Lexical transfer using a vector-space model

Eiichiro SUMITA
ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika, Soraku
Kyoto 619-0288, Japan
sumita@slt.atr.co.jp

Abstract

Building a bilingual dictionary for transfer in a machine translation system is conventionally done by hand and is very time-consuming. In order to overcome this bottleneck, we propose a new mechanism for lexical transfer, which is simple and suitable for learning from bilingual corpora. It exploits a vector-space model developed in information retrieval research. We present a preliminary result from our computational experiment.

Introduction

Many machine translation systems have been developed and commercialized. When these systems are faced with unknown domains, however, their performance degrades. Although there are several reasons behind this poor performance, in this paper, we concentrate on one of the major problems, i.e., building a bilingual dictionary for transfer.

A bilingual dictionary consists of rules that map a part of the representation of a source sentence to a target representation by taking grammatical differences (such as the word order between the source and target languages) into consideration. These rules usually use case-frames as their base and accompany syntactic and/or semantic constraints on mapping from a source word to a target word.

For many machine translation systems, experienced experts on individual systems compile the bilingual dictionary, because this is a complicated and difficult task. In other words, this task is knowledge-intensive and labor-intensive, and therefore, time-consuming.

Typically, the developer of a machine translation system has to spend several years building a general-purpose bilingual dictionary. Unfortunately, such a general-purpose dictionary is not almighty, in that (1) when faced with a new domain, unknown source words may emerge and/or some domain-specific usages of known words may appear and (2) the accuracy of the target word selection may be insufficient due to the handling of many target words simultaneously.

Recently, to overcome these bottlenecks in knowledge building and/or tuning, the automation of lexicography has been studied by many researchers: (1) approaches using a *decision tree*: the ID3 learning algorithm is applied to obtain transfer rules from case-frame representations of simple sentences with a thesaurus for generalization (Akiba *et. al.*, 1996 and Tanaka, 1995); (2) approaches using *structural matching*: to obtain transfer rules, several search methods have been proposed for maximal structural matching between trees obtained by parsing bilingual sentences (Kitamura and Matsumoto, 1996; Meyers *et. al.*, 1998; and Kaji *et. al.*,1992).

1 Our proposal

1.1 Our problem and approach

In this paper, we concentrate on lexical transfer, i.e., target word selection. In other words, the mapping of structures between source and target expressions is not dealt with here. We assume that this structural transfer can be solved on top of lexical transfer.

We propose an approach that differs from the studies mentioned in the introduction section in that:

I) It use not structural representations like *case frames* but *vector-space* representations.

II) The weight of each element for constraining the ambiguity of target words is determined automatically by following the *term frequency* and *inverse document frequency* in information retrieval research.

III) A word alignment that does not rely on parsing is utilized.

IV) Bilingual corpora are clustered in terms of target equivalence.

1.2 Background

The background for the decisions made in our approach is as follows:

A) We would like to reduce human interaction to prepare the data necessary for building lexical transfer rules.

B) We do not expect that mature parsing systems for multi-languages and/or spoken languages will be available in the near future.

C) We would like the determination of the importance of each feature in the target selection to be automated.

D) We would like the problem caused by errors in the corpora and data sparseness to be reduced.

2 Vector-space model

This section explains our trial for applying a *vector-space model* to lexical transfer starting from a basic idea.

2.1 Basic idea

We can select an appropriate target word for a given source word by observing the environment including the context, world knowledge, and target words in the neighborhood. The most influential elements in the environment are of course the other words in the source sentence surrounding the concerned source word.

Suppose that we have translation examples including the concerned source word and we know in advance which target word corresponds to the source word.

By measuring the similarity between (1) an unknown sentence that includes the concerned source word and (2) known sentences that include the concerned source word, we can select the target word which is included in the most similar sentence.

This is the same idea as example-based machine translation (Sato and Nagao, 1990 and Furuse *et. al.*, 1994).

Group1: 辛口 (not sweet)
source sentence 1: This beer is **drier** and full-bodied. target sentence 1: □□□□□□□□**辛口**□□□□□□□□ source sentence 2: Would you like **dry** or sweet sherry? target sentence 2: **辛口**□□□□□□□□□□□□□□□□□□□□□□□□ source sentence 3: A **dry** red wine would go well with it. target sentence 3: □□□□**辛口**□□□□□□□□□□□□
Group2: 乾燥 (not wet)
source sentence 4: Your skin feels so **dry**. target sentence 4: □□□□□**乾燥**□□□□□ source sentence 5: You might want to use some cream to protect your skin against the **dry** air. target sentence 5: **乾燥**□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Table 1 Portions of English “dry” into Japanese for an aligned corpus

Listed in Table 1 are samples of English-Japanese sentence pairs of our corpus including the source word “dry.” The upper three samples of group 1 are translated with the target word “辛口 (not sweet)” and the lower two samples of group 2 are translated with the target word “乾燥 (not wet).” The remaining portions of target sentences are hidden here because they do not relate to the discussion in the paper. The underlined words are some of the cues used to select the target words. They are distributed in the source sentence with several different grammatical relations such as *subject*, *parallel adjective*, *modified noun*, and so on, for the concerned word “dry.”

2.2 Sentence vector

We propose representing the sentence as a **sentence vector**, i.e., a vector that lists all of the words in the sentence. The sentence vector of the first sentence of Table 1 is as follows:

<this, beer, is, dry, and, full-body>

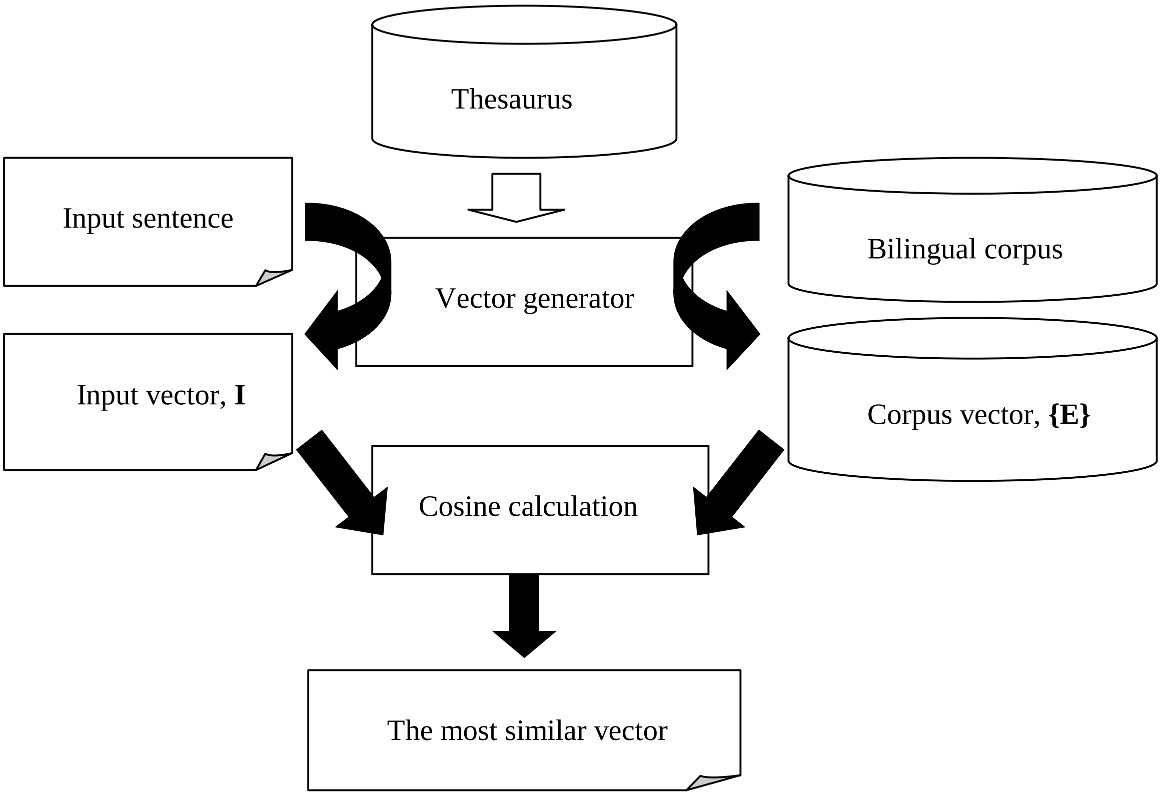

Figure 1 System Configuration

Figure 1 outlines our proposal. Suppose that we have the sentence vector of an input sentence **I** and the sentence vector of an example sentence **E** from a bilingual corpus.

We measure the similarity by computing the **cosine** of the angle between **I** and **E**.

We output the **target word** of the example sentence whose cosine is **maximal**.

2.3 Modification of sentence vector

The naïve implementation of a sentence vector that uses the occurrence of words themselves suffers from data sparseness and unawareness of relevance.

2.3.1 Semantic category incorporation

To reduce the adverse influence of data sparseness, we count occurrences by not only the words themselves but also by the semantic categories of the words given by a thesaurus. For example, the “辛口 (not sweet)” sentences of

Table 1 have the different cue words of “beer,” “sherry,” and “wine,” and the cues are merged into a single semantic category *alcohol* in the sentence vectors.

2.3.2 Grouping sentences and weighting dimensions

The previous subsection does not consider the relevance to the target selection of each element of the vectors; therefore, the selection may fail due to non-relevant elements.

We exploit the *term frequency* and *inverse document frequency* in information retrieval research. Here, we regard a *group* of sentences that share *the same target word* as a *document*.” Vectors are made not sentence-wise but group-wise. The relevance of each dimension is the *term frequency* multiplied by the *inverse document frequency*. The *term frequency* is the frequency in the document (group). A repetitive occurrence may indicate the importance of the word. The *inverse document frequency* corresponds to the discriminative power of the target selection. It is usually calculated as a logarithm of *N* divided by *df* where *N* is the number of the documents (groups) and *df* is the frequency of documents (groups) that include the word.

Cluster 1: a piece of paper money, C(**紙幣**)
source sentence 1: May I have change for a ten dollar **bill**? target sentence 1: □□□□□**紙幣**□□□□□□□□□□ source sentence 2: Could you change a fifty dollar **bill**? target sentence 2: □□□□札□□□□□□□□□□
Cluster 2: an account, C(**勘定**)
source sentence 3: I've already paid the **bill**. target sentence 3: □□**勘定**□□□□□□□ source sentence 4: Isn't my **bill** too high? target sentence 4: □□料金□□□□□□□□□□ source sentence 5: I'm checking out. May I have the **bill**, please? target sentence 5: □□□□□□□□□□会計□□□□□

Table 2 Samples of groups clustered by target equivalence

3 Pre-processing of corpus

Before generating vectors, the given bilingual corpus is pre-processed in two ways (1) words are aligned in terms of translation; (2) sentences are clustered in terms of target equivalence to reduce problems caused by data sparseness.

3.1 Word alignment

We need to have source words and target words aligned in parallel corpora. We use a word alignment program that does not rely on parsing (Sumita, 2000). This is not the focus of this paper, and therefore, we will only describe it briefly here.

First, all possible alignments are hypothesized as a matrix filled with occurrence similarities between source words and target words.

Second, using the occurrence similarities and other constraints, the most plausible alignment is selected from the matrix.

3.2 Clustering by target words

We adopt a clustering method to avoid the sparseness that comes from variations in target words.

The translation of a word can vary more than the meaning of the target word. For example, the English word “bill” has two main meanings: (1) a piece of paper money, and (2) an account. In Japanese, there is more than one word for each meaning. For (1), “札” and “紙幣” can correspond, and for (2), “勘定,” “会計,” and “料金” can correspond.

The most frequent target word can represent the cluster, e.g., “紙幣” for (1) a piece of paper money; “勘定” for (2) an account. We assume that *selecting a cluster is equal to selecting the target word*.

If we can merge such equivalent translation variations of target words into clusters, we can improve the accuracy of lexical transfer for two reasons: (1) doing so makes the mark larger by neglecting accidental differences among target words; (2) doing so collects scattered pieces of evidence and strengthens the effect.

Furthermore, word alignment as an automated process is incomplete. We therefore need to filter out erroneous target words that come from alignment errors. Erroneous target words are considered to be low in frequency and are expected to be semantically dissimilar from correct target words based on correct alignment. Clustering example corpora can help filter out erroneous target words.

By calculating the semantic similarity between the semantic codes of target words, we perform clustering according to the simple algorithm in subsection 3.2.2.

3.2.1 Semantic similarity

Suppose each target word has semantic codes for all of its possible meanings. In our thesaurus, for example, the target word “札” has three decimal codes, 974 (label/tag), 829 (counter) and 975 (money) and the target word “紙幣” has a single code 975 (money). We represent this as a code vector and define the similarity between the two target words by computing the cosine of the angle between their code vectors.

3.2.2 Clustering algorithm

We adopt a simple procedure to cluster a set of n target words $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$. $\boldsymbol{X}$ is sorted in the descending order of the frequency of X_n in a sub-corpus including the concerned source word.

We repeat (1) and (2) until the set $\boldsymbol{X}$ is empty.

(1) We move the leftmost X_l from $\boldsymbol{X}$ to the new cluster $C(X_l)$.

(2) For all m (m>l) , we move X_m from $\boldsymbol{X}$ to $C(X_l)$ if the cosine of X_l and X_m is larger than the threshold T.

As a result, we obtain a set of clusters $\{C(X_l)\}$ for each meaning as exemplified in Table 2.

The threshold of semantic similarity T is determined empirically. T in the experiment was 1/2.

4 Experiment

To demonstrate the feasibility of our proposal, we conducted a pilot experiment as explained in this section.

Number of sentence pairs (English-Japanese)	19,402
Number of source words (English)	156,128
Number of target words (Japanese)	178,247
Number of source content words (English)	58,633
Number of target content words (Japanese)	64,682
Number of source different content words (English)	4,643
Number of target different content words (Japanese)	6,686

Table 3 Corpus statistics

4.1 Experimental conditions

For our sentence vectors and code vectors, we used hand-made thesauri of Japanese and English covering our corpus (for a travel arrangement task), whose hierarchy is based on that of the Japanese commercial thesaurus *Kadokawa Ruigo Jiten* (Ohno and Hamanishi, 1984).

We used our English-Japanese phrase book (a collection of pairs of typical sentences and their translations) for foreign tourists. The statistics of the corpus are summarized in Table 3. We word-aligned the corpus before generating the sentence vectors.

We focused on the transfer of content words such as *nouns*, *verbs*, and *adjectives*. We picked out six polysemous words for a preliminary evaluation: “ bill, ” “ dry, ” “ call ” in English and “熱,” “悪い,” “飲む” in Japanese.

We confined ourselves to a *selection between two major clusters of each source word* using the method in subsection 3.2

	#1&2	#1	baseline	#correct	vsm
bill [noun]	47	30	64%	40	85%
call [verb]	179	93	52%	118	66%
dry [adjective]	6	3	50%	4	67%
熱 [noun]	19	13	68%	14	73%
飲む [verb]	60	42	70%	49	82%
悪い [adjective]	26	15	57%	16	62%

Table 4 Accuracy of the baseline and the VSM systems

4.2 Selection accuracy

We compared the accuracy of our proposal using the *vector-space model* (*vsm system*) with that of a *decision-by-majority model* (*baseline system*). The results are shown in Table 4.

Here, the accuracy of the baseline system is #1 (the number of target sentences of the most major cluster) divided by #1&2 (the number of target sentences of clusters 1 & 2). The accuracy of the vsm system is #correct (the number of vsm answers that match the target sentence) divided by #1&2.

	#all	#1&2	Coverage
bill [noun]	63	47	74%
call [verb]	226	179	79%
dry [adjective]	8	6	75%
熱 [noun]	22	19	86%
飲む [verb]	77	60	78%
悪い [adjective]	38	26	68%

Table 5 Coverage of the top two clusters

Judging was done mechanically by assuming that the aligned data was 100% correct.[1] Our vsm system achieved an accuracy from about 60% to about 80% and outperformed the baseline system by about 5% to about 20%.

4.3 Coverage of major clusters

One reason why we clustered the example database was to filter out noise, i.e., wrongly aligned words. We skimmed the clusters and we saw that many instances of noise were filtered out. At the same time, however, a portion of correctly aligned data was unfortunately discarded. We think that such discarding is not

[1] This does not necessarily hold, therefore, performance degrades in a certain degree.

fatal because the coverage of clusters 1&2 was relatively high, around 70% or 80% as shown in Table 5. Here, the coverage is #1&2 (the number of data not filtered) divided by #all (the number of data before discarding).

5 Discussion

5.1 Accuracy

An experiment was done for a restricted problem, i.e., *select the appropriate one cluster (target word) from two major clusters (target words)*, and the result was encouraging for the automation of the lexicography for transfer.

We plan to improve the accuracy obtained so far by exploring elementary techniques: (1) Adding new features including extra linguistic information such as the role of the speaker of the sentence (Yamada et al., 2000) (also, the topic that sentences are referring to) may be effective; and (2) Considering the physical distance from the concerned input word, which may improve the accuracy. A kind of *window function* might also be useful; (3) Improving the word alignment, which may contribute to the overall accuracy.

5.2 Data sparseness

In our proposal, deficiencies in the naïve implementation of vsm are compensated in several ways by using a thesaurus, grouping, and clustering, as explained in subsections 2.3 and 3.2.

5.3 Future work

We showed only the translation of content words. Next, we will explore the translation of function words, the word order, and full sentences.

Our proposal depends on a *handcrafted* thesaurus. If we manage to do without craftsmanship, we will achieve broader applicability. Therefore, automatic thesaurus construction is an important research goal for the future.

Conclusion

In order to overcome a bottleneck in building a bilingual dictionary, we proposed a simple mechanism for lexical transfer using a vector space.

A preliminary computational experiment showed that our basic proposal is promising. Further development, however, is required: to use a *window function* or to use a better alignment program; to compare other statistical methods such as *decision trees*, *maximal entropy*, and so on.

Furthermore, an important future work is to create a full translation mechanism based on this lexical transfer.

Acknowledgements

Our thanks go to Kadokawa-Shoten for providing us with the Ruigo-Shin-Jiten.

References

Akiba, O., Ishii, M., ALMUALLIM, H., and Kaneda, S. (1996) *A Revision Learner to Acquire English Verb Selection Rules*, Journal of NLP, 3/3, pp. 53-68, (in Japanese).

Furuse, O., Sumita, E. and Iida, H. (1994) *Transfer-Driven Machine Translation Utilizing Empirical Knowledge*, Transactions of IPSJ, 35/3, pp. 414-425, (in Japanese).

Kaji, H., Kida, Y. and Morimoto, Y. (1992) *Learning translation templates from bilingual text*, Proc. of Coling-92, pp. 672-678.

Kitamura, M. and Matsumoto, Y. (1996) Automatic *Acquisition of Translation Rules from Parallel Corpora*, Transactions of IPSJ, 37/6, pp. 1030-1040, (in Japanese).

Meyers, A., Yangarber, R., Grishman, R., Macleod, C., and Sandoval, A. (1998) Deriving *Transfer rules from dominance-preserving alignments*, Coling-ACL98, pp. 843-847.

Ohno, S. and Hamanishi, M. (1984) *Ruigo-Shin-Jiten*, Kadokawa, p. 932, (in Japanese).

Sato, S. and Nagao, M. (1990) *Toward memory-based translation*, Coling-90, pp. 247-252.

Sumita, E. (2000) *Word alignment using matrix* PRICAI-00, 2000, (to appear).

Tanaka H. (1995) *Statistical Learning of "Case Frame Tree" for Translating English Verbs*, Journal of NLP, 2/3, pp. 49-72, (in Japanese).

Yamada, S., Sumita, E. and Kashioka, H. (2000) *Translation using Information on Dialogue Participants*, ANLP-00, pp. 37-43.