-based oil and gas producer said the securities include $30 *million* of *subordinated* **debentures** an
Moody's assigned a Baa-3 rating to a proposed $100 *million* *convertible* *subordinated* debenture issue
Valero's *subordinated* **debentures** to Ba-1 from single-B-2 and depository preferred to Ba-1 from single-B-

million for the second quarter in connection with redeeming some **debentures** outstanding
and its 12 1/2% senior *subordinated* **debentures** at par
instead of the previous $150 *million* of notes and $50 *million* of *convertible* **debentures**
$20 *million* of *convertible* **debentures** *due* June 1
on the company's *convertible* *subordinated* **debentures** and liquid yield option notes
issues of $110 *million* of senior notes *due* 1997 and $115 *million* of *convertible* **debentures** *due* 2012
convert the stock in two years into **debentures** at the rate of one share for $25 of **debentures**
said it reached an agreement with holders of $30 *million* of its *convertible* *subordinated* **debentures**
and preferred shareholders will receive *subordinate* **debentures** with an interest rate of 7 3/8%
downgraded the *subordinated* **debentures** of Bank of Montreal
7 *million* of 12% *subordinated* sinking fund **debentures** *due* 1999 was oversubscribed by about $28 *millio*
subordinated discount **debentures** *due* 1999 and $100 *million* of
announced an offering of $25 *million* principal amount of 7% *convertible* **debentures** *due* 2012
5 *million* charge from the redemption of **debentures**
common shares and $35 *million* of *convertible* **debentures** *due* 2012
$35 *million* of *convertible* **debentures** *due* May 15
financed with $450 *million* of new Western Union senior secured **debentures** to be placed by Drexel
Commission to issue as much as $125 *million* of 30-year **debentures** packaged with common stock
redeem its entire $55 *million* face amount of 8 3/4% *convertible* *subordinated* **debentures** *due* 2011
holders may convert their **debentures** into common stock at a price of $6
6 *million* of tax-exempt sinking fund **debentures** to Ba-2 from single-A-3
rating of B1 to a $50 *million* *convertible* *subordinated* debenture issue to be offered by this Stamford
an offering of $400 *million* of senior notes *due* 1994 and $300 *million* of *subordinated* **debentures** *due* 19
on senior *convertible* **debentures** and Eurodebentures to Ba-2 from Baa-3

word rank in A	frequency in A	frequency in B	word rank in B	frequency in A	frequency in B
million	154	115	due	126	123
due	126	123	million	154	115
convertible	95	77	convertible	95	77
subordinated	75	69	subordinated	75	69
said	38	32	said	38	32
common	27	21	common	27	21
senior	26	16	share	16	16
offering	17	16	senior	26	16
each	17	15	offering	17	16
stock	16	14	each	17	15
shares	16	7	stock	16	14
share	16	16	amount	9	14
notes	15	13	notes	15	13
sinking	13	11	face	8	13
fund	13	11	par	11	12
preferred	12	8	sinking	13	11
issue	12	8	fund	13	11
sold	11	9	priced	6	10
par	11	12	holders	2	10
redeem	10	6	sold	11	9
March	10	3	preferred	12	8
April	10	2	outstanding	6	8

Figure 2: Common content words in the context of *debentures* in non-parallel texts with similar frequency ranking.

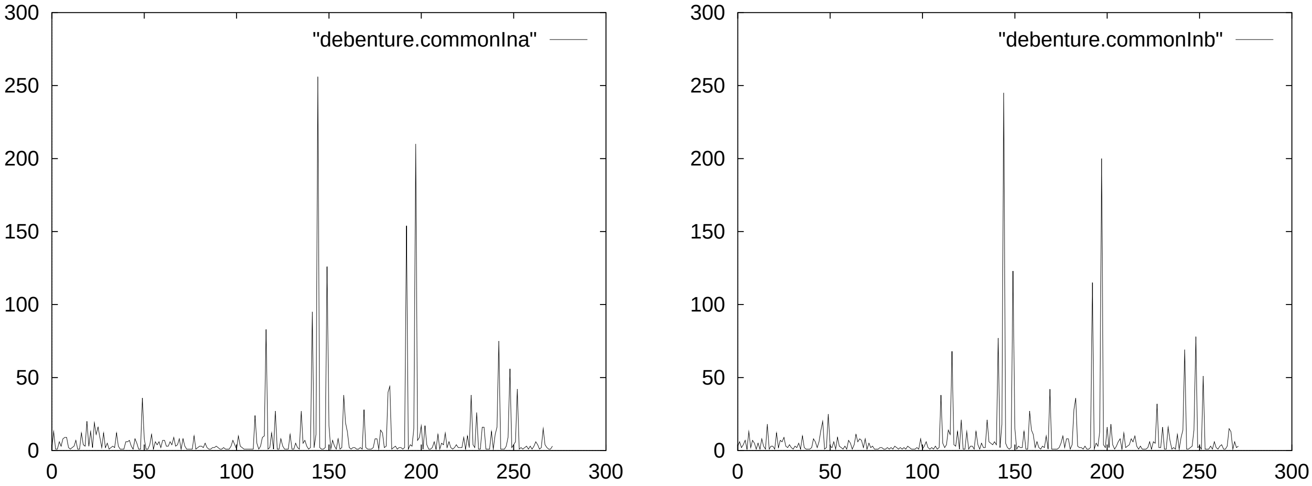


Figure 3: Common word associations for *debenture* in both texts. The horizontal axis represents the common words. The vertical axis is the associatin score.

6. AN ALGORITHM FOR FINDING BILINGUAL WORD PAIRS FROM NON-PARALLEL CORPORA

Following the postulations we just presented, we propose the following algorithm of finding domain word or term translation pairs from a set of known *seed words*:

1. Given a bilingual list of known word pairs (i.e. seed words)

2. For every unknown word or term e in language 1, find its $association1$ with every word in the seed word list in language 1 $\Rightarrow$ relation vector $WoRM1$

3. Similarly for unknown words c in language 2, find its $association1$ with every word in the seed word list in language 2 $\Rightarrow$ relation vector $WoRM2$

4. Compute $association2(WoRM1, WoRM2)$; if it is high, e and c are translated word pairs, otherwise not.

Section 7 describes the type of seed words used in the algorithm. $association1$ is an association measure between *monolingual* term and word pairs. We will show how segment sizes affect $association1$ scores. In section 8, we discuss using Mutual Information or the relative occurrence frequency as this measure. $association2$ is a association measure between *bilingual* pairs of domain-specific terms. In section 9, we discuss the choice between using the Cosine Measure or the Euclidean Distance for this purpose. Overall, it is important to (1) obtain a suitable set of seed word pairs, (2) suitable segment sizes, (3) a suitable $association1$ and (4) a suitable $association2$ which will give us bilingual term pairs from non-parallel corpora of different categories.

7. BILINGUAL SEED WORDS

The above algorithm can be carried out if we knew all mappings of words in the contexts A and B for all unknown words or terms. This does not seem impossible at first since there are a large number of online dictionaries which can be used to first translate these context words. For the English/Japanese WSJ/Nikkei non-parallel corpus, we employed a Japanese/English online dictionary, EDICT, to carry out this first step. It contains 57,885 entries, with one Japanese word or term mapped to multiple English translations.

However, we quickly found out that there are some obstacles to using online dictionary entries to map words in context A to context B:

1. Most dictionary entries have multiple translations. Mononsemous words can have multiple translations, especially in languages across linguistic families. Figure 4 shows that near 50% of the 57,885 entries have two translation candidates.

2. Some of the words are polysemous. A polysemous word like "*interest*" has multiple sense translations in Japanese. Each of these Japanese translations can in turn be polysemous with multiple English correspondences. This many-way polysemy can result in a large set of words unrelated to each other being linked together. As shown in Figure 5, we found that across English/Japanese, about 48% of 22052 English words are translations of multiple Japanese words or terms, about 37% of nouns and adjectives are translations of multiple Japanese words or terms. This means that many Japanese words or terms in EDICT share at least one common English word as one of the senses.

3. Since the two texts are non-parallel, words appearing in one might not appear in the other text. In our experiment with the contexts for *debentures*, there are 272 common words out of 541 words in context A and 585 words in context B.

4. Words in the corpora do not conform to dictionary formats. e.g. In EDICT, all Japanese verbs are in plain form and the English verbs begin with *"to ..."*. No inflections of verbs or adjectives have been included, except in idiomatic expressions. However, verbs occuring in the texts have various conjugations.

5. English and other European languages also have case differences for the same word, and singular and plural forms. Across European language pairs, such as English and French, these case differences roughly have a one-to-one mapping between languages (e.g. plural nouns in English map to plural nouns in French). Across language pairs such as English and Japanese or Chinese, such differences in English do not translate as these languages do not have case differences or consistent singular/plural forms (plural forms are only employed to people in Chinese and Japanese, in both cases it is indicated by a single character "men" or "tachi").

6. In many cases, the author of the dictionary chose to use Oxford (British) standard spelling (-our, -ize) whereas our WSJ corpus uses American spelling.

candidate number	number Japanese words	percentage
1	21164	36.6%
2	28788	49.7%
3	5715	9.9%
4	1394	2.4%
5	423	0.73%
6	177	0.31%
7	93	0.16%
8	39	0.067%
9	32	0.055%
10	19	0.033%
11	11	0.019%
12	10	0.017%
13	6	0.01%
14+	14	0.02%

Figure 4: Most Japanese words in EDICT have more than one translation candidates.

frequency	percentage words	percentage nouns/adjectives
1	52.16%	63.18%
2	16.75%	18.71%
3	7.96%	8.44%
4	4.96%	4.07%
5	3.28%	2.34%
6	2.39%	1.32%
7	1.64%	0.86%
8	1.28%	0.35%
9	1.11%	0.24%
10	0.79%	0.2%

Figure 5: 48% of the English words are translations of more than one Japanese words.

Disambiguating all words in the Wall Street Journal/Nikkei texts, and lemmatizing these words into the dictionary formats are in themselves no trivial tasks. Instead, we propose to use a more reliable, less ambiguous subset of these dictionary entries as the *seed word list*.

Considering the above factors, we use shell scripts with the following criteria to filter EDICT entries into a seed word list:

1. Choose the entries where the Japanese word occurs in the Nikkei text and the English ones occur in the WSJ corpus.
2. Choose those which occurr with frequency between 100 and 10,000 in the WSJ corpus.
3. Content words are more reliable seed words than function words since the latter appear almost everywhere in the text and thus co-occurr with most words in a non-discriminative manner. So choose those words in the dictionary which are nouns, verbs, or adjectives, in their basic word stem forms, without inflections.
4. Choose those English words which are themselves unique translations of certain Japanese word in the EDICT dictionary. This reduces the chance of these entries being polysemous.
5. Even words which are not polysemous can have many translations in the other language. It is useful to include multiple translations rather than one-to-one mapping per entry. Thus, at this point, the seed word list has one English word and multiple Japanese words per entry.

The resulting seed word list we choose from EDICT have 1,416 entries. By manually going through this list, we again filter out some translations which are improbable in our corpora.

In our experiment with English/English WSJ corpus from different years, the seed words we chose are 307 words with frequency between 400 and 3900.

Such seed words are the textual anchor points in non-parallel corpora. From these seed words found in the dictionary, we can obtain statistical word features for each new word or term not found in the dictionary. In general, it is better to have *as many seed words as possible* for deducing the remaining unknown word or term. Meanwhile, our statistical feature would be more reliable if we can find context segments which contain *multiple occurrences of the new word.*

8. A WORD IN RELATION TO SEED WORDS

In the previous section, we presented the relationship between the word *debentures* to the words which appear in the same context in plots shown in Figure 3. In the plots, the y-axis represents the relative frequency of a context word or a seed word to that of *debentures.* In general, this represents the relation between the unknown word or term to the seed word. Hence, the matrix is called the Word Relation Matrix (WoRM).

As we described in previous sections, word relations are important statistical information which has been successfully employed to find monolingual collocations, words and terms from unsegmenated Asian language texts, and to find bilingual word pairs from parallel corpora. Word relations $W(w_s, w_t)$ are computed from some general likelihood scores based on the co-occurrence of words in some common segments. Segments are either sentences, paragraphs, or string groups delimited by some anchor points. We repeat the marginal and joint probabilities of the bilingual word pair below:

$$\begin{aligned}
\Pr(w_s = 1) &= \frac{a+b}{a+b+c+d} \\
\Pr(w_t = 1) &= \frac{a+c}{a+b+c+d} \\
\Pr(w_s = 1, w_t = 1) &= \frac{a}{a+b+c+d} \\
\text{where } a &= \text{number of segments where both words occur} \\
b &= \text{number of segments where only } w_s \text{ occur} \\
c &= \text{number of segments where only } w_t \text{ occur} \\
d &= \text{number of segments where neither words occur}
\end{aligned}$$

All relation measures use the above likelihood scores in different formulations. In our Word Relation Matrix (WoRM) representation, the relation measure $W(w_s, w_t)$ is between a seed word w_s and an unknown word w_x. a, b, c and d are computed from the segments in the monolingual text of the non-parallel corpus.

$W(w_x, w_s)$ can be any relation measure such as average mutual information, the Dice coefficient, or weighted mutual information. In our algorithm, we choose to use the weighted mutual information score:

$$\Pr(w_x = 1, w_s = 1) \log_2 \frac{\Pr(w_x = 1, w_s = 1)}{\Pr(w_x = 1) \Pr(w_s = 1)}$$

Given n seed words $(w_{s1}, w_{s2}, \ldots, w_{sn})$, we thus obtain a Word Relation Matrix for w_x to be:

$$(W(w_x, w_{s1}), W(w_x, w_{s2}), \ldots, W(w_x, w_{sn}))$$

As an initial step, all $\Pr(w_s = 1)$ are pre-computed for the seed words in both languages. We have experimented with various segment sizes, ranging from phrases delimited by all punctuations, a sentence, to an entire paragraph.

From our experiment results, we conclude that the right segment size is a function of the frequency of the seed words[1]:

$$\text{segment size} \propto \frac{1}{\text{frequency}(W_s)}$$

For example, content words do not occur as frequently as function words. If the seed words are mostly content words, then they would not co-occur very often with the new words in the same segments. However, if the segments are large, such as the size of an entire paragraph, then the chances of co-occurrence between content seed words and new words would be higher. On the other hand, if the segments are as small as that between any two punctuations, then the chances of co-occurrence are too low.

If the seed words include some frequent words, and if the segment size is as large as a paragraph size, then these frequent seed words could occur in every single segment. In this case, everywhere the new words appear, these frequent seed words would also appear. The chances for co-occurrence between such seed words and all new words are very high, close to one. Such seed words are too biasing in large segments. If smaller segment size is chosen, then the chances of co-occurrence between frequent seed words and new words are

[1] To a lesser extent, segment size is also dependent on the language pairs, and the writing style of the texts. However, these factors are not deducible without empirical studies involving different sets of corpora.

lower, and therefore less biasing.

Consequently, we chose to use the paragraph as the context window size for our experiment on Wall Street Journal/Nikkei Corpus since all the seed words are mid-frequency content words. We computed all binary vectors of the 1,416 seed words w_s where the i-th dimension of the vector is 1 if the seed word occurs in the i-th paragraph in the text, zero otherwise.

We chose to use smaller segment size – that between any two punctuations, to be the context window size for the Wall Street Journal English/English corpus since many of the seed words are frequent words.

We vary segments sizes according to the frequency of seed words. Alternatively, we could also set the segment size, and then choose the type of seed words accordingly. In general, mid-frequency content words make better seed words as they are less ambiguous. So we suggest using paragraphs as context segments, and mid-frequency content words as seed words when possible.

Next, $Pr(w_x = 1)$ are computed for all unknown words x in both texts. The WoRM vectors are then sorted according to $W(w_x, w_{si})$. The most correlated seed word w_{sj} will have the top scoring $W(w_x, w_{sj})$.

As an example, using 307 seed word pairs in the WSJ/WSJ corpus, we obtain the following most correlated seed words with *debentures* in two different years of Wall Street Journal as shown in Figure 6. In both texts, the same set of words correlate with *debenture* closely. Note that the set of words in Figure 6. a subset of the list which is partly shown in Figure 2.

Whereas in Figure 2, the "seed words" are all common words which co-occur with *debentures* in the two texts, the seed words in this figure include some which co-occur with *debentures* only in one text, but not in the other. It also includes some which does not co-occur with *debentures* in either text. This is a natural consequence of postulation #3.

WoRM plots of *debentures* and *administration* are in Figures 7 and 8 respectively. The horizontal axis has 307 points representing the seed words, vertical axis has the value of the relation scores between these 307 seed words and our example words. These figures show that the Word Relation Matrix of the same words are similar to each other, and different between different words. This follows postulations #1, #2 and #4.

As another example, the 50 seed words which correlate most closely with *Nikkei* in the Wall Street

seed word	corr1(text1)	seed word	corr1(text2)
amount	1083.35	amount	1083.35
July	695.58	offered	646.30
offered	646.30	preferred	551.50
Canadian	596.42	July	695.58
preferred	551.50	June	393.14
June	393.14	exchange	387.16
exchange	387.16	issue	373.80
issue	373.80	notes	229.45
notes	229.45	gas	158.60
gas	158.60	Capital	157.64

Figure 6: The sets of most closely related seed words with *debentures* in two texts are very similar.

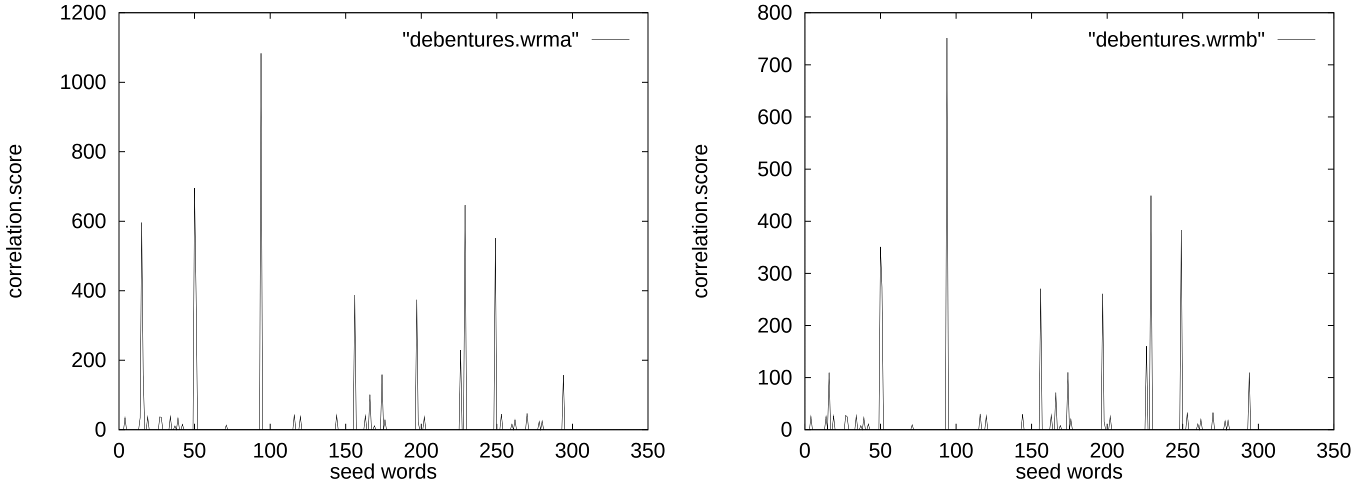


Figure 7: Word relation matrix for *debenture* in both texts.

Journal/Nikkei corpus are in Figure 9. Seed words marked with a # are those which occur in both lists. This example again illustrates postulations #1 and #2.

From these examples and others, we are able to follow postulations #1 to #4 to propose the Word Relation Matrix as a discriminative feature for matching bilingual lexical pairs in non-parallel corpora.

9. MATCHING WORD RELATION MATRICES

When all unknown words are represented in WoRMs, a matching function is needed to find the best WoRM pairs as bilingual lexicon entries. There are many similarity measures and metrics we can use to measure the closeness of two WoRMs (Tou and Gonzalez, 1974). One assumption we make is that the seed

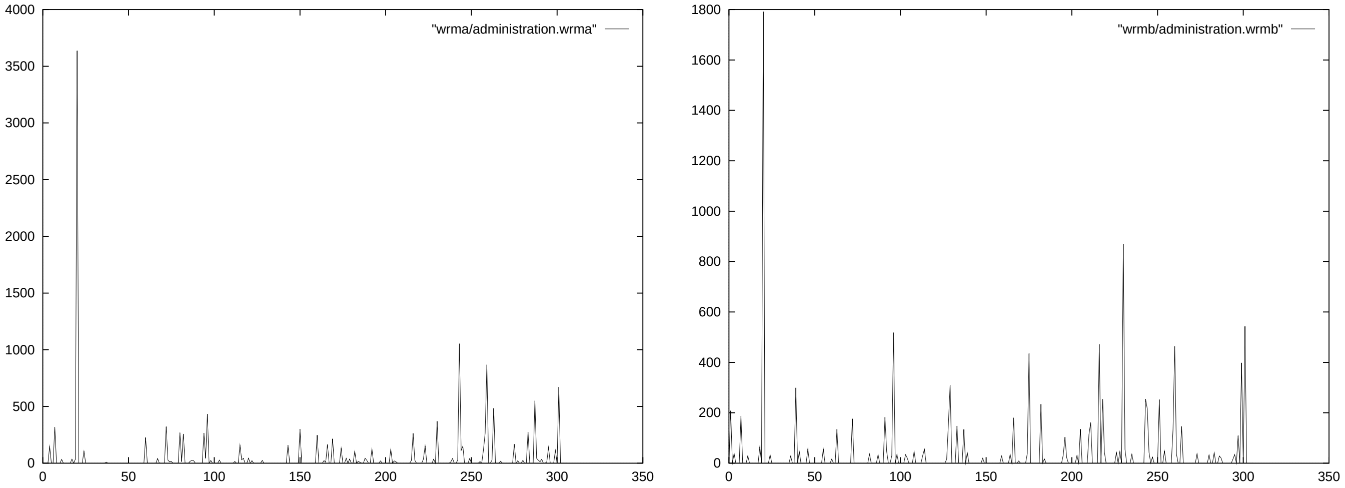


Figure 8: Word relation matrix for *administration* in both texts.

words are independent of each other. That is to say that each dimension of the WoRMs should be treated independently. Another assumption we make is positive definiteness, i.e. the similarity between two vectors is never negative. A third assumption is symmetry, i.e. the similarity score between WoRM A and WoRM B is the same as that between B and A. These assumptions lead to our particular choices of similarity measures.

When matching vectors are very similar such as those in Figure 3, it seems a simple measure like the Euclidean Distance could be used to find those matching pairs, where the Euclidean Distance between two WoRM vectors w_s, w_t is:

$$\mathcal{E} = \sqrt{\Sigma_{1<i<n}(w_{s_i} - w_{t_i})^2}$$

However, most word pairs in the corpus look more like those in Figure 7. It is not a necessary condition for a bilingual word pair to be associated to the exact same extent with the same seed words. It is sufficient that they share a significant number of closely related seed words . The y value of a new word is high when there is a x-th seed word which co-occur with it significantly. If a pair of bilingual words are supposed to be translations of each other, they should share the most significant y values. From this observation, we suggest that perhaps a measure like the Cosine Measure would be more appropriate where:

seed word	score1(text1)	seed word	score1(text2)
morning#	33300	mean	290414
session	31886	average#	203908
volume	24974	balance	203289
close#	19472	futures#	131510
index#	18347	late	100084
trading	16094	old	89855
day#	14556	securities	80196
rose	11512	bonds	77357
futures#	6752	coupon	77291
buying	5444	bond#	76956
overseas	5036	ticket	72212
market	4987	value	66989
benchmark	3168	price	66101
stock#	3064	stock#	64510
stocks#	3055	cost	58930
reform	3053	speculation	56998
closed	2784	worth	55505
political	2306	stocks#	53738
rally	2265	products	53738
yen#	2264	trade#	52130
advanced	2184	end	50795
profits	2135	yen#	49304
highest	2012	transactions	45899
foreign	1921	investment	44868
level	1857	share	33811
high	1846	thing	30541
lost	1716	many	27929
late	1580	put	27470
end	1514	index#	26974
settlement	1457	arbitration	26894
final	1419	day#	26362
trust	1400	trend	24435
thin	1276	week#	23909
pressure	1272	newspaper	23363
finish	1224	movement	23092
average#	1182	morning#	23059
gain	1171	money	22886
bond#	1144	recover	22761
early	1125	brand	22323
opening	1092	handle	22203
strength	1066	rebound	21046
year	1019	design	21043
push	959	phase	20995
week#	957	house	20462
currency	934	engine	19698
further	848	advantage	19276
momentum	843	interest	19146
belief	812	reaction	18835
strong	804	steep	18290
bit	793	sudden	18259

Figure 9: Some common words among the 50 seed words most closely related to *Nikkei* in the WSJ/Nikkei corpus. The scores are multiplied by 10^6.

$$\mathcal{C} = \frac{\Sigma_{1<i<n}(w_{s_i} \cdot w_{t_i})}{\sqrt{\Sigma w_{s_i^2} \cdot \Sigma w_{t_i^2}}}$$

The Cosine Measure has long been used by researchers in information retrieval tasks (Salton and McGill, 1983; Croft, 1984; Turtle and Croft, 1992; Boostein, 1983; Korfhage, 1995) to compare features vectors characterizing various text contents. In our case, the Cosine Measure gives the highest value to vector pairs which share the most non-zero y values. Therefore, it favors word pairs which share the most number of closely related seed words. However, the Cosine Measure is also directly proportional to another parameter, namely the actual $(w_{s_i} \times w_{t_i})$ values. Consequently, if w_s has a high y value everywhere, then the Cosine Measure between any w_t and this w_s would be high. This violates postulation #4 in that although w_s and w_t might not correlate closely with the same set of seed words, the matching score would be nevertheless high. This is another supporting reason for choosing mid-frequency content words as seed words.

There are perhaps various heuristics one can use to remedy the pitfall of the Cosine Measure. One can even propose other metrics for matching these WoRM vectors.

However, in this paper, we will provide a first evaluation with the basic form of the Cosine Measure.

10. EVALUATION

We have evaluated the WoRM feature on two sets of non-parallel corpora (1) Wall Street Journal material from 1993 and 1994 in English/English; (2) Wall Street Journal material in English, from January to March, 1994 and Nikkei Financial News material in Japanese, from the same time period.

10.1. Matching English words to English

The evaluation on the WSJ/WSJ English/English corpus is intended as a pilot test on the discriminative power of the Word Relation Matrix: (1) the contents of the two texts are similar, minimizing the effect of different topics, different styles and different vocabulary set. (2) There is a minimum amount of polysemous mismatching between word pairs—if *interest* has two meanings in the first text, it also has two meanings in the second text. (3) The seed word list is easy to compile, by choosing a set of English words without having to find the translations. (4) In addition, using an English/English test set, the output can be evaluated automatically—a translated pair is considered correct if they are identical English words.

Since this corpus is English/English, we do not have the constraint of choosing seed words from a bilingual dictionary. In order to obtain content words as seeds, 307 seed words are chosen according to their occurrence frequency (mid-ranged) to minimize the number of function words. However, occurence frequencies of some of these seed words might still be higher than those of content words in a true non-parallel bilingual corpus. As a result, we have chosen to use a smaller segment length to be the context window size, according to our analysis in Section 7. Two words co-occur in the same segment delimited by two punctuations. As explained in Section 9, the frequent nature of the seed words led to our choice of the Euclidean Distance, instead of the Cosine Measure to avoid the effect of frequent seed words. The choices of segment size, seed words, and Euclidean Distance measure are all direct consequences of the atypical nature of the English/English pilot test set. We will show later that different choices are made for a more typical non-parallel corpus.

We used a test set of 582 by 687 single words. Some of the 687 words obviously have no correspondence in the 582 word set. We computed the WoRM feature for each of these test words and compute the Euclidean Distance between every word in Set A and every word in Set B. For each word e in Set A, we sort the list of Set B words according to their $relation2$ score with e. We then calculate the accuracy by counting the number of e words whose top one candidate is identical to itself. In other words, we do not count collocation translations such as *North/Korea* as correct. Thus, by this most stringent measure, we obtain a precision of 29%.

By allowing N-top candidates, the accuracy improves as shown in in the graphs for 582 words output in Figure 10. In other words, for N-top candidate precision calculation, a translation is correct if it appears among the first N candidates. If we find the correct translation among the top 100 candidates, we obtain a precision of around 58%.

N-top candidate lists are useful as translator aids. Translators can use candidate word lists to facilitate translations of technical and domain terms they are not very familiar with.

This evaluation result is in some sense an upper bound of the WoRM feature. However, we project that the real upper bound would be higher if the test words are actually unambiguous terms. In this

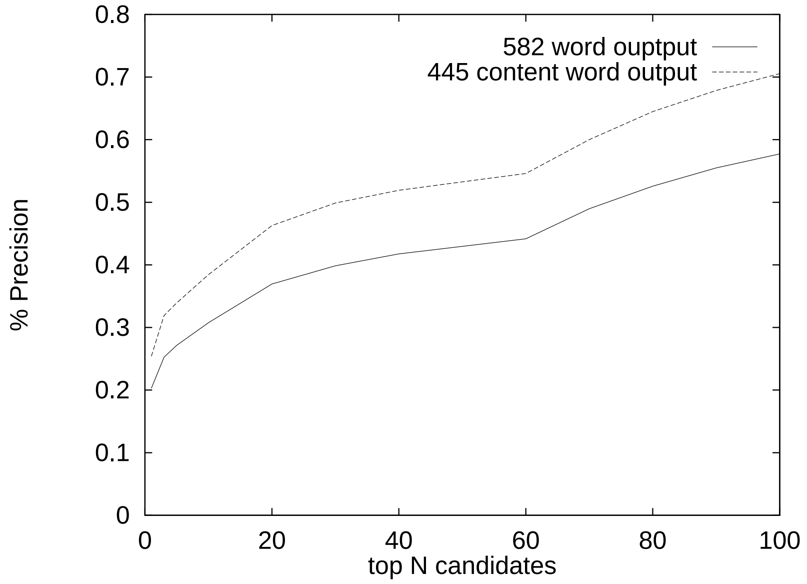


Figure 10: Evaluation results of WoRM in 1993/94 Wall Street Journal.

English/English pilot corpus, the test words are chosen according to their frequencies, and are mostly polysemous. To reduce the ambiguity of the test words, we manually filter out the non-content words from the 582-word set. We obtain 445 content words. The precisions at different top N candidates for this 445-word set are higher. We believe the accuracy would be even higher if we only look at really unambiguous test words, such as an entire technical term. It is well known that polysemous words only have one sense when used as part of a colocation or technical term (Yarowsky, 1993). As we shall see in the following section describing the evaluation on a true non-parallel corpus, we will indeed match full technical terms across languages rather than individual words.

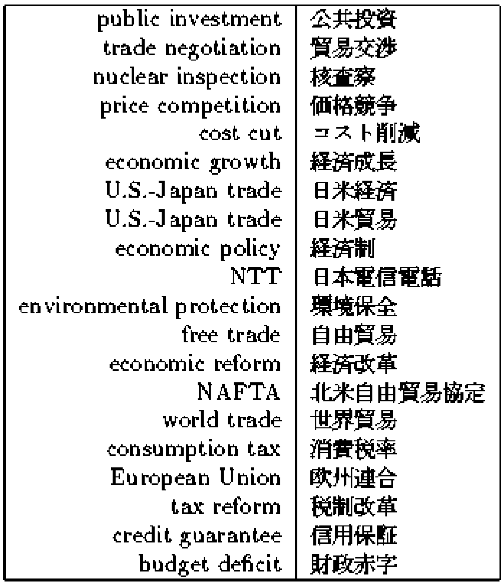
10.2. Matching Japanese terms to English

Evaluations are also carried out on the Wall Street Journal and Nikkei Financial News corpus, matching technical terms in Japanese to their counterparts in English. This evaluation is carried out in a worst-case scenario where (1) the two languages, English and Japanese, are across language groups; (2) the two texts, Wall Street Journal and Nikkei Financial News, do not focus on the same topics; (3) the two texts are not written by the same authors.

Seed words for this test are chosen according to the descriptions in Section 7. The 1,416 entries from the

Japanese/English online dictionary EDICT with occurrence frequencies between 100 and 1000. Since these seed words have relatively low to mid-range occurrence frequencies compared to the corpus size of around 7 million words for the WSJ text, we chose the segment size to be that of an entire paragraph. For the same reason, the Cosine Measure is chosen as a matching function. We believe this represents a more realistic scenario of truly non-parallel, bilingual corpora.

For evaluation, we need to select a test set of known technical term translations. Since most of the articles from these two newspaper do not describe a common topic, among the large set of technical terms extracted from the Nikkei Financial News corpus, we selected a set we believe also occur in the English Wall Street Journal. We then hand-translated these terms into English and look them up in the Wall Street Journal text. Among these test sets, nineteen terms have their translations in Japanese. They are shown in Figure 11.



public investment	公共投資
trade negotiation	貿易交渉
nuclear inspection	核査察
price competition	価格競争
cost cut	コスト削減
economic growth	経済成長
U.S.-Japan trade	日米経済
U.S.-Japan trade	日米貿易
economic policy	経済制
NTT	日本電信電話
environmental protection	環境保全
free trade	自由貿易
economic reform	経済改革
NAFTA	北米自由貿易協定
world trade	世界貿易
consumption tax	消費税率
European Union	欧州連合
tax reform	税制改革
credit guarantee	信用保証
budget deficit	財政赤字

Figure 11: The 19 term test set for the WSJ/Nikkei corpus

Three evaluations were carried out. Test I tries to find the correct translation for each of the nineteen Japanese term amongst the nineteen English terms. A translation is correct if the top candidate is the right one. To increase the candidate numbers, test II is carried out on the nineteen Japanese terms with their English counterparts plus 293 other English terms. For each Japanese term in test II, there are now 312 possible candidates. A translation is counted correct if the right one is on the top. The third test set

III consists of the nineteen Japanese terms paired with their translations and 383 single English words in addition. A translation is counted as correct if this best candidate is the same as the translation in Figure 11. The accuracies for the three test sets are shown in the table in Figure 12. The precision for these three tests range from 52.6% to 21.1%. We notice that many of the incorrect term translations are terms or words which overlap with the source term. For example, *economic reform* mapped to *tax reform.*

Test set	I (19)	II (312)	III (402)
Precision of best candidate	10/19=52.6%	4/19=21.1%	6/19=31.6%

Figure 12: Precisions for the best candidate translation in the WSJ/Nikkei corpus

Despite the low precision, we are still interested in seeing the ranking of the *true translations* among all the candidates for all nineteen cases for the purpose of a possible translator-aid. The result showing the ranking of the correct translation is in Figure 13. It shows that most of the correct translations can be found among the top 20 candidates.

Since these nineteen terms are first randomly chosen from the list of technical terms from the Nikkei text, with the only constraint being their English counterpart does not occur too few times in the WSJ text, and since we are computing the accuracy of the best translation candidate for each of the nineteen terms independently of each other, we consider the evaluation on these nineteen terms to be representative of an expected case scenario.

10.3. Translator-aid results

The previous two evaluations show that the precision of best-candidate translation using our algorithm is around 30% on average. While it is far from ideal, this is the first result of terminology translation from non-parallel corpora. Meanwhile, we have found that the correct translation is often found among the top 20 candidates. This leads us to conjecture that the output from this algorithm can be used as a translator-aid.

To evaluate this, we again chose the nineteen English/Japanese terms from the WSJ/Nikkei non-parallel corpus as a test set. We chose three evaluators who are all native Chinese speakers with bilingual knowledge in English and Chinese. Chinese speakers are able to recognize most Japanese technical terms since they

English translation	rank in test I (19)	rank in test II (312)	rank in test III (402)
public investment	14	128	61
trade negotiation	4	8	5
nuclear inspection	12	139	76
price competition	11	54	16
cost cut	2	2	2
economic growth	2	2	2
U.S.-Japan trade	4	17	5
U.S.-Japan trade	4	17	1
economic policy	1	1	5
NTT	4	18	5
environmental protection	4	17	5
free trade	4	16	83
economic reform	12	139	1
NAFTA	1	1	1
world trade	1	4	1
consumption tax	1	4	1
European Union	1	1	1
tax reform	1	1	1
credit guarantee	3	11	4
budget deficit	1	1	2

Figure 13: Rank of the correct translations for WSJ/Nikkei evaluations

are very similar to Chinese. We ask them to translate these nineteen Japanese terms into English. This is akin to asking Italian speakers to translate French technical terms into English [2]. The translators have some general knowledge of international news. However, none of them specializes in economics or finance, which is the domain of the WSJ/Nikkei corpus. The translators are asked to translate these nineteen terms from Japanese to English first, without using dictionaries or any other reference material. Their output is in SET A. Our system then propose two sets of outputs: (1) for each Japanese term, our system proposes the top-20 candidates from the set of 312 noun phrases. Using this candidate list, the translators again translate the nineteen terms. Their output based on this set is in SET B; (2) for each Japanese term, our system proposes the top-20 candidates from the set containing 383 single words plus the nineteen terms. The result of human translation based on this candidate list is in SET C. SET A, B and C are all compared to the original translation in the corpus. If the translation is the same as in the corpus, then it is judged as correct.

[2] Although Japanese and Chinese share common or very similar technical terms and words, the languages themselves are more different than, say, Italian and French.

The result is in Figure 14. It shows that evaluators on average are able to translate 8 terms out of 19 by themselves, whereas they can translate 18 terms on average with aid. This evaluation shows that translation precision increased from 42

Translator	SET A	without aid	SET B		SET C		increase
A	10/19	52.6%	18/19	94.7%	18/19	94.7%	42.1%
B	7/19	36.8%	17/19	89.5%	18/19	94.7%	57.9%
C	7/19	36.8%	17/19	89.5%	17/19	89.5%	52.7%
Average	8/19	42.1%	17.3/19	91.2%	17.7/19	93.0%	50.9%

Figure 14: Translator improvement on term translation

All three evaluators had no difficulty understanding the Japanese terms, even though some characters are different from that in Chinese. For example, the character for *America, U.S.* in Japanese is different from that in Chinese. Yet, all evaluators are able to understand the Japanese term for *U.S.-Japan trade.* Their translations for this term however vary from *Japanese-American trade* to *Japan-U.S. tradings.* Another interesting case is the term *cost cut* in Japanese which contains *Katakana* unrecognizable to Chinese speakers. However, the translators are able to select the correct English translation from the proposed list because they recognize the Chinese character for *cut* in the Japanese term.

These evaluations justified our conjecture about the usefulness of our system output as translator-aid.

11. CONCLUSION

We have described a statistical word signature feature, the Word Relation Matrix, that can be extracted from monolingual texts, and can be used to find matching pairs of content words or terms from lists of technical words and terms in a pair of same-domain non-parallel bilingual texts. The evaluation results of using this statistical feature alone for terminology translation show a precision of about 30%. Even though this result pales next to the high precisions obtained from parallel corpora, it is nevertheless quite useful output. In fact, we showed that humans are able to translate twice as many Japanese technical terms into English when our system output is used, compared to their own effort on translating a random set of 19 Japanese terms. It is also a significant initial result for lexical translation from truly non-parallel corpora, especially across language groups.

We have also brought up the issue of finding reliable seed word pairs from common online dictionaries as anchors for the World Relation Feature. We propose that rather than using one-to-one seed word pairs, multiple candidates of a dictionary entry should all be considered. To increase the reliability of seed words, only content words such as nouns, verbs, and adjectives should be used. In addition, to reduce the chance of the seed words being polysemous, we propose using only dictionary entries where the source word and the candidate words all appear only once.

In addition to the evaluation results, we have made the following important discoveries:

- The content words in the same segment with a word or term all contribute to the occurrence of this word. This feature represents some of the long-distance relations between the word and multiple other words which are not its immediate neighbors.
- The occurernce relationship between a content word or term with a list of anchor point seed words is consistent across languages in the same domain. If A is closely correlated to a set of words $B_1, B_2, \ldots, B_n$ to *varying degrees*, then its counterpart in another language, A', is closely correlated to a set of words $B_1, B_2, \ldots, B_n$ to *similar* varying degrees.

The information from the first discovery implies that the Word Relation Matrix can be used in language modeling in addition to the currently popular N-gram models and word trigger pairs. The second discovery implies that the Word Relation Matrix feature can be used as part of a system for matching technical terms extracted from same-domain non-parallel texts.

12. DISCUSSION

Various improvements are still needed both on the Word Relation Matrix feature, and on the system algorithm for translating technical terminologies from non-parallel corpora.

The dimensionality of WoRM vectors we have chosen is not optimal. Just as in text categorization tasks using content words, or sense disambiguation tasks using context words (Shütze, 1992; Yarowsky, 1995), a high dimensionality of vectors is favorable (Gale and Church, 1994). In our case, this means that if we have more seed word pairs, then we can be more sure of the translation of an unknown word or term. On the other hand, high dimensionality can also lead to noise, in addition to computational complexity. Therefore, dimensionality reduction methods such as the Singular Value Decomposition (Shütze, 1992) or clustering is often used. In our case, this means that we should choose a subset of highly discriminative seed word pairs. It is conceivable to use a Maximum Likelihood training algorithm to select the best seed words according to their discriminative power for a known set of bilingual lexical pairs. Perhaps individual seed words can be given different weights after training.

To improve the overall system performance, the Word Relation Matrix could be used in combination with other word signature features for non-parallel corpora. Word Relation Matrix might be sensitive to certain characteristics of words and terms, while oblivious about others. Other features may focus on these other characteristics.

Finally, finding non-parallel corpora is an issue. We suggested that monolingual texts of the same domain are easier to come by than parallel corpora of translated texts.

In our, due to our limitations in time and human power, we were only able to use one realistic non-parallel corpus, namely the Wall Street Journal/Nikkei Financial News material from the same time period. However, this corpus belongs to the most non-parallel category where there is not a large overlap of topics. In real applications of domain term translation, one might wish to choose from pairs of texts with closer domain resemblance.

Some possibilities include the following:

- Same domain texts from MULTEXT (Multilingual text tools and corpora), a project funded in the Commission of European Communities Linguistic Research and Engineering Program. This corpus consists of 2 million words per language from six languages (English, French, Germain, Italian, Spanish, Dutch), composed of comparable types of texts from two or three different domains.

- Part of the ECI/MCI Corpus 1 (European Corpus Initiative Multilingual Corpus 1) which contains approximately 97 million words in 27 (mainly European) languages. We plan to use newspaper texts from the same time period in different languages, assuming that the newspapers report on similar topics in the same time period.

- Wall Street Journal articles from various time periods, and part of the Nihon Kezai Shimbun texts consisting of 30 million words from the largest Japanese financial news daily newspaper. These two corpora can be used in conjunction as a non-parallel corpus. The latter is available from the Linguistic Data Consortium.

- The AP Newswire material in English and French from the same period, to form a non-parallel news domain corpus.
- In the long term, perhaps the most promising resource will be the various news magazines and home pages online, such as TimesOnLine, PC magazine, Financial News. This type of online material can be easily found by using any Internet search engine, such as Yahoo. Web material have almost all been classified by the search engines, which greatly facilitates the collection of multilingual material of specific domains. The popularity of the World Wide Web makes online data more and more accessible to individual researchers. Scripts can be easily written to download daily news, home pages and collect them over time to form large corpora. It is also easy to collect news material from a specific period of time as most of these news Web sites have time-stamped, archival material.

It is clear that we have just begun the work in the new area of terminology translation using non-parallel corpora. There are still a lot to be done in the future and many potential applications to be explored by using statistical word signature features.

REFERENCES

A. Boostein. 1983. Explanation and generalization of vector models in information retrieval. In *Proceedings of the 6th Annual International Conference on Research and Development in Information Retrieval*, pages 118–132.

P. Brown, J. Lai, and R. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics.*

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Stanley Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, June.

Kenneth Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 1–8, Columbus, Ohio, June.

W. Bruce Croft. 1984. A comparison of the cosine correlation and the modified probabilistic model. In *Information Technology*, volume 3, pages 113–114.

Ido Dagan and Kenneth W. Church. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 34–40, Stuttgart, Germany, October.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. In *Computational Linguistics*, pages 564–596.

Ido Dagan. 1990. Two languages are better than one. In *Proceedings of the 28th Annual Conference of the Association for Computational Linguistics*, Berkeley, California.

Pascale Fung and Dekai Wu. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, pages 69–85, Kyoto, Japan, June.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

William A. Gale and Kenneth W. Church. 1994. Discrimination decisions in 100,000 dimensional spaces. *Current Issues in Computational Linguisitcs: In honour of Don Walker*, pages 429–550.

W. Gale, K. Church, and D. Yarowsky. 1992a. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th Conference of the Association for Computational Linguistics.* Association for Computational Linguistics.

W. Gale, K. Church, and D. Yarowsky. 1992b. Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of TMI 92.*

W. Gale, K. Church, and D. Yarowsky. 1992c. Work on statistical methods for word sense disambiguation. In *Proceedings of AAAI 92.*

W. Gale, K. Church, and D. Yarowsky. 1993. A method for disambiguating word senses in a large corpus. In *Computers and Humanities*, volume 26, pages 415–439.

M. Hearst. 1991. Noun homograph disambiguation using local context in large text corpora. In *Using Corpora*, Waterloo, Canada.

Robert Korfhage. 1995. Some thoughts on similarity measures. In *The SIGIR Forum*, volume 29, page 8.

I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, Boston, Massachusettes.

Frederick Mosteller and David L. Wallace. 1968. *Applied Bayesian and Classical Inference - The Case of The Federalist Papers.* Springer Series in Satistics, Springer-Verlag.

Isadore Pinchuck. 1977. *Scientific and technical translation.* Andre Deutsch.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 35th Conference of the Association of Computational Linguistics, student session*, pages 321–322, Boston, Mass.

Juan C. Sager. 1990. *A Practical Course in Terminology Processing.* John Benjamins B.V.

G. Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.

Hinrich Shütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92.*

M. Simard, G Foster, and P. Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Forth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada.

Frank Smadja, Kathleen McKeown, and Vasileios Hatzsivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 21(4):1–38.

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

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proceedings of COLING 96*, Copenhagan, Danmark.

J.T. Tou and R.C Gonzalez. 1974. *Pattern Recognition Principles.* Addison-Wesley Publishing Company.

Howard R. Turtle and W. Bruce Croft. 1992. A comparison of text retrieval methods. In *The Computer Journal*, volume 35, pages 279–290.

Dekai Wu and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213, Columbia, Maryland, October.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, pages 80–87, Las Cruces, New Mexico, June.

D. Yarowsky. 1993. One sense per collocation. In *Proceedings of ARPA Human Languag Technology Workshop.*

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Conference of the Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics.