Resolving Translation Ambiguity using Non-parallel Bilingual Corpora

Genichiro KIKUI
NTT Cyberspace Laboratories
1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847, JAPAN
e-mail: kikui@isl.ntt.co.jp

Abstract

This paper presents an unsupervised method for choosing the correct translation of a word in context. It learns disambiguation information from non-parallel bilingual corpora (preferably in the same domain) free from tagging.

Our method combines two existing unsupervised disambiguation algorithms: a word sense disambiguation algorithm based on distributional clustering and a translation disambiguation algorithm using target language corpora.

For the given word in context, the former algorithm identifies its meaning as one of a number of predefined usage classes derived by clustering a large amount of usages in the source language corpus. The latter algorithm is responsible for associating each usage class (i.e., cluster) with a target word that is most relevant to the usage.

This paper also shows preliminary results of translation experiments.

1 Introduction

Choosing the correct translation of a content word in context, referred to as "translation disambiguation (of content word)", is a key task in machine translation. It is also crucial in cross-language text processing including cross-language information retrieval and abstraction.

Due to the recent availability of large text corpora, various statistical approaches have been tried including using 1) parallel corpora (Brown et al., 1990), (Brown et al., 1991), (Brown, 1997), 2) non-parallel bilingual corpora tagged with topic area (Yamabana et al., 1998) and 3) un-tagged mono-language corpora in the target language (Dagan and Itai, 1994), (Tanaka and Iwasaki, 1996), (Kikui, 1998).

A problem with the first two approaches is that it is not easy to obtain sufficiently large parallel or manually tagged corpora for the pair of languages targeted.

Although the third approach eases the problem of preparing corpora, it suffers from a lack of useful information in the source language. For example, suppose the proper name, "Dodgers", provides good context to identify the usage of "hit" in the training corpus in English. If the translation of "Dodgers" rarely occurs in the target language corpora, it does not contribute to target word selection.

The method presented in this paper solves this problem by choosing the target word that corresponds to the usage identified in the source language corpora. This method is totally unsupervised in the sense that it acquires disambiguation information from non-parallel bilingual corpora (preferably in the same domain) free from tagging.

It combines two unsupervised disambiguation algorithms: one is the word sense disambiguation algorithm based on distributional clustering(Schuetze, 1997) and the other is the translation disambiguation algorithm using target language corpora(Kikui, 1998). For the given word in context, the former algorithm identifies its usage as one of several predefined usage classes derived by clustering a large amount of usages in the source language corpus. The latter algorithm is responsible for associating each usage class (i.e., cluster) with a target word that best expresses the usage.

The following sections are organized as follows. In Section 2, we overview the entire method. The following two sections (i.e., Section 3 and 4) then introduce the two major components of the method including the two unsupervised disambiguation algorithms. Section 5 and 6 are devoted respectively to a preliminary evaluation and discussions on related research.

2 Overview of the method

Figure 1 shows an overview of the entire method.

Components inside the dotted line on the left represent word-sense disambiguation (WSD) in the source language. There are two sub-processes: *distributional clustering* and *categorizing*. The former automatically identifies different usages (or senses) of the given source word (shown at top center) in the

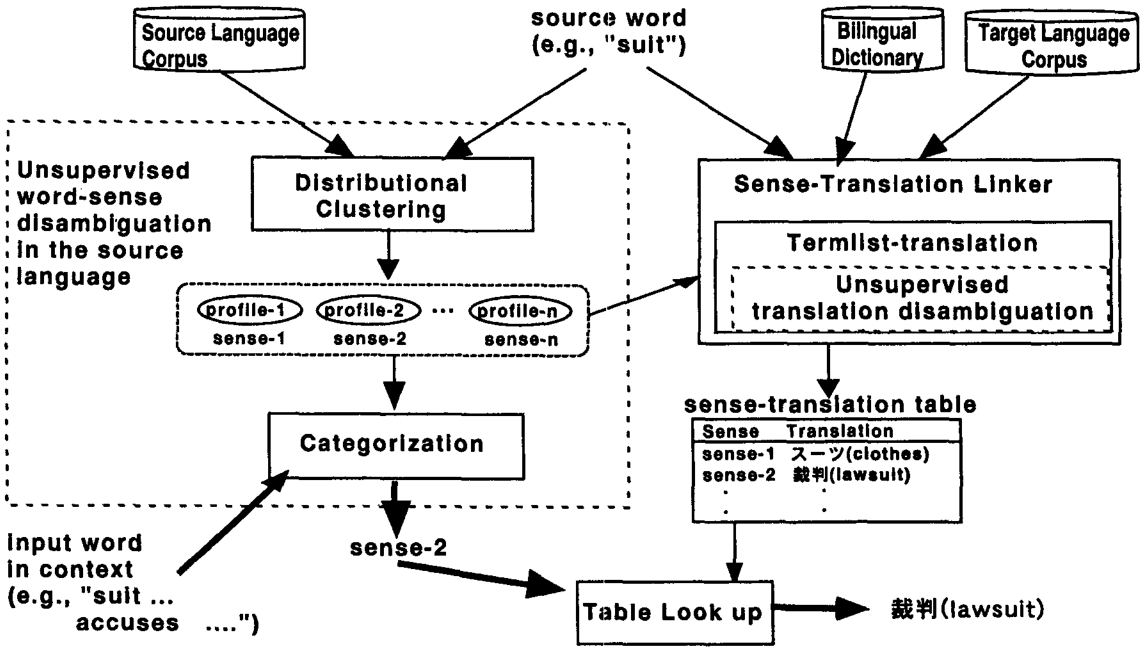

Figure 1: Overview of the disambiguation method

source language corpus and creates a profile, referred to as the "sense profile" for each class. The categorization process chooses the profile most relevant to the input word whose sense is implicitly given by its surrounding context.

Located to the right is what we call the *sense-translation linking* process. It is responsible for associating each semantic profile with the most likely translation of the source word (for which the semantic profile is derived). The result of this process is registered in the *sense-translation table*.

The table look-up process, bottom center, simply retrieves the translation associated with the sense identified by the categorization process.

3 Unsupervised Word Sense Disambiguation

We adopted the unsupervised word-sense disambiguation (WSD) algorithm based on distributional clustering (Schuetze, 1997). The underlying idea is that the sense of a word is determined by its co-occurring words. For example, the word "suit" co-occurring with "jacket" and "pants" tends to mean a set of clothes, whereas the same word co-occurring with "file" and "court" means "lawsuit".

As stated in Section 2, the WSD algorithm comprises two parts: distributional clustering and categorization.

The former learns the relation between sense and co-occurring words in the following steps:

1. Collecting contexts of the word in the corpus, then

2. Clustering them into small coherent groups (clusters).

Table 1 shows sample contexts surrounding "suit" as extracted by the first step from actual news articles. These contexts are expected to be clustered into two sets: (1,4) and (2,3) by the second step.

Since each cluster corresponds to a particular sense (or usage) of the word, it is called a "sense profile".

The latter part of the WSD algorithm is responsible for choosing the cluster "closest" or relevant to a new context of the same word (in this case, "suit"). The selected cluster is the "sense" in the new context.

3.1 Distributional Clustering

The above idea is implemented using the multi-dimensional vector space derived from word co-occurrence statistics in the source language corpus.

We first map each word, w, in the corpus onto a vector, $\vec{v}(w)$, referred to as the "word vector". A word vector is created from a large corpus(Schuetze, 1997)by the following procedure:

1. Choose a set of content bearing words (typically 1,000 most frequent content words).

Table 1: Contexts surrounding "suit"

No	Sample occurrences of "suit" in context		
1	..without fear of a libel	**suit**	A California court recently ..
2	..fitted jacket of the Chanel	**suit**	Referring to the push-up bras ..
3	..double-breasted dark blue	**suit**	Robinson was asked if he ...
4	.. the remaining	**suit**	accuses the hospital of turning...

2. Make a co-occurrence matrix whose (i, j) element corresponds to the occurrence count of the j-th content bearing word in the context of every occurrence of word-i[1] in the corpus. For simplicity we employ the *sliding window* approach where neighboring n words are judged to be the context.
3. Apply singular value decomposition (SVD) to the matrix to reduce its dimensionality.
4. The vector representation of word-i is the i-th row vector of the reduced matrix.

Second, the context of each occurrence of the word is also mapped to a multi-dimensional vector, called the "context vector". The context vector is the sum of every word vector within the context (again, neighboring n words) weighted by its *idf score*. Formally, the context vector $\vec{cxt}$ of word set W is defined as follows:

$$\vec{cxt}(W) = \sum_{w \in W} idf_w \vec{v}(w) \quad (1)$$

$$idf_w = \log(\frac{N}{N_w}) \quad (2)$$

$$N = \textit{the total number of documents in the collection} \quad (3)$$

$$N_w = \textit{the number of documents containing } w.^2 \quad (4)$$

Finally, derived context vectors are clustered by applying a clustering algorithm. We used the group-average agglomerative clustering algorithm called Buckshot(Cutting et al., 1992). In this algorithm, the proximity, *prox*, between two context vectors, $\vec{a}, \vec{b}$ is measured by cosine of the angle between these two vectors:

$$prox(\vec{a}, \vec{b}) = (\vec{a} \bullet \vec{b}) / (|\, \vec{a} \,||\, \vec{b} \,|) \quad (5)$$

Since we hypothesize that each translation alternative corresponds to at least one usage of the source word, the number of clusters is determined to be the number of the translation alternatives plus some fixed number (e.g., 3).

3.2 Categorization

The task of this step is to determine which semantic profile (i.e., context cluster) is "closest" to the word in a new context. In this step, the "closeness" between a semantic profile and the context is measured by the proximity, defined by (5), between the former vector and the representative vector of the latter called the *sense vector*. The sense vector of a sense profile is the centroid of all the context vectors in the cluster.

Unlike the original algorithm, we used only a portion (e.g., 70%) of the context vectors closest to the centroid for computing the sense vector since these central vectors contain less noise in terms of representing the cluster(Cutting et al., 1992).

4 Linking Sense to Translation

The WSD algorithm introduced in the previous section represents the sense of a given word, w, as a cluster of contexts (i.e., co-occurring words) in the source language. If each cluster is associated with one translation, then the result of the WSD can directly be maped to the translation.

Our method for associating each cluster with a translation consists of the following two major steps:

1. Extracting characteristic words from the cluster, then
2. Applying the termlist translation (disambiguation) algorithm(Kikui, 1998) to the list of words consisting of these characteristic words and the given word, w.

The termlist translation algorithm employed in the second step chooses the translation, from possible translation alternatives, that is most relevant to the context formed by the entire input (i.e., word

[1] Every word (type) is assigned a sequential id number.

[2] The document unit may be a paragraph, a text, an article etc.

Table 2: Characteristics words of sense profiles for "suit"

sense id	characteristic words
1	wearing(34.6), blue(28.5), designer(21.6), white(21.0), dark(20.7), shoes(20.6), hat(18.5), shirt(17.6), ...
2	fees(51.3), defendants(47.6), filed(45.6), court(43.8), District Court(35.8), fund(33.2), ...
3	

list). Thus, the second step is expected to translate the given source word w into the target word relevant to the sense represented by the cluster.

4.1 Extracting Characteristics Words

We applied IR[3] techniques to extract characteristic words as follows.

1. Let S be a sense profile of a source word w .
2. Extract central elements (i.e., contexts or context vectors) of S in the same way as described in Section 3.2.
3. Calculate the *tf-idf* score for each word in the extracted contexts, where *tf-idf* score of a word w is the frequency of w (term frequency) multiplied by the *idf* value defined by (3) in Section 3.1.
4. Choose the topmost m words (m is typically 10 to 20).

Table 2 shows the output of the above procedure applied to two sense profiles for "suit" using the training data shown in Section 5.1.

The extracted words for each cluster are combined with the source word to form a term-list. The term-list of length 8 for the first sense in Table 2 is :

> (suit, wearing, blue, designer, white, dark, shoes. hat).

Each term-list is then sent to the term-list translation module. The resulting translation of the source word is associated with the cluster and stored in the sense-translation table, shown in Figure 1.

4.2 Term-list Translation using Target Language Corpus

The termlist translation algorithm(Kikui, 1998) aims at translating a list of words that characterize a consistent text or a concept. It is an unsupervised algorithm in the sense that it relies only on a mono-lingual corpus free from manual tagging.

The algorithm first retrieves all translation alternatives of each word from a bilingual dictionary (*Dictionary Lookup*), then tries to find the most coherent (or semantically relevant) combination of the translation alternatives in the target language corpus (*Disambiguation*), detailed as follows:

1. Dictionary Lookup:

   For each word in the given term-list, all the alternative translations are retrieved from a bilingual dictionary. A combination of one translation for each input word is called ***a translation candidate***. For example, if the input is (book, library), then a translation candidate in French is (livre, bibliothèque).

2. Disambiguation:

   In this step, all possible translation candidates are ranked by the 'similarity' score of a candidate. The top ranked candidate is the output of the entire algorithm.

The *similarity score* of a translation candidate (i.e., a set of target words) is defined again by using the multi-dimensional vector space introduced in Section 3.1. Each target word in the translation candidates is, first, mapped to a word vector derived from the target language corpus. The similarity score *sim* of a set of (target) words W, is the average distance of word vectors from the centroid $\vec{c}$ of them as shown below.

$$\begin{aligned} sim(W) &= \frac{1}{\mid W \mid} \sum_{w \in W} prox(\vec{v}(w), \vec{c}(W)) && (6) \\ \vec{c}(W) &= \sum_{w \in W} \vec{v}(w) && (7) \\ \mid W \mid &= the\, number\, of\, words\, in\, W && (8) \end{aligned}$$

5 Evaluation and Discussion

We conducted English-to-Japanese translation experiments using newspaper articles. The results of the proposed algorithm were compared against those of the previous algorithm which relies solely on target language corpora(Kikui, 1998).

5.1 Experimental Data

The bilingual dictionary, from English to Japanese, was an inversion of the EDICT(Breen, 1995), a free Japanese-to-English dictionary.

The co-occurrence statistics were extracted from the 1994 New York Times (420MB) for English and 1994 Mainichi Shinbun (Japanese newspaper) (90MB) for Japanese. Note that 100 articles were

[3] IR = Information Retrieval

Table 3: Result of Translation for Test-NYT

Method	success/ambiguous (rate)
previous (Target Corpus Only)	91/120 (75.8%)
proposed	95/120 (79.1%)

randomly separated from the former corpus as the test set described below.

Although these two kinds of newspaper articles were both written in 1994, topics and contents greatly differ. This is because each newspaper publishing company edited its paper primarily for domestic readers. Note that the domains of these texts range from business to sports.

The initial size of each co-occurrence matrix was 50000-by-1000, where rows and columns correspond to the 50,000 and 1000 most frequent words in the corpus [4]. Each initial matrix was then reduced by using SVD into a matrix of 50000-by-100 using SVDPACKC(Berry et al., 1993).

Test data, a set of word-lists, were automatically generated from the 120 articles from the New York Times separated from the training set. A word-list was extracted from an article by choosing the topmost n words ranked by their *tf-idf* scores, given in Section 5. In the following experiments, we set n to 6 since it gave the "best" result for this corpus.

5.2 Results

In order to calculate success rates, translation 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, we then excluded them in calculating the success-rate.

The success rates of the proposed method and the previous algorithm are shown in Table 3.

5.3 Discussion

Although our method produced higher accuracy than the previous method, we cannot tell whether or not the difference is quantitatively significant. Further experiments with more data might be required.

From a qualitative view point, the proposed method successfully learned useful knowledge for choosing the correct target word. An example is shown in Table 4.

One advantage of the proposed method is that it is applicable to interactive disambiguation. The acquired disambiguation knowledge gives clues for choosing a target word in terms of the source language. For example, Table 4 enables English speakers to choose their preferred translation.

Table 4: An example of acquired disambiguation knowledge

Significant Word in Cluster	Translation
train, suburban, tracks, man, suicide, Glendale, brakes, ..	tsuitotsu (collision)
at-bats, game, homers, home, runs, ...	hitto (hit in baseball)
graceful, Lion King, comedy, Jackson, idea	hitto (becoming popular)

Another contribution of this research is that it gives one criteria for determining the number of clusters.

6 Related Work

Since we have referred to previous work in the area of statistical target word selection in Section 1 ((Brown et al., 1990), (Brown et al., 1991), (Brown, 1997), (Yamabana et al., 1998), (Dagan and Itai, 1994), (Tanaka and Iwasaki, 1996), (Kikui, 1998)), this section focuses on other related research.

Fung et al. ((Fung and K., 1997),(Fung and Yee, 1998)) presented interesting results on bilingual lexicon construction from "comparable corpora". which is non-parallel bilingual corpora in the same domain. Since their algorithm does not resolve word-sense ambiguity in the source language, it would be interesting to combine unsupervised disambiguation in the same way as we did.

Although we employed the distributional clustering algorithm for resolving word sense ambiguity, different algorithms are also applicable. Among them, the unsupervised algorithm using decision-trees (Yarowsky, 1995) has achieved promising performance. An interesting approach is to use the output of our sense-translation linking process as the "seeds" required by that algorithm.

7 Concluding Remarks

This paper presented an unsupervised method for choosing the correct translation of a source word in context. Preliminary experiments have shown that this method achieved 79% success rate. The method generates associations between word usages and their corresponding translations, which are useful for interactive machine translation.

Future directions include extending the proposed method with the decision-tree based word-sense dis-

[4] Stopwords are ignored.

ambiguation algorithm, and applying it to situations in which a reliable bilingual dictionary is not available.

8 Acknowledgment

A part of this work was done when the author was at Center for the Study of Language and Information(CSLI), Stanford University. The author acknowledges the fruitful discussions with the staff of the Computational Semantics Laboratory at CSLI. Comments from anonymous reviewers helped to improve the paper.

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):79–85.

P. Brown, V. Della Pietra, and R.L. Mercer. 1991. Word sense disambiguation using statisical methods. In *Proceedings of ACL-91*, pages 264–270.

R. D. Brown. 1997. Automated dictionary extraction for "knowledge-free" example-based translation. In *Proceedings of Theoretical and Methodological Issues in Machine Translation.*

D. R. Cutting, D. R. Karger, J. O. Pedersen, and J.W. Tukey. 1992. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of ACM SIGIR-92.*

I. Dagan and A. Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):564–596.

P. Fung and McKeown K. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora.*

P. Fung and L. Y. Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of COLING-ACL-98.*

G. Kikui. 1998. Term-list translation using monolingual co-occurence vectors. In *Proceedings of COLING-ACL-98.*

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.*

K. Yamabana, K. Muraki, S. Doi, and S. Kamei. 1998. A language conversion front-end for cross-language information retrieval. In G. Grefenstette, editor, *Cross-langauge Information Retrieval*, pages 93–104. Kluwer Academic Publishers.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL-95*, pages 189–195.