Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora

Takaaki TANAKA
NTT Communication Science Laboratories
2-4 Hikari-dai, Seika-cho, Soraku-gun,
Kyoto, 619-0237, JAPAN
takaaki@cslab.kecl.ntt.co.jp

Abstract

This paper presents a method that measures the similarity between compound nouns in different languages to locate translation equivalents from corpora. The method uses information from unrelated corpora in different languages that do not have to be parallel. This means that many corpora can be used. The method compares the contexts of target compound nouns and translation candidates in the word or semantic attribute level. In this paper, we show how this measuring method can be applied to select the best English translation candidate for Japanese compound nouns in more than 70% of the cases.

1 Introduction

Many electronic documents in various languages are distributed via the Internet, CD-ROM, and other computer media. Cross-lingual natural language processing such as machine translation (MT) and cross-lingual information retrieval (CLIR) is becoming more important.

When we read or write documents in a foreign language, we need more knowledge than what is provided in an ordinary dictionary, such as terminology, words relevant to current affairs, etc. Such expressions can be made up of multiple words, and there are almost infinite possible variations. Therefore, so it is quite difficult to add them and their translations to a dictionary.

Many approaches have tried to acquire translation equivalents automatically from parallel corpora (Dagan and Itai, 1994; Fung, 1995). In parallel corpora, effective features that have obvious correlations between these corpora can be used – e.g., similarity of position and frequency of words.

However, we cannot always get enough parallel corpora to extract the desired information. We propose a method of measuring the similarity to acquire compound noun translations by corpus information, which is not restricted to parallel corpora. Co-occurrence information is obtained as context where target nouns appear from the corpora. In a specific domain (e.g., financial news), a target word and its translation are often used in a similar context. For example, in a financial newspaper, *price competition* may appear with products (electric appliances, clothes, and foods), stores and companies more often than with nations and public facilities.

2 Extraction of Translations from Non-Parallel Corpora

In parallel corpora, positions and frequencies of translation equivalents are correlated; therefore, when we try to find translation equivalents from parallel corpora, this information provides valuable clues. On the other hand, in non-parallel corpora, positions and frequencies of words cannot be directly compared. Fung assumed that co-occurring words of translation equivalents are similar, and compared distributions of the co-occurring words to acquire Chinese-English translations from comparable corpora (Fung, 1997). This method generates co-occurring words vectors for target words, and judges the pair of words whose similarity is high to be translation equivalents. Rapp made German and English association word vectors and calculated the similarity of these vectors to find translations (Rapp, 1999). K.Tanaka and Iwasaki (1996) also assumed the resemblance between co-occurring words in a source language and those in a target language, and performed experiments to find irrelevant translations intentionally added to a dictionary.

In fact, finding translation equivalents from non-parallel corpora is a very difficult problem, so it is not practical to acquire all kinds of translations in the corpora. Most technical terms are composed of known words, and we must collect these words to translate them correctly because new terms can be infinitely created by combining several words. We focus on translations of compound nouns here. First, we collect the translation candidates of a target compound, and then measure the similarity between them to choose an appropriate candidate.

In many cases, translation pairs of compound nouns in different languages have corresponding component words, and these can be used as strong clues to finding the translations (Tanaka and Matsuo, 1999). However, these clues are sometimes insufficient for determining which is the best translation for a target compound noun when two or more candidates exist. For example, 営業利益 *eigyo rieki*, which means earnings before interest and taxes, can be paired with *operating profits* or *business interest*. Both pairs have common components, and we cannot judge which pair is better using only this information. A reasonable way to discriminate their meanings and usages is to see the context in which the compound words appear. In the following example, we can judge *operating profits* is a numerical value and *business interest* is a kind of group.

- ... its fourth-quarter *operating profit* will fall short of expectations ...
- ... the powerful coalition of *business interests* is pumping money into advertisements ...

Thus contextual information helps us discriminate words' categories. We use the distribution of co-occurring words to compare the context.

This paper describes a method of measuring semantic similarity between compound nouns in different languages to acquire compound noun translations from non-parallel corpora. We choose Japanese and English as the language pairs. The English translation candidates of a Japanese compound c_J that are tested for similarity can be collected by the method proposed by T.Tanaka and Matsuo (1999). The summary of the method is as follows except to measure the similarity in the third stage.

1. Collect English candidate translation equivalents C_E from corpus by part-of-speech (POS) patterns.
2. Make translation candidates set T_E by extracting the compounds whose component words are related to the components of c_J in Japanese from C_E.
3. Select a suitable translation c_E of c_J from T_E by measuring the similarity between c_J and each element of T_E.

In the first stage, this method collects target candidates C_E by extracting all units that are described by a set of POS templates. For example, candidate translations of Japanese compound nouns may be English noun–noun, adjective–noun, noun–of–noun, etc. T.Tanaka and Matsuo (2001) reported that 60% of Japanese compound nouns in a terminological dictionary are noun-noun or noun-suffix type and 55% of English are noun-noun or adjective-noun. Next, it selects the compound nouns whose component words correspond to those of nouns of the original language c_J, and makes a set of translation candidates T_E. The component words are connected by bilingual dictionaries and thesauri. For example, if c_J is 営業利益 *eigyo rieki*, the elements of T_E are $\{$*business interest*, *operating profits*, *business gain*$\}$.

The original method selects the most frequent candidate as the best one, however, this can be improved by using contextual information. Therefore we introduce the last stage; the proposed method calculates the similarity between the original compound noun c_J and its translation candidates by comparing the contexts, and chooses the most plausible translation c_E.

In this paper, we describe the method of selecting the best translation using contextual information.

3 Similarity between two compounds

Even in non-parallel corpora, translation equivalents are often used in similar contexts. Figure 1 shows parts of financial newspaper articles whose contents are unrelated to each other. In the article, 価格競争 *kakaku-kyousou* appears with 激化 *gekika* "intensify", 営業 *eigyo* "business", 利益 *rieki* "profit", 予想 *yosou* "prospect", etc. Its translation *price competition* is used with similar or relevant words – *brutal*, *business*, *profitless*, etc., although the article is not related to the Japanese one at all. We use the similarity of co-occurring words of target compounds in different languages to measure the similarity of the compounds. Since co-occurring words in different languages cannot be directly compared, a bilingual dictionary is used as a bridge across the corpora. Some other co-occurring words have similar meanings or are related to the same concept – 利益 "profit" and

海外旅行を中心に**価格競争**が激化，営業利益が従来の黒字予想から五億円程度の赤字になる見通し．(In particular, price competition of overseas travel has become intense. The operating profits are likely to show a five hundred million yen deficit, although they were expected to show a surplus at first.)

Price competition has become so brutal in a wide array of businesses – cellular phones, disk drives, personal computers – that some companies are stuck in a kind of profitless prosperity, selling loads of product at puny margins.

Figure 1: Examples of newspaper articles

価格競争 *kakaku kyoso*	1030	price competition	158	price control	100
激化する	223	- intensify	1		0
入札する	174	- bid	4		0
激しい	86	- severe	2		0
対抗する	21	- rival	5		1
安価だ	8	- cheap	1		0

Table 1: Common co-occurring words

margin, etc. These can be correlated by a thesaurus.

The words frequently co-occurring with 価格競争 are listed in Table 1. Its translation *price competition* has more co-occurring words related to these words than the irrelevant word *price control*. The more words can be related to each other in a pair of compounds, the more similar the meanings.

4 Context representation

In order to denote the feature of each compound noun, we use the context in the corpora. In this paper, context means information about words that co-occur with the compound noun in the same sentence.

4.1 Word co-occurrence

Basically, the co-occurring words of the target word are used to represent its features. Such co-occurring words are divided into two groups in terms of syntactic dependence, and are distinguished in comparing contexts.

1. Words that have syntactic dependence on the target word.
   (subject-predicate, predicate-object, modification, etc.)
   - ... fierce *price competition* by exporters ...
   - ... *price competition* was intensifying in this three months ...
2. Words that are syntactically independent of the target word.
   - ... intense *price competition* caused margins to shrink ...

Japanese			English		
CN		**N**	CN	(prep)	**N**
CN	の (*no*)	**N**	CN		**V**
CN	が (*ga*)	**V**	CN	(be)	**Adj**
CN	を (*o*)	**V**	**N**		CN
CN	に (*ni*)	**V**	**Adj**		CN
CN	が (*ga*)	**Adj**	**Ving**		CN

CN: a target compound, N: noun,
V: verb, Adj: adjective

Figure 2: Templates for syntactic dependence (part)

The words classified into the first class represent the direct features of the target word: attribute, function, action, etc. We cannot distinguish the role using only POS since it varies – attributes are not always represented by adjectives nor actions by verbs (compare intense *price competition* with *price competition is intensifying this month.*).

On the other hand, the words in the second class have indirect relations, e.g., association, with the target word. This type of word has more variation in the strength of the relation, and includes noise, therefore, they are distinguished from the words in the first class.

For simplicity of processing, words that have dependent relations are detected by word sequence templates, as shown in Figure 2. Kilgarriff and Tugwell collect pairs of words that have syntactic relations, e.g., *subject-of*, *modifier-modifiee*, etc., using finite-state techniques (Kilgarriff and Tugwell, 2001). The templates shown in Figure 2 are simplified versions for pattern matching. Therefore, the templates cannot detect all the dependent words; however, they can retrieve frequent and dependent words that are relevant to a target compound.

4.2 Semantic co-occurrence

Since finding the exact translations of co-occurring words from unrelated corpora is harder than from parallel corpora, we also compare the contexts at a more abstract level. In the example of “price competition”, a 電話 *denwa* “telephone” corresponds to a *fax* in term of communications equipment, as well as its exact translation, "telephone".

We employ semantic attributes from *Nihongo Goi-Taikei – A Japanese Lexicon* (Ikehara et al., 1997) to abstract words. *Goi-Taikei* originated from a Japanese analysis dictionary for the Japanese-English MT system ALT-J/E (Ikehara et al., 1991). This lexicon has about 2,700 semantic attributes in a hierarchical structure (maximum 12 level), and these attributes are attached to three hundred thousand Japanese words. In order to abstract English words, the bilingual dictionary for ALT-J/E was used. This dictionary has the same semantic attributes as *Goi-Taikei* for pairs of Japanese and English. We use 397 attributes in the upper 5 levels to ignore a slight difference between lower nodes. If a word has two or more semantic attributes, an attribute for a word is selected as follows.

1. For each set of co-occurring words, sum up the frequency for all attributes that are attached to the words.
2. For each word, the most frequent attribute is chosen. As a result each word has a unique attribute.
3. Sum up the frequency for an attribute of each word.

In the following example, each word has one or more semantic attributes at first. The number of words that have each attribute are counted: three for `[374]`, and one for `[494]` and `[437]`. As the attribute `[374]` appears more frequently than `[494]` among all words in the corpus, `[374]` is selected for "bank".

bank	:	`[374: enterprise/company]`, `[494: embankment]`
store	:	`[374: enterprise/company]`
hotel	:	`[437: lodging facilities]`, `[374: enterprise/company]`

4.3 Context vector

A simple representation of context is a set of co-occurring words for a target word. As the strength of relevance between a target compound noun t and its co-occurring word r, the feature value of r, $\mu_w(t,r)$ is defined by the log likelihood ratio (Dunning, 1993) [1] as follows.

$$\mu_w(t,r) = \begin{cases} L(t,r) & : f(t,r) \neq 0 \\ 0 & : f(t,r) = 0 \end{cases} \tag{1}$$

$$\begin{aligned} L(t,r) &= \sum_{i,j \in 1,2} k_{ij} \log \frac{k_{ij}N}{C_i R_j} \\ &= k_{11} \log \frac{k_{11}N}{C_1 R_1} + k_{12} \log \frac{k_{12}N}{C_1 R_2} \\ &+ k_{21} \log \frac{k_{21}N}{C_2 R_1} + k_{22} \log \frac{k_{22}N}{C_2 R_2} \end{aligned} \tag{2}$$

$$\begin{aligned} k_{11} &= f(t,r) \\ k_{12} &= f(t) - k_{11} \\ k_{21} &= f(r) - k_{11} \\ k_{22} &= N - k_{11} - k_{12} - k_{21} \\ C_1 &= k_{11} + k_{12} \\ C_2 &= k_{21} + k_{22} \\ R_1 &= k_{11} + k_{21} \\ R_2 &= k_{12} + k_{22} \end{aligned} \tag{3}$$

where $f(t)$ and $f(r)$ are frequencies of compound noun t and co-occurring word r, respectively. $f(t,r)$ is the co-occurring frequency between t and r, and N is the total frequencies of all words in a corpus.

The context of a target compound t can be represented by the following vector (context word vector 1, $\mathbf{c}_{w1}$), whose elements are the feature values of t and its co-occurring words r_i.

$$\boldsymbol{c}_{w1}(t) = (\mu_w(t,r_1), ..., \mu_w(t,r_n)) \tag{4}$$

Note that the order of the elements are common to all vectors of the same language. Moreover, translation matrix T, described in K.Tanaka and Iwasaki (1996), can convert a vector to another vector whose elements are aligned in the same order as that of the other language ($T\boldsymbol{c}_w$). The element t_{ij} of T denotes the conditional probability that a word r_i in a source language is translated into another word r_j in a target language.

We discriminate between words that have syntactic dependence and those that do not because the strengths of relations are different as mentioned in Section 4.1. In order to intensify the value of dependent words, $f(t,r)$ in equation(3) is replaced with the following $f'(t,r)$ using the weight w determined by the frequency of dependence.

$$f'(t,r) = wf(t,r) \tag{5}$$

[1] This formula is the faster version proposed by Ted Dunning in 1997.

$$w = 1 + \frac{f_d(t,r)}{f(t,r)} * const \tag{6}$$

Here, $f_d(t,r)$ is the frequency of word r that has dependency on t. The constant is determined experimentally, and later evaluation is done with $const = 2$. We define a modified vector (context word vector 2, $\mathbf{c}_{w2}$), which is a version of $\mathbf{c}_{w1}$.

Similarly, another context vector is also defined for semantic attributes to which co-occurring belong by using the following feature value μ_a instead of μ_w (context attribute vector, $\mathbf{c}_a$). L_a in equation (8) is the semantic attribute version of L in equation (2). $f(t,r)$ and $f(t)$ are replaced with $f(a,r)$ and $f(a)$, respectively, where a indicates an attribute of a word.

$$\boldsymbol{c}_a(t) = (\mu_a(t,a_1), ..., \mu_a(t,a_m)) \tag{7}$$

$$\mu_a(t,a) = \begin{cases} L_a(t,a) & : f(t,a) \neq 0 \\ 0 & : f(t,a) = 0 \end{cases} \tag{8}$$

5 Comparison of context

As described in Section 3, the contexts of a compound noun and its translation are often alike in the corpora of a similar domain. Figure 2 shows a comparison of co-occurrence words and semantic attributes of three compound nouns – 営業利益 and its translation *operating profit*, and an irrelevant word, *business interest*. Each item corresponds to an element of context vector c_w or c_a, and the words in the same row are connected by a dictionary. The high μ_w words in the class of "independent words" include words associated indirectly to a target word, e.g., 見込み *mikomi –expectation*, 配当 *haito –share*. Some of these words are valid clues for connecting the two contexts; however, others are not very important. On the other hand, words in the class of "dependent words" are directly related to a target word, e.g., *increase of operating profits*, *estimate operating profits*. The variety of these words are limited in comparison to the "independent words" class, whereas they often can be effective clues.

More co-occurring words of *operating profits* that mark high μ_w are linked to those of 営業利益 rather than *business interest*. As for semantic attributes, *operating profit* shares more upper attributes with 営業利益 than *business interest*.

営業利益	μ_w/μ_a	operating profit	μ_w/μ_a	business interest	μ_w/μ_a
[independent words]					
利益	3478	profit	117		
削減する	654	slash	16.3	reduce	7.8
見込み	508	expectation	137		
合理化	455			rationalize	46.8
模様	363	design	11.2		
配当	353	share	130		
[dependent words]					
増える	1866	increase	49.7	increase	6.3
減少する	727	decline	5.9	diminish	14.1
予想する	709	estimate	51.2	estimate	3.6
部門	422	division	49.2	division	7.1
比べる	321	contrast	3.2	compete	9.4
連結	266	connect	4.9	link	5.4
[semantic attributes](*)					
[2262]	8531		131		11
[1867]	7290		321		93
[2694]	4936		13		19
[1168]	3855				83
[1395]	3229		695		110
[1920]	1730		810		428

(*) [2262:increase/decrease],[1867:transaction], [2694:period/term],[1168:economic system], [1395:consideration],[1920:labor]

Table 2: Comparison of co-occurrence word and semantic attributes

The similarity $S_w(t_s, t_t)$ between compound nouns t_s in the source language and t_t in the target language is defined by context word vectors and translation matrix T as follows.

$$S_w(t_s, t_t) = T\mathbf{c}_w(t_s)\mathbf{c}_w(t_t) \tag{9}$$

Similarly, the semantic attribute based similarity $S_a(t_s, t_t)$ is defined as follows.

$$S_a(t_s, t_t) = \mathbf{c}_a(t_s)\mathbf{c}_a(t_t) \tag{10}$$

6 Evaluation

In order to evaluate this method, an experiment on the selection of English translations for a Japanese compound noun is conducted. We use two Japanese corpora, Nihon Keizai Shimbun CD-ROM 1994 (NIK, 1.7 million sentences) and Mainichi Shimbun CD-ROM 1995 (MAI, 2.4 million sentences), and two English corpora, The Wall Street Journal 1996 (WSJ, 1.2 million sentences) and Reuters Corpus 1996 (REU, 1.9 million sentences [2] Reuters (2000)) as contextual information. Two of them, NIK and WSJ, are financial newspapers, and the rest are general newspapers and news archives. All combinations of Japanese and English corpora are examined to reduce the bias of the combinations.

[2] Only part of the corpus is used because of file size limitation in the data base management system in which the corpora are stored.

First, 400 Japanese noun-noun type compounds c_J that appear frequently in NIK (more than 15 times) are randomly chosen. Next, the translation candidates T_E for each c_J are collected from the English corpus WSJ as described in Section 2. The bilingual dictionary for MT system ALT-J/E, *Goi-Taikei* and a terminological dictionary (containing about 105,000 economic and other terms) are used to connect component words. As a result, 393 Japanese compound nouns and their translation candidates are collected and the candidates for 7 Japanese are not extracted. Note that we link component words widely in collecting the translation candidates because components in different languages do not always have direct translations, but do have similar meanings. For instance, for the economic term 設備投資 *setsubi toushi* and its translation *capital investment*, while 投資 *toushi* means *investment*, 設備 *setsubi*, which means *equipment* or *facility*, is not a direct translation of *capital.* The *Goi-Taikei* and the terminological dictionary are employed to link such similar component words. Each Japanese word has a maximum of 5 candidates (average 3 candidates). We judge adequacy of chosen candidates by referring to articles and terminological dictionaries. More than 70% of Japanese have only one clearly correct candidate and many incorrect ones (e.g. *securities company* and *paper company* for 証券会社 *shouken gaisha*). The others have two or more acceptable translations.

Moreover, if all of the translation candidates of compound c_J are correct (45 Japanese), or all are incorrect (86 Japanese), c_J and its translation candidates are removed from the test set. For each c_J in the remainder of the test set (262 Japanese compound nouns, set 1), a translation c_E that is judged the most similar to c_J is chosen by measuring the similarity between the compounds. Set 1 is divided into two groups by the frequency of the Japanese word: set 1H (more than 100 times) and set 1L (less than 100 times) to examine the effect of frequency. In addition, the subset of set 1 (135 Japanese compound nouns, set 2), whose members also appear more than 15 times in MAI, is extracted, since set 1 includes compounds that do not appear frequently in MAI. On the other hand, the candidate that appears the most frequently in the English corpus can be selected as the best translation of c_J. This simple procedure is the baseline that is compared to the proposed method.

sets	corpora	word1	word2	attr	freq
		(c_{w1})	(c_{w2})	(c_a)	(WSJ)
[1H]	NIK-WSJ	73.4	74.2	66.4	65.6
[1L]	NIK-WSJ	53.3	53.3	43.0	46.7
[1]	NIK-WSJ	63.0	63.4	54.2	55.7
[2]	NIK-WSJ	71.1	72.6	65.9	64.4
[2]	NIK-REU	71.9	71.9	66.7	
[2]	MAI-WSJ	58.5	58.5	63.7	
[2]	MAI-REU	57.0	56.3	65.2	

Table 3: Precision of selecting translation candidates

Table 3 shows the result of selecting the appropriate English translations for the Japanese compounds when each pair of corpora is used. The column of "freq(WSJ)" is the result of choosing the most frequent candidates in WSJ. Since the methods based on word context reach higher precision in set 1, this suggests that word context vectors (c_w) can efficiently describe the context of the target compounds. For almost all sets, context word vector 2 provides higher precision than context word vector 1. However, the effect of consideration of syntactic dependency is minimal in this experiment.

The precisions of word context vector in both MAI-WSJ and MAI-REU are low. This main reason is that many Japanese compounds in the test set appear less frequently in MAI than in NIK, since the frequent compounds in NIK are chosen for the set (the average frequency in NIK is 417, but that in MAI is 75). Therefore, less common co-occurrence words are found in MAI and the English corpora than in NIK and them. For instance, 25 Japanese compounds share no co-occurrence words with their translation candidates in MAI-WSJ while only one Japanese shares none in NIK-WSJ. In spite of this handicap, the method based on semantic context (c_a) of MAI-WSJ/REU has the high precision. This result suggests that an abstraction of words can compensate for lack of word information to a certain extent.

The proposed method based on word context (c_w) surpasses the baseline method in precision in measuring the similarity between relatively frequent words. Our method can be used for compiling dictionary or machine translations.

Table 4 shows examples of translation candi-

Japanese	English candidates	$S_{w1} \times 10^{-2}$
技術移転	+ technology transfer	24433
	technology share	20849
	policy move	9173
為替レート	+ exchange rate	530509
	market rate	111712
	bill rate	46417
電力会社	energy company	730323
	+ power company	441790

Table 4: Examples of translation candidates

電力会社	power company	energy company
電力	power	power
エレクトリック	electric	electric
停電	blackout	
余る	remain	remain
料金	charge	charge
貯蔵		storage

Table 5: Similar co-occurring words of hypernyms and hyponyms

dates and their similarity scores. The mark "+" indicates correct translations. Some hyponyms and hypernyms or antonyms cannot be distinguished by this method, for these words often have similar co-occurring words. As shown in Table 5, using the example of *power company* and *energy company*, co-occurring words are very similar, therefore, their context vectors cannot assist in discriminating these words. This problem cannot be resolved by this method alone. However, there is still room for improvement by combining other information, e.g., the similarity between components.

7 Conclusion

We proposed a method that measures the similarity between compound nouns in different languages by contextual information from non-parallel corpora. The method effectively selects translation candidates although it uses unrelated corpora in different languages. It measures the similarity between relatively frequent words using context word vector. As for less common co-occurrence words, context attribute vectors compensate for the lack of information. For future work, we will investigate ways of integrating the method and other information, e.g., similarity of components, to improve precision.

Acknowledgements

This research was supported in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. The author would like to thank Timothy Baldwin of CSLI and Francis Bond of NTT for their valuable comments.

References

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Pascale Fung. 1995. A pattern method for finding noun and proper noun translation from noisy parallel corpora. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics.*

Pascale Fung. 1997. Finding terminology translations from non-parallel corpora. In *Proc. of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing – effects of new methods in ALT-J/E –. In *Proc. of the 3rd Machine Translation Summit*, pages 101–106.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Nihongo Goi-Taikei – A Japanese Lexicon.* Iwanami Shoten.

Adam Kilgarriff and David Tugwell. 2001. WASP-bench: an MT lexicographers' workstation supporting state-of-art lexical disambiguation. In *Proc. of the 8th Machine Translation Summit.*

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proc. of the 37th Annual Meeting of the Association of Computational Linguistics*, pages 1–17.

Reuters. 2000. Reuters corpus (1996–1997).

Kumiko Tanaka and Hideya Iwasaki. 1996. Extraction of lexical translations from non-aligned corpora. In *Proc. of the 16th International Conference on Computational Linguistics*, pages 580–585.

Takaaki Tanaka and Yoshihiro Matsuo. 1999. Extraction of translation equivalents from non-parallel corpora. In *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 109–119.

Takaaki Tanaka and Yoshihiro Matsuo. 2001. Extraction of compound noun translations from non-parallel corpora (in Japanese). *In Trans. of the Institute of Electronics, Information and Communication Engineers*, 84-D-II(12):2605–2614.