Term-list Translation using Mono-lingual Word Co-occurrence Vectors*

Genichiro Kikui
NTT Information and Communication Systems Labs.
1-1 Hikarinooka, Yokosuka-Shi, Kanagawa, Japan
e-mail: kikui@isl.ntt.co.jp

Abstract

A term-list is a list of content words that characterize a consistent text or a concept. This paper presents a new method for translating a term-list by using a corpus in the target language. The method first retrieves alternative translations for each input word from a bilingual dictionary. It then determines the most 'coherent' combination of alternative translations, where the coherence of a set of words is defined as the proximity among multi-dimensional vectors produced from the words on the basis of co-occurrence statistics. The method was applied to term-lists extracted from newspaper articles and achieved 81% translation accuracy for ambiguous words (i.e., words with multiple translations).

1 Introduction

A list of content words, called a term-list, is widely used as a compact representation of documents in information retrieval and other document processing. Automatic translation of term-lists enables this processing to be cross-linguistic. This paper presents a new method for translating term-lists by using co-occurrence statistics in the target language.

Although there is little study on automatic translation of term-lists, related studies are found in the area of target word selection (for content words) in conventional full-text machine translation (MT).

Approaches for target word selection can be classified into two types. The first type, which has been adopted in many commercial MT systems, is based on hand assembled disambiguation rules, and/or dictionaries. The problem with this approach is that creating these rules requires much cost and that they are usually domain-dependent [1].

The second type, called the statistics-based approach, learns disambiguation knowledge from large corpora. Brown et al. presented an algorithm that relies on translation probabilities estimated from large bilingual corpora (Brown et al., 1990)(Brown et al., 1991). Dagan and Itai (1994) and Tanaka and Iwasaki (1996) proposed algorithms for selecting target words by using word co-occurrence statistics in the target language corpora. The latter algorithms using mono-lingual corpora are particularly important because, at present, we cannot always get a sufficient amount of bilingual or parallel corpora.

Our method is closely related to (Tanaka and Iwasaki, 1996) from the viewpoint that they both rely on mono-lingual corpora only and do not require any syntactic analysis. The difference is that our method uses "coherence scores", which can capture associative relations between two words which do not co-occur in the training corpus.

This paper is organized as follows, Section 2 describes the overall translation process. Section 3 presents a disambiguation algorithm, which is the core part of our translation method. Section 4 and 5 give experimental results and discussion.

2 Term-list Translation

Our term-list translation method consists of two steps called *Dictionary Lookup* and *Disambiguation*.

1. Dictionary Lookup:
   For each word in the given term-list, all the alternative translations are retrieved from a bilingual dictionary.
   A translation candidate is defined as a combination of one translation for each input word. For example, if the input term-list consists of two words, say w_1 and w_2, and their translation include w_{11} for w_1 and w_{23} for w_2, then (w_{11}, w_{23}) is a translation candidate. If w_1 and w_2 have two and three alternatives respectively then there are 6 possible translation candidates.
2. Disambiguation:
   In this step, all possible translation candidates are ranked according to a measure that reflects the 'coherence' of each candidate. The top ranked candidate is the translated term-list.

* This research was done when the author was at Center for the Study of Language and Information(CSLI), Stanford University.

[1] In fact, this is partly shown by the fact that many MT systems have substitutable domain-dependent (or "user") dictionaries .

In the following sections we concentrate on the disambiguation step.

3 Disambiguation Algorithm

The underlying hypothesis of our disambiguation method is that a plausible combination of translation alternatives will be semantically coherent.

In order to find the most coherent combination of words, we map words onto points in a multidimensional vector space where the 'proximity' of two vectors represents the level of coherence of the corresponding two words. The coherence of n words can be defined as the order of spatial 'concentration' of the vectors.

The rest of this section formalizes this idea.

3.1 Co-occurrence Vector Space: WORD SPACE

We employed a multi-dimensional vector space, called WORD SPACE (Schuetze, 1997) for defining the coherence of words. The starting point of WORD SPACE is to represent a word with an n-dimensional vector whose i-th element is how many times the word w_i occurs close to the word. For simplicity, we consider w_i and w_j to occur close in context if and only if they appear within an m-word distance (i.e., the words occur within a window of m-word length), where m is a predetermined natural number.

Table 1 shows an artificial example of co-occurrence statistics. The table shows that the word *ginko* (bank, where people deposit money) co-occurred with *shikin* (fund) 483 times and with *hashi* (bridge) 31 times. Thus the co-occurrence vector of *ginko* (money bank) contains 483 as its 89th element and 31 as its 468th element. In short, a word is mapped onto the row vector of the co-occurrence table (matrix).

Table 1: An example of co-occurrence statistics.

col. no. *word* (Eng.)	...	89 *shikin* (fund)	...	468 *hashi* (bridge)
ginko (bank:money)	...	483		31
teibo (bank:river)	...	8	...	120

Using this word representation, we define the proximity, *prox*, of two vectors, $\vec{a}, \vec{b}$, as the cosine of the angle between them, given as follows.

$$prox(\vec{a}, \vec{b}) = (\vec{a} \bullet \vec{b})/(\mid \vec{a} \mid\mid \vec{b} \mid) \quad (1)$$

If two vectors have high proximity then the corresponding two words occur in similar context, and in our terms, are coherent.

This simple definition, however, has problems, namely its high-dimensionality and sparseness of data. In order to solve these problems, the original co-occurrence vector space is converted into a condensed low dimensional real-valued matrix by using SVD (Singular Value Decomposition). For example, a 20000-by-1000 matrix can be reduced to a 20000-by-100 matrix. The resulting vector space is the WORD SPACE [2].

3.2 Coherence of Words

We define the *coherence* of words in terms of a geometric relationship between the corresponding word vectors.

As shown above, two vectors with high proximity are coherent with respect to their associative properties. We have extended this notion to n-words. That is, if a group of vectors are concentrated, then the corresponding words are defined to be coherent. Conversely, if vectors are scattered, the corresponding words are in-coherent. In this paper, the concentration of vectors is measured by the average proximity from their centroid vector.

Formally, for a given word set W, its coherence $coh(W)$ is defined as follows:

$$coh(W) = \frac{1}{\mid W \mid} \sum_{w \in W} prox(\vec{v}(w), \vec{c}(W)) \quad (2)$$

$$\vec{c}(W) = \sum_{w \in W} \vec{v}(w) \quad (3)$$

$$\mid W \mid = \textit{the number of words in } W \quad (4)$$

3.3 Disambiguation Procedure

Our disambiguation procedure is simply selecting the combination of translation alternatives that has the largest $coh(W)$ defined above. The current implementation exhaustively calculates the coherence score for each combination of translation alternatives, then selects the combination with the highest score.

3.4 Example

Suppose the given term-list consists of *bank* and *river*. Our method first retrieves translation alternatives from the bilingual dictionary. Let the dictionary contain following translations.

[2]The WORD SPACE method is closely related to Latent Semantic Indexing (LSI)(Deerwester et al., 1990), where document-by-word matrices are processed by SVD instead of word-by-word matrices. The difference between these two is discussed in (Schuetze and Pedersen, 1997).

source		translations
bank	→	ginko (bank:money), teibo(bank:river)
interest	→	rishi (interest:money), kyoumi(interest:feeling)

Combining these translation alternatives yields four translation candidates:

$(ginko, risoku), (ginko, kyoumi),$
$(teibo, risoku), (teibo, kyoumi).$

Then the coherence score is calculated for each candidate.

Table 2 shows scores calculated with the co-occurrence data used in the translation experiment (see. Section 4.4.2). The combination of *ginko* (bank:money) and *risoku*(interest:money) has the highest score. This is consistent with our intuition.

Table 2: An example of scores

rank	candidate	score (*coh*)
1	$(ginko, risoku)$	0.930
2	$(teibo, kyoumi)$	0.897
3	$(ginko, kyoumi)$	0.839
4	$(teibo, risoku)$	0.821

4 Experiments

We conducted two types of experiments: re-translation experiments and translation experiments. Each experiment includes comparison against the baseline algorithm, which is a unigram-based translation algorithm. This section presents the two types of experiments, plus the baseline algorithm, followed by experimental results.

4.1 Two Types of Experiments

4.1.1 Translation Experiment

In the translation experiment, term-lists in one language, e.g., English, were translated into another language, e.g., in Japanese. In this experiment, humans judged the correctness of outputs.

4.1.2 Re-translation Experiment

Although the translation experiment recreates real applications, it requires human judgment [3]. Thus we decided to conduct another type of experiment, called a re-translation experiment. This experiment translates given term-lists (e.g., in English) into a second language (e.g., Japanese) and maps them back onto the source language (e.g., in this case, English). Thus the correct translation of a term list, in the most strict sense, is the original term-list itself.

[3] If a bilingual parallel corpus is available, then corresponding translations could be used for correct results.

This experiment uses two bilingual dictionaries: a forward dictionary and a backward dictionary.

In this experiment, a word in the given term-list (e.g. in English) is first mapped to another language (e.g., Japanese) by using the forward dictionary. Each translated word is then mapped back into original language by referring to the backward dictionary. The union of the translations from the backward dictionary are the translation alternatives to be disambiguated.

4.2 Baseline Algorithm

The baseline algorithm against which our method was compared employs unigram probabilities for disambiguation. For each word in the given term-list, this algorithm chooses the translation alternative with the highest unigram probability in the target language. Note that each word is translated independently.

4.3 Experimental Data

The source and the target languages of the translation experiments were English and Japanese respectively. The re-translation experiments were conducted for English term-lists using Japanese as the second language.

The Japanese-to-English dictionary was EDICT(Breen, 1995) and the English-to-Japanese dictionary was an inversion of the Japanese-to-English dictionary.

The co-occurrence statistics were extracted from the 1994 New York Times (420MB) for English and 1990 Nikkei Shinbun (Japanese newspaper) (150MB) for Japanese. The domains of these texts range from business to sports. Note that 400 articles were randomly separated from the former corpus as the test set.

The initial size of each co-occurrence matrix was 20000-by-1000, where rows and columns correspond to the 20,000 and 1000 most frequent words in the corpus[4]. Each initial matrix was then reduced by using SVD into a matrix of 20000-by-100 using SVDPACKC(Berry et al., 1993).

Term-lists for the experiments were automatically generated from texts, where a term-list of a document consists of the topmost n words ranked by their tf-idf scores [5]. The relation between the length n of term-list and the disambiguation accuracy was also tested.

We prepared two test sets of term-lists: those extracted from the 400 articles from the New York Times mentioned above, and those extracted from

[4] Stopwords are ignored.

[5] The tf-idf score of a word w in a text is $tf_w log(\frac{N}{N_w})$, where $tf_w$is the occurrence of w in the text, N is the number of documents in the collection, and N_w is the number of documents containing w.

articles in Reuters(Reuters, 1997), called Test-NYT, and Test-REU, respectively.

4.4 Results

4.4.1 re-translation experiment

The proposed method was applied to several sets of term-lists of different length. Results are shown in Table 3. In this table and the following tables, "ambiguous" and "success" correspond to the total number of ambiguous words, not term-lists, and the number of words that were successfully translated[6]. The best results were obtained when the length of term-lists was 4 or 6. In general, the longer a term-list becomes, the more information it has. However, a long term-list tends to be less coherent (i.e., contain different topics). As far as our experiments are concerned, 4 or 6 was the point of compromise.

Table 3: Result of Re-translation for Test-NYT

length	success/ambiguous	(rate)
2	98/141	(69.5%)
4	240/329	(72.9%)
6	410/555	(73.8%)
8	559/777	(71.9%)
10	691/981	(70.4%)
12	813/1165	(69.8%)

Then we compared our method against the baseline algorithm that was trained on the same set of articles used to create the co-occurrence matrix for our algorithm (i.e., New York Times). Both are applied to term-lists of length 6 made from test-NYT. The results are shown in Table 4. Although the absolute value of the success rate is not satisfactory, our method significantly outperforms the baseline algorithm.

Table 4: Result of Re-translation for Test-NYT

Method	success/ambiguous (rate)
baseline	236/555 (42.5%)
proposed	410/555 (73.8%)

We, then, applied the same method with the same parameters (i.e., cooccurence and unigram data) to Test-REU. As shown in Table 5, our method did better than the baseline algorithm although the success rate is lower than the previous result.

Table 5: Result of re-translation for Test-REU

Method	success/ambiguous (rate)
baseline	162/565 (28.7%)
proposed	351/565 (62.1%)

[6]If 100 term-lists were processed and each term-list contains 2 ambiguous words, then the "total" becomes 200.

Table 6: Result of Translation for Test-NYT

Method	success/ambiguous (rate)
baseline	74/125 (72.6%)
proposed	101/125 (80.8%)

4.4.2 translation experiment

The translation experiment from English to Japanese was carried out on Test-NYT. The training corpus for both proposed and baseline methods was the Nikkei corpus described above. Outputs were compared against the "correct data" which were manually created by removing incorrect alternatives from all possible alternatives. If all the translation alternatives in the bilingual dictionary were judged to be correct, then we counted this word as unambiguous.

The accuracy of our method and baseline algorithm are shown on Table6.

The accuracy of our method was 80.8%, about 8 points higher than that of the baseline method. This shows our method is effective in improving translation accuracy when syntactic information is not available. In this experiment, 57% of input words were unambiguous. Thus the success rates for entire words were 91.8% (proposed) and 82.6% (baseline).

4.5 Error Analysis

The following are two major failure reasons relevant to our method [7].

The first reason is that alternatives were semantically too similar to be discriminated. For example, "*share*" has at least two Japanese translations: "*shea*"(market share) and "*kabu*"(stock). Both translations frequently occur in the same context in business articles, and moreover these two words sometimes co-occur in the same text. Thus, it is very difficult to discriminate them. In this case, the task is difficult also for humans unless the original text is presented.

The second reason is more complicated. Some translation alternatives are polysemous in the target language. If a polysemous word has a very general meaning that co-occurs with various words, then this word is more likely to be chosen. This is because the corresponding vector has "average" value for each dimension and, thus, has high proximity with the centroid vector of multiple words.

For example, alternative translations of "*stock*" includes two words: "*kabu*" (company share) and "*dashi*" (liquid used for food). The second translation "*dashi*" is also a conjugation form of the Japanese verb "*dasu*", which means "put out" and "start". In this case, the word, "*dashi*", has a cer-

[7]Other reasons came from errors in pre-processing including 1) ignoring compound words, 2) incorrect handling of capitalized words etc.

tain amount of proximity because of the meaning irrelevant to the source word, e.g., *stock*.

This problem was pointed out by (Dagan and Itai, 1994) and they suggested two solutions 1) increasing the size of the (mono-lingual) training corpora or 2) using bilingual corpora. Another possible solution is to resolve semantic ambiguities of the training corpora by using a mono-lingual disambiguation algorithm (e.g., (?)) before making the co-occurrence matrix.

5 Related Work

Dagan and Itai (1994) proposed a method for choosing target words using mono-lingual corpora. It first locates pairs of words in dependency relations (e.g., verb-object, modifier-noun, etc.), then for each pair, it chooses the most plausible combination of translation alternatives. The plausibility of a word-pair is measured by its co-occurence probability estimated from corpora in the target language.

One major difference is that their method relies on co-occurrence statistics between tightly and locally related (i.e., syntactically dependent) word pairs, whereas ours relies on associative properties of loosely and more globally related (i.e., co-occurring within a certain distance) word groups. Although the former statistics could provide more accurate information for disambiguation, it requires huge amounts of data to cover inputs (the data sparseness problem).

Another difference, which also relates to the data sparseness problem, is that their method uses "row" co-occurrence statistics, whereas ours uses statistics converted with SVD. The converted matrix has the advantage that it represents the co-occurrence relationship between two words that share similar contexts but do not co-occur in the same text[8]. SVD conversion may, however, weaken co-occurrence relations which actually exist in the corpus.

Tanaka and Iwasaki (1996) also proposed a method for choosing translations that solely relies on co-occurrence statistics in the target language. The main difference with our approach lies in the plausibility measure of a translation candidate. Instead of using a "coherence score", their method employs proximity, or inverse distance, between the two co-occurrence matrices: one from the corpus (in the target language) and the other from the translation candidate. The distance measure of two matrices given in the paper is the sum of the absolute distance of each corresponding element. This definition seems to lead the measure to be insensitive to the candidate when the co-occurrence matrix is filled with large numbers.

[8] "Second order co-occurrence". See (Schuetze, 1997)

6 Concluding Remarks

In this paper, we have presented a method for translating term-lists using mono-lingual corpora.

The proposed method is evaluated by translation and re-translation experiments and showed a translation accuracy of 82% for term-lists extracted from articles ranging from business to sports.

We are planning to apply the proposed method to cross-linguistic information retrieval (CLIR). Since the method does not rely on syntactic analysis, it is applicable to translating users' queries as well as translating term-lists extracted from documents.

A future issue is further evaluation of the proposed method using more data and various criteria including overall performance of an application system (e.g., CLIR).

Acknowledgment

I am grateful to members of the Infomap project at CSLI, Stanford for their kind support and discussions. In particular I would like to thank Stanley Peters and Raymond Flournoy.

References

M.W. Berry, T. Do, G. O'Brien, V. Krishna, and S. Varadhan. 1993. *SVDPACKC USER'S GUIDE.* Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN,.

J.W. Breen. 1995. *EDICT, Freeware, Japanese-to-English Dictionary.*

P. Brown, J. Cocke, V. Della Pietra, F. Jelinek, R.L. Mercer, and P. C. Roosin. 1990. A statistical approach to language translation. *Computational Linguistics*, 16(2).

P. Brown, V. Della Pietra, and R.L. Mercer. 1991. Word sense disambiguation using statisical methods. In *Proceedings of ACL-91.*

I. Dagan and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics.*

S. Deerwester, S.T. Dumais, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of American Society for Information Science.*

Reuters. 1997. *Reuters-21578, Distribution 1.0.* available at http://www.research.att.com/~lewis.

H. Schuetze and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management.*

H. Schuetze. 1997. *Ambiguity Resolution in Language Learning.* CSLI.

K. Tanaka and H. Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proceedings of COLING-96.*