Integrated Use of Internal and External Evidence in the Alignment of Multi-Word Named Entities

Takeshi KUTSUMI
SHARP Corporation
492 Minosho-cho, Yamatokoriyama-shi,
Nara, 639-1186, JAPAN
kutsumi.takeshi@sharp.co.jp

Takehiko YOSHIMI
Ryukoku University
1-5 Yokotani, Setaoe-cho, Otsu-shi, Shiga,
520-2925, JAPAN
yoshimi@rins.ryukoku.ac.jp

Katsunori KOTANI
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto,
619-0289, JAPAN
kat@khn.nict.go.jp

Ichiko SATA
SHARP Corporation
492 Minosho-cho, Yamatokoriyama-shi,
Nara, 639-1186, JAPAN
sata@isl.nara.sharp.co.jp

Hitoshi ISAHARA
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, JAPAN
isahara@nict.go.jp

Abstract

This paper proposes a method of extracting English multi-word named entities and their Japanese equivalents from a parallel corpus. The aim of our research is to extract multi-word named entities which are not listed in a dictionary of an English-to-Japanese MT system and appear infrequently in a parallel corpus. Our method makes its alignment on the basis of two kinds of external evidence provided by the context in which a bilingual pair appears, as well as two kinds of internal evidence within the pair. Each evidence is accompanied by a score, and the aggregate score is computed as a weighted sum of the scores. The appropriate weights are estimated with the logistic regression analysis. An experiment using a parallel corpus of Yomiuri Shimbun and The Daily Yomiuri satisfactorily found that 86.36% of the extracted bilingual pairs with the highest scores were judged to be correct.

1 Introduction

Over the past few decades, English-to-Japanese bilingual dictionaries have been expanded both in quality and in quantity. The development has led to the practical use of MT systems which have the dictionaries of more than two millions of entries. But words which are not found even in such large-scale dictionaries can appear in real life texts. Since they are not found in the large dictionaries, they would appear infrequently in corpora.

Many researches have pursued the extraction of bilingual pairs from a parallel corpus in which sentences are aligned (Eijk 1993; Kupiec 1993; Dekai & Xia 1994; Smadja & McKeown 1996; Ker & Chang 1997; Le et al. 1999; Huang et al. 2003), as well as from a comparable corpus in which sentences are not aligned (e.g. Tanaka 1999; Nakagawa 2001; Chiao 2002). Furthermore, some researchers have focused on the extraction of multi-word named entity translations (Al-Onaizan 2002; Moore 2003). The methods of extracting the bilingual pairs from a parallel corpus seem to have been established. However, in the case of aiming at extraction of the bilingual pair of a word with low frequency of occurrence and its counterpart, as is well known, statistical methods often encounter serious problems.

Although literature on this task is available (Dagan & Church 1994; Tsuji 2001), it calls for further investigation to determine what kinds of linguistic information we should use and how we should combine them.

We propose a method of extracting the bilingual pair of an English multi-word named entity (hereafter MWNE) and its Japanese counterpart. Here we focus our attention on the MWNEs which appear just once in a parallel corpus and are not listed in the dictionary of a practical English-to-Japanese machine translation (MT) system.

Our method keeps four kinds of scores for a bilingual pair based on the internal and external evidence of the pair. Internal evidence means what a word itself contains, while external evidence means what is provided by the context in which the word appears.[1] It calculates the overall score as the weighted sum of the four kinds of scores. The appropriate weights are obtained with the logistic regression analysis.

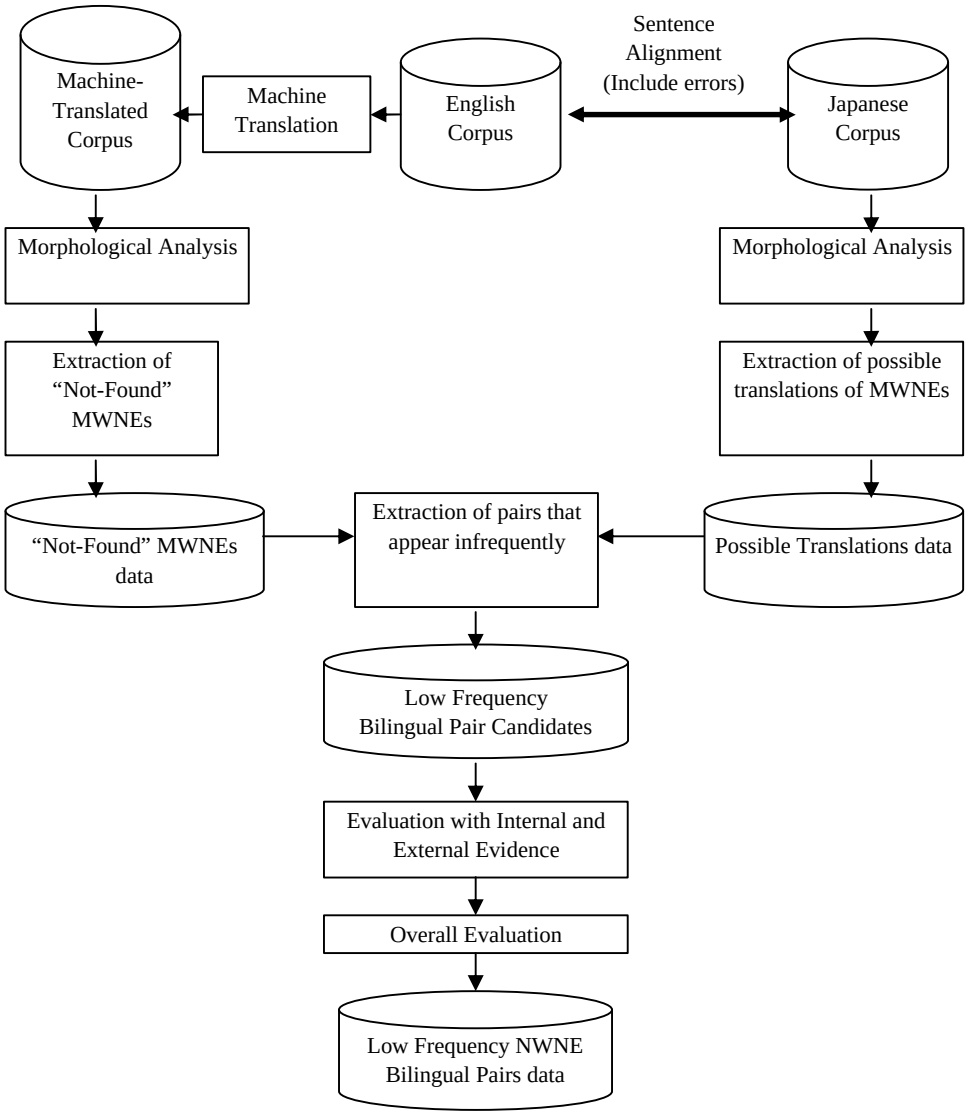

Figure 1: Outline of our MWNE pairs acquisition system

[1] The idea of consulting the internal and external evidences is found in the literature (Kaji01), but it only gives a mention, and does not present any results of experiments based on the idea.

2 Outline of Aligning Multi-Word Named Entities

Figure 1 shows the outline of our method for extracting MWNE pairs.

The outline of our method is as follows.

1. We translated the English sentences in the parallel corpus with our MT system. If an English word is not found in the dictionary of our MT system, it is left untranslated in a machine-translated (MT) sentence.
2. We extracted from the MT sentences a sequence of words beginning with capital letters, which we assume to be an MWNE. Hereafter, we call a "not-found" MWNE simply as an MWNE.
3. We performed morphological analysis of the MT sentences and of the Japanese sentences in the parallel corpus.[2] Moreover, we extracted a sequence of nouns as a possible translation of an MWNE from the Japanese sentence that corresponds to an MT sentence.
4. Out of the bilingual pairs obtained by the step 2 and 3, we focused on the pairs of an MWNE and its possible translation where 1) the MWNE appears just once in the MT sentences, 2) the possible translation appears just once in the Japanese sentences, and 3) the MWNE and its possible translation cooccur just once in the aligned sentences[3].
5. We gave a score to each bilingual pair based on external evidence as well as internal evidence. We discuss details in Section 3.1 through Section 3.4.
6. We computed an overall score for each bilingual pair by integrating individual scores given in the step 5, and sorted the pairs in descending order of the overall scores. We show the way of obtaining the overall score in Section 3.5.

3 Internal and External Evidences for Extracting Bilingual Pairs

In this section, we will explain the four kinds of evidence we exploit to extract the bilingual pairs.

3.1 Similarity of Literal Translations

Similarity could be seen between the translation of a whole MWNE and the expression obtained by translating the components of the MWNE literally (Kumano & Hirakawa 94). Based on this knowledge, we give a score to each bilingual pair of an MWNE and its possible translation.

We use the Jaccard coefficient to estimate the similarity. Let E be an MWNE, and J be a possible translation of E. And let X be a set consisting of the components of J, and Y be a set consisting of the components of the literal translation of E. Then the score of the similarity of literal translation for the bilingual pair, S_1, is calculated according to the following formula (1):

$$S_1(E,J)=\frac{|X\cap Y|}{|X\cup Y|} \quad (1)$$

Note that we do not take into account the order of the components.

Consider, for example, the bilingual pair of the MWNE, "Disaster Prevention Law", which is not an entry of our MT dictionary, and its possible translation **災害対策基本法** *saigai taisaku kihon ho* "disaster measure base law". The set X consists of the four words **災害** *saigai* "disaster", **対策** *taisaku* "measure", **基本** *kihon* "base" and **法** *ho* "law". Our MT system translates each component of the MWNE, "Disaster Prevention Law" as **災害** *saigai* "disaster", **防止** *boshi* "prevention" and **法** *ho* "law" respectively. These three words are the members of the set Y. Therefore the members of the intersection of the two sets X and Y are **災害** *saigai* "disaster" and **法** *ho* "law", and the members of the union of the

[2] We used the morphological analyzer *Chasen* (http://chasen.aist-nara.ac.jp/index.html.en).

[3] In the case of aiming at extraction of the bilingual pair of a word and its counterpart which do not appear frequently in a corpus, statistical methods often encounter serious problems. Therefore, it is an important issue to develop a method of extracting entities which appear infrequently. Our method can also be applied to the MWNEs whose frequency is more than one.

two sets are **災害** *saigai* "disaster", **対策** *taisaku* "measure", **基本** *kihon* "base", **法** *ho* "law" and **防止** *boshi* "prevention". We obtain the score of S_1(Disaster Prevention Law, **災害対策基本法** *saigai taisaku kihon ho* "disaster measure base law") equal to 2/5.

3.2 Phonological Correspondence

The evidence discussed in Section 3.1 would be reliable in finding a translation of an MWNE, whose components are listed separately in the MT dictionary although its entry is not found. However, there is a case where even the separate components of an MWNE are not listed, being typically true in finding the translation of proper names.

The English part of the parallel corpus we used, The Daily Yomiuri, consists of many articles about Japanese domestic events and involves many Japanese proper names. Some of them can be easily transliterated by preparing the Japanese syllabary. For example, the English word "Nara", the name of a city in Japan, is transliterated as **ナラ** *nara* "nara" in Japanese. Note that transliterating such a word as "festival" fails because our current method just scans the Japanese syllabary. In case of failure, we regard the English word itself as a transliteration. Thus, ``Nara festival'' comes out as ``**ナラ** festival''.

The morphological analyzer we used is able to yield the pronunciation of a Japanese word. For example, it yields the pronunciation **ナラ マツリ** *nara matsuri* "nara festival" for the input word **奈良祭り** *nara matsuri* "nara festival".

We again use the Jaccard coefficient to evaluate phonological correspondences of a bilingual pair. Let X be a set consisting of the transliteration of an MWNE E, and Y be a set consisting of the pronunciation of its possible translation J. Then we can obtain the score of phonological correspondences for the bilingual pair, S_2, according to such a formula as (1). In the case of the above example, the member of the intersection of the two sets X and Y is **ナラ** *nara* "nara", and the members of the union of the two sets are **ナラ** *nara* "nara", "festival" and **マツリ** *matsuri* "festival". We obtain the score of S_2(Nara festival, **奈良祭り** *nara matsuri* "nara festival") equal to 1/3.

3.3 Similarity of Neighboring Nouns

In a parallel corpus, bilingual pairs of words commonly appear in similar contexts (Kaji & Aizono 1996; Fung 1998). This means that a similar context would encourage the likelihood of a bilingual pair.

We represent the context of an MWNE by a set of nouns surrounding it, which we refer to as neighboring nouns. We suppose that the number of the neighboring nouns of an MWNE is proportional to the total number of nouns in the sentence in which the MWNE appears: the number of the neighboring nouns is a quarter of the total number of nouns which locate on the right and left of the MWNE respectively. We do the same for the possible translations of an MWNE.

(2) a. Germany provided a new impetus to European integration, terminating in the Maastrich Treaty.

b. **ドイツは** Maastrich Treaty **で** **終了して**
doitsu-ha *maasutorihito toriitii-de* *shuryo-shite*
Germany-SBJ Maastrich Treaty in terminate
欧州統合に **新しい** **刺激を** **した**
oshutogo-ni *atarashii* *shigeki-wo* *shita*
European integration to new impetus do

c. **ドイツは** **マーストリヒト条約として** **結実する**
doitsu-ha *maasutorihito joyaku-toshite* *ketsujitsu-suru*
Germany-SBJ Maastrich Treaty in terminate
欧州統合に **新たな** **弾み を** **与えた**
oshutogo-ni *aratana* *hazumi-wo* ataeta
European integration to new impetus provide

Consider the following example. The English sentence (2a) is translated into the MT sentence (2b), and is aligned with the Japanese sentence (2c) in the parallel corpus. The MWNE "Maastrich Treaty", being not listed in our MT dictionary, is left untranslated in (2b). Its possible translations in (2c) are **マーストリヒト条約** *maasutorihito joyaku* "Maastrich Treaty" and **弾み** *hazumi* "impetus".

Since the total number of nouns in the sentence (2b) is four[4], the neighboring nouns of the NWNE "Maastrich Treaty" are one (=4/4) noun on its left and one (=4/4) noun on its right, namely **ドイツ** *doitsu* "Germany" and **欧州統合** *oshutogo* "European integration". The sentence (2c) has four nouns, too[5]. Therefore the neighboring nouns of the possible translation **マーストリヒト条約** *maasutorihito joyaku* "Maastrich Treaty" are **ドイツ** *doitsu* "Germany" and **欧州統合** *oshutogo* "European integration", and the neighboring noun of the possible translation **弾み** *hazumi* "impetus" is **欧州統合** *oshutogo* "European integration".

We adopt the Jaccard coefficient to estimate the similarity of the set of the neighboring nouns. Let X be a set of the neighboring nouns of an NWNE E and let Y be a set of the neighboring nouns of its possible translation J, then the score of the similarity of neighboring nouns, S_3, is calculated according to the formula (1).

In the above example, let X be a set of the neighboring nouns of the MWNE "Maastrich Treaty" and let Y be a set of the neighboring nouns of **マーストリヒト条約** *maasutorihito joyaku* "Maastrich Treaty'". then the size of the intersection of the two sets X and Y is 2 and that of the union of them is 2. This leads the score of S_3 (Maastrich Treaty, **マーストリヒト条約** *maasutorihito joyaku* "Maastrich Treaty'" to 2/2. Let Y be a set of the neighboring nouns of **弾み** *hazumi* "impetus", then the size of the intersection of the two sets X and Y is 1 and that of the union of them is 2, leading the score of S_3(Maastrich Treaty, **弾み** *hazumi* "impetus") to 1/2. This results in preference **マーストリヒト条約** *maasutorihito joyaku* "Maastrich Treaty" over **弾み** *hazumi* "impetus" as the translation of "Maastrich Treaty".

3.4 Nonexistence of the Equivalent Nouns

Suppose there exists a noun in an MT sentence that is equivalent to the possible translation of an MWNE. In this situation, the possible translation is likely to pair with the equivalent noun rather than with the MWNE.

Accordingly, we give a lower score to the bilingual pair of an MWNE and its possible translation, which has the equivalent noun in the MT sentence. In the experiments, we define the score on the equivalent noun, S_4, as 0 in the case where a possible translation has the equivalent noun. In the case where such a noun does not exist, we define the score as 0.5.

$$S_4 = \begin{cases} 0 & \textit{if the equivalent noun exists} \\ 0.5 & \textit{otherwise} \end{cases}$$

3.5 Overall Evaluation

Each bilingual pair receives the separate scores from the external and internal evidences. The different types of evidence may be at conflict. That is to say, there is no guarantee that the best score for one evidence will be the best score for another evidence.

It is necessary to arrive at a relative weighting between the separate evidences. The overall evaluation is based on the four kinds of scores according to the formula (3).

$$S = C + \sum_{i=1}^{4} (W_i \times S_i) \qquad (3)$$

[4] The nouns are **ドイツ** *doitsu* "Germany", Maastrich Treaty, **欧州統合** *oshutogo* "European integration", and **刺激** *shigeki* "impetus".

[5] The nouns are **ドイツ** *doitsu* "Germany", **マーストリヒト条約** *maasutorihito joyaku* "Maastrich Treaty", **欧州統合** *oshutogo* "European integration", and **弾み** *hazumi* "impetus".

Here C stands for a constant value, and W_i for relative weighting of the score S_i.

There is a way of obtaining and maintaining the values of C and W_i empirically. Unlike this conventional way, we propose a way of achieving the values by performing the logistic regression analysis.[6]

4 Experiment and Discussion

In this section we present the results of the experiments we carried out and clarify the factors leading to incorrect alignments. We also discuss the effectiveness of each evidence. Furthermore, we examine how much the degree of independence among the evidences is.

4.1 Method

Our experiments use the parallel corpus of Yomiuri Shimbun and The Daily Yomiuri, sentences of which are aligned automatically (Utiyama & Isahara 2003). We limit the scope of the articles which cover the period from September 1989 to the mid July 1996. Utiyama and Isahara have given each sentence pair a alignment score. While the corpus involves incorrect alignments, an alignment with a high score is likely to be correct. We use only the sentence pairs with the top 10\% of the entire corpus.

Based on preliminary observation, we empirically assumed that the reliability of the four kinds of evidence increases in the following order.

1. Similarity of neighboring nouns (S_3)
2. Similarity of literal translation (S_1), and nonexistence of the equivalent nouns (S_4)
3. Phonological correspondences (S_2)

Therefore, we first decided C and W_i in the formula (3) as the values shown in the row “ED” (empirical decision) in Table 1.

On the other hand, we performed the logistic regression analysis to obtain those values as shown in the row “LRA” (logistic regression analysis) in Table 1. We made a training data set by let an evaluator give the judgment of “correct” or “incorrect” to the 1734 bilingual pairs. The four kinds of evidence were treated as the explanatory variables, while the judgment was considered as the dependent variable (0 or 1).

Table 1: Values of Weights

	C	W_1	W_2	W_3	W_4
ED	0	2	3	1	2
LRA	-4.58	20.75	15.04	3.58	2.81

The comparison of the weight values decided empirically with the ones obtained by the logistic regression analysis shows the following differences.

- In the case of the empirical decision, the gaps between the weights for the internal evidences (W_1 and W_2) and the ones for the external evidences (W_3 and W_4) are small. Unlike the case of the empirical decision, the gaps between the weights obtained by the logistic regression analysis are large.
- While we assumed empirically that the reliability of each evidence becomes higher in the order of $S_3 < S_1 = S_4 < S_2$, the result of the logistic regression analysis shows that the appropriate order is $S_4 < S_3 < S_2 < S_1$.

Each bilingual pair is given one of the two evaluations: ``correct'' or ``incorrect''. A bilingual pair would be judged to be correct when it can be registered in the MT dictionary without any modification of Japanese words: for example, the bilingual pair of “Comprehensive Security Board” and **総合安全保障審議会** *sogo anzen hosho shingikai* “comprehensive safety security council” is judged to

[6] Regression analysis has been used for NLP research such as text summarization (Watanabe 1996).

be correct. The judgment "incorrect" would be given to a pair when it needs adding, deleting or replacing of Japanese words.

4.2 Evaluation

For evaluation, we randomly sampled 264 MWNEs from the full results. The number of possible translations for these 264 MWNEs were 1086. Each MWNE has 4.11 (1086/264) possible translations on average.

Table 2 shows the results of the two experiments: one is carried out with the use of the weights decided empirically (ED), and the other with the use of the weights obtained by the logistic regression analysis (LRA). The numbers in the parentheses are the numbers of the pairs being judged as correct. Table 2 indicates that the rate of correct alignments achieved by the logistic regression analysis surpasses the one by the empirical decision, that is to say, our proposed method is more effective than the conventional way of weighting.

Table 2: Rate of Correct Alignments

	Top (solo)	Top (inc. tie)	Top Two
ED	74.24% (196)	83.71% (221)	92.80% (245)
LRA	77.65% (205)	86.36% (228)	95.08% (251)

Table 3 shows how the order of output of the correct alignments changes between the empirical decision and the logistic regression analysis. The distribution indicates that the logistic regression analysis is superior to the empirical decision in 15 bilingual pairs, while the former is inferior to the latter in 5 bilingual pairs. Of the 15 pairs, 2 pairs went up from tie for first place to the solo possession of first place, 7 pairs from second place to the solo possession of first place, 5 pairs from third place and under to the solo possession of first place, and 1 pair from third place and under to second place.

Table 3: Distribution of Ranking

ED \ LRA	Top (solo)	Top(tie)	Second	Third and under	Total
Top (solo)	191	0	5	0	196
Top (tie)	2	23	0	0	25
Second	7	0	17	0	24
Third and under	5	0	1	13	19
Total	205	23	23	13	264

4.3 Causes of Errors

Table 4 shows the reasons why the correct bilingual pairs were ranked below second place. It indicates that the evidence on the phonological correspondences and the evidence on the nonexistence of the equivalent nouns are the major causes of the incorrect alignments.

Table 4: Causes of Errors

Causes	ED	LRA
Literal Translation	2	6
Phonological Correspondences	12	11
Neighboring Nouns	7	7
Equivalent Noun	19	9
Multiple Causes	3	3
Total	43	36

4.4 Validity of Each Evidence

In order to see how much each evidence contributes to the correct alignments, we repeated experiments four times. In each experiment, in order to obtain weight values, we performed the logistic regression analysis with leaving out one of the four kinds of evidence, and then we made alignment with the use of the weight values.

Table 5 shows the rate of the correct alignments achieved in the condition of leaving out the individual evidence one by one. The numbers in the parentheses are the numbers of the pairs being judged as correct.

Table 5: Validity of Each Evidence

	Top (solo)	Top (inc. tie)	Top Two
With All Evidences	77.65% (205)	86.36% (228)	95.08% (251)
Without Literal Translation	63.64% (168)	81.44% (215)	90.15% (238)
Without Phonological Correspondences	58.33% (154)	79.92% (211)	90.90% (240)
Without Neighboring Nouns	74.24% (196)	91.67% (242)	95.45% (252)
Without Equivalent Noun	76.89% (203)	86.74% (229)	94.70% (250)

The rates of the correct alignments where the evidence on the similarity of literal translations was left out are lower than those where all evidences are used. The drop of the solo possession of first place is particularly noticeable. The same goes for the case where the evidence on the phonological correspondences was not used. Accordingly the two kinds of evidence contribute to improve the rate of the correct alignments.

As compared the case where all evidences were used with the case where the evidence on the similarity of neighboring nouns was left out, the rate at top (including tie) by the latter is higher than that by the former, and the rates at the solo possession of first place and top two by the latter are slightly lower than that by the former. The same goes for the case where the evidence on the existence of equivalent nouns was not used.

The two kinds of evidence contributing the improvement are internal, while those which do not are external. One of the future studies is to explore external evidences which will work effectively.

4.5 Independence among Evidences

In a method of extracting bilingual pairs with the use of several kinds of evidence, it is desirable that the degree of independence among the evidences should be high. We calculated Spearman's rank correlation coefficients to see the degree of independence. The result is shown in Table 6.

Table 6: Spearman's Rank Correlation Coefficient

	LT	P	NN	EN
Literal Translation (LT)	---	0.057	0.048	0.039
Phonological Correspondences (P)	---	---	0.099	0.133
Neighboring Nouns (NN)	---	---	---	-0.034
Equivalent Noun (EN)	---	---	---	---

Table 6 tells that all the coefficients among the evidences are low enough, which suggests the degree of independence is high.

5 Conclusion

This paper proposed a method of extracting English MWNEs which are not found in the dictionary of an English-to-Japanese MT system and appear infrequently in a parallel corpus, and their Japanese counterparts. Our method makes its alignment on the basis of the external evidences provided by the context in which a bilingual pair appears, as well as the internal evidences within the pair.

Each evidence is accompanied by a score and the aggregate score is computed as the weighted sum of the scores. We applied the logistic regression analysis to obtain the appropriate weights.

The experiments with the parallel corpus of Yomiuri Shimbun and the Daily Yomiuri found that 86.36% of the extracted bilingual pairs with the highest scores and the 95.08% with the top two scores were judged to be correct. The results surpass those achieved by using the weight values decided empirically.

References

Al-Onaizan, Y. and Kinght, K.: 2002, Translating Named Entities Using Monolingual and Bilingual Resources, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 400-408.

Chiao, Y. and Zweigenbaum, P.: 2002, Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora, in *Proceedings of the 19th COLING*, pp.1208-1212.

Dagan, I. and Church, K.: 1994, Termight: Identifying and Translating Technical Terminology, in *Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 34-40.

Dekai, W. and Xia, X.: 1994, Learning an English-Chinese Lexicon from a Parallel Corpus, in *Proceedings of the Annual Conference of AMTA*, pp. 206-213.

Eijk, P.: 1993, Automating the Acquisition of Bilingual Terminology, in *Proceedings of the 6th Conference of the EACL*, pp. 113-119.

Fung, P.: 1998, Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora, *Lecture Notes in Artificial Intelligence*, 1529: 1-17.

Huang, F., Vogel, S. & Waibel, A.: 2003, *Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-Feature Cost Minimization.*

Kaji, H. and Aizono, T.: 1996, Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information, in *Proceedings of the 16th COLING*, pp. 23-28.

Kaji, H. and Aizono, T.: 2001, Kyokigo-shugo-no ruijido-nimotoduku taiyaku-kopasu-karano taiyakugo-chushutsu: Extracting Word Translations from Bilingual Corpora Based on Similarity of Co-occurring Word Sets, *Johoshori-gakkai-ronbunshi*, 42(9): 2248-2258, (in Japanese).

Ker, S. and Chang, J.: 1997, A Class-based Approach to Word Alignment, *Computational Linguistics*, 23(2): 312-343.

Kumano, A. and Hirakawa, H.: 1994, Building an MT Dictionary from Parallel Texts Based on Linguistic and Statistical Information, in *Proceedings of the 15th COLING*, pp. 76-81.

Kupiec, J.: 1993, An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora, in *Proceedings of the 31st Annual Meeting of the ACL*, pp. 17-22.

Le, S., Youbing, J., Lin, D. and Yufang, S.: 1999, Word Alignment of English-Chinese Bilingual Corpus Based on Chunks, in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 110-116.

Moore, R. C.: 2003, Learning Translations of Named-Entity Phrases from Parallel Corpora, in *Proceedings of the 41st Annual Meeting of the ACL*.

Nakagawa, H.: 2001, Disambiguation of Compound Noun Translations Extracted from Bilingual Comparable Corpora, in Proceedings of NLPRS, pp.67-74.

Smadja, F. and McKeown, K.: 1996, Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics*, 22(1): 1-38.

Tanaka, T. and Matuo, Y.: 1999, Extraction of Translation Equivalents from Non-Parallel Corpora, in *Proceedings of TMI*, pp.109-119.

Tsuji, K.: 2001, Taiyaku-kopasu-karano teihindo-yakugo-tsui-no chushutsu: hanji/hindo-joho-no togoteki-riyo, in *Dai 49 kai nihon-toshokan-joho-gakkai kenkyu-taikai happyo yoko*, pp. 59-62, (in Japanese).

Utiyama, M. and Isahara, H.: 2003, Reliable Measures for Aligning Japanese-English News Articles and Sentences, in *Proceedings of the 41st Annual Meeting of the ACL*, pp. 72-79.

Watanabe, H.: 1996, A Method for Abstracting Newspaper Articles by Using Surface Clues, in *Proceedings of the 16th COLING*, pp. 974-979.