# A Comparison of Unsupervised Bilingual Term Extraction Methods Using Phrase-Tables

**Masamichi Ideue**†
**Kazuhide Yamamoto**†
†Department of Electrical Engineering,
Nagaoka University of Technology
1603-1 Kamitomioka, Ngaoka,
Niigata 940-2188, Japan
{ideue,yamamoto}@jnlp.org

**Masao Utiyama**‡
**Eiichiro Sumita**‡
‡National Institutre of Information
and Communications Technology
3-5, Hikaridai, Seika, Soraku,
Kyoto 619-0289, Japan
{mutiyama,eiichiro.sumita}
@nict.go.jp

## Abstract

Automatic bilingual term extraction is essential for providing a consistent bilingual term list for human translators engaged in translating a set of documents. We compare three statistical measures for extracting bilingual terms from a phrase-table built from a parallel corpus. We show that these measures extract different bilingual term candidates and a combination of these measures ranks valid bilingual terms highly.

## 1 Introduction

Automatic bilingual term extraction methods have been studied extensively, because bilingual terms are essential in supporting human and machine translation. For example, Itagaki et al. (2007) have proposed a supervised method for extracting bilingual terms from a phrase-table built from parallel corpus using standard statistical machine translation techniques. Tonoike et al. (2005) have used an existing bilingual dictionary to estimate translations for technical terms.

Our objective for extracting bilingual terms from a phrase-table is to provide a consistent bilingual term list for human translators engaged in translating a set of documents. In this case, we have a parallel corpus created from a set of past related documents (e.g., computer manuals). From this parallel corpus, we can extract a set of bilingual terms that should be used by the translators for the documents being translated.

Since these documents are often domain specific, we usually do not have annotated data for training supervised methods nor bilingual dictionaries specific to the documents under translation. Consequently, we need to develop unsupervised methods for extracting bilingual terms from a phrase-table.

In this paper, we compare three measures for extracting bilingual terms. We also propose a term counting method which is suitable for extracting bilingual terms.

## 2 Related Work

There has been a lot of work done on extracting bilingual terms from parallel or comparable corpora (Robitaille et al., 2006; Hjelm, 2007; Fan et al., 2009; Lee et al., 2010). For example, Itagaki et al. (2007) proposed a supervised method for extracting bilingual terms from the phrase-table built from a parallel corpus. They first extracted bilingual term candidates from the phrase-table. Then, they annotated these candidates as *valid* or *invalid* terms. They used this annotated data to train a classifier for discriminating valid or invalid terms in the extracted term candidates. Tonoike et al. (2005) proposed a method using an existing bilingual dictionary. They translated the components of each source language term using the bilingual dictionary and combined these translations to form term candidates. They then validated these term candidates using statistics obtained from comparable corpora.

Since, as stated in the introduction, we need to develop unsupervised methods for extracting bilingual terms, we extract bilingual terms from the phrase-table using statistical measures. In addition, we compare three statistical measures for extracting bilingual terms. Note that Macken et al. (2008)

have also used statistical measures to filter out invalid bilingual term candidates; however, they did not compare their statistical measures against other measures.

## 3 Bilingual term extraction

We extract bilingual terms from a Japanese-English parallel corpus.[1] The overview of our bilingual term extraction method is as follows:

(1) Extract the term candidates, which match specific part-of-speech(POS) patterns (e.g., a single noun or a noun sequence), from the Japanese and English sentences of the given parallel corpus.

(2) Build the phrase-table from the parallel corpus using the Moses toolkit (Koehn et al., 2007).

(3) Extract bilingual term candidates from the phrase table that are included in the term candidates obtained in (1).

(4) Calculate a statistical measure for each candidate term.

(5) Rank the candidates according to the statistical measure, and extract the highly-ranked candidates as valid bilingual terms.

We compare three statistical measures, Score$_F$, Score$_L$ and Score$_C$, for extracting correct bilingual terms.

### 3.1 Extraction of term candidates

We use functions implemented in TermExtract [2] to extract term candidates. TermExtract is a Perl module for extracting terms. We slightly modified the POS patterns used in TermExtract for our purposes. For example, we allowed plural nouns for our term candidates. Note that the POS patterns are specific to each language.

Examples of Japanese and English POS patterns and term candidates are as follows, where the Japanese term candidates are romanized and written in **bold**.

**A noun:** **suso** (hem), underwear

**Noun sequence:** **rinen sozai** (linen material), **purinto kizi** (printed fabric), boxer shorts, chest pocket

**Adjective+Noun:** metallic color, checkered coat

### 3.2 Fisher's exact test

Fisher's exact test has been used by Johnson et al. (2007) to select valid phrase pairs from the phrase-table for statistical machine translation. We use the statistic of Fisher's exact test as Score$_F$ to measure the validity of each bilingual term candidate. The statistic used in Fisher's exact test is defined as Score$_F$ as follows. First, we obtain the contingency table, shown below, for a bilingual term candidate $T_{J,E}$ consisting of Japanese term $J$ and English term $E$

| | |
|---|---|
| $C(J,E)$ | $C(J)-C(J,E)$ |
| $C(E)-C(J,E)$ | $N-C(J)-C(E)+C(J,E)$ |

where $C(J,E)$, $C(J)$, $C(E)$, and $N$ are the numbers of parallel sentences containing $J$ and $E$, Japanese sentences containing $J$, English sentences containing $E$, and all parallel sentences, respectively. Score$_F$ of $T_{J,E}$ is defined as

$$\text{Score}_F = -\log(p\text{-}value)$$

where

$$p\text{-}value = \sum_{k=C(J,E)}^{\infty} P_h(k)$$

$$P_h(C(J,E)) = \frac{\binom{C(J)}{C(J,E)}\binom{N-C(J)}{C(E)-C(J,E)}}{\binom{N}{C(E)}}$$

Note that $P_h(C(J,E))$ is the probability of observing the contingency table under the null hypothesis of $J$ and $E$ being independent of each other. Consequently, Score$_F$ has a high value when they are not independent each other.

### 3.3 Log-likelihood ratio

Fisher's exact test treats the Japanese and English terms in each candidate bilingual term as units for counting $C(J,E)$. In this section, we propose using the word alignments of each candidate term to measure the validity of that term.

Let $T_{J,E}$ be a bilingual term candidate consisting of Japanese term $J$ and English term $E$. Let $J$ and $E$ be composed of $j_1, j_2, \ldots, j_k$ and $e_1, e_2, \ldots, e_l$,

[1] However, our method is language independent.

[2] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

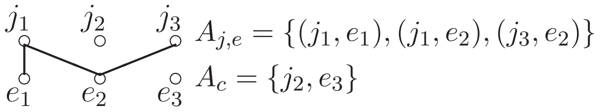


Figure 1: Example of word alignments.

respectively. Then, $A_{j,e}$ is the set of the word alignments in $T_{J,E}$, and $A_c$ is the set of words having no corresponding words. Figure 1 illustrates an example of the word alignments in $T_{J,E}$ composed of $J = \{j_1, j_2, j_3\}$ and $E = \{e_1, e_2, e_3\}$.

Score$_L$ of $T_{J,E}$ is defined as follows:

$$\text{Score}_L(T_{J,E}) = \sum_{(j_k,e_l)\in A_{j,e}} LLR_{j,e}(j_k,e_l|T_{J,E}) + \sum_{c\in A_c} LLR_{c,\varphi}(c,\varphi|T_{J,E})$$

$$LLR_{j,e}(j_k,e_l|T_{J,E}) = \log \frac{P(+1|j_k,e_l)}{1-P(+1|j_k,e_l)}$$

$$P(+1|j_k,e_l) = \frac{C_A(j_k,e_l)+\alpha}{C(j_k,e_l)+2\alpha}$$

$$LLR_{c,\varphi}(c,\varphi|T_{J,E}) = \log \frac{P(+1|c,\varphi)}{1-P(+1|c,\varphi)}$$

$$P(+1|c,\varphi) = \frac{C_\varphi(c)+\alpha}{C(c)+2\alpha}$$

where $C_A(j_k, e_l)$ and $C(j_k, e_l)$ are the numbers of parallel sentences containing the word alignment $(j_k, e_l)$ and parallel sentences containing $j_k$ and $e_l$, respectively. The log-likelihood ratio defined in $LLR_{j,e}(j_k, e_l \mid T_{J,E})$ has a high value when $j_k$ and $e_l$ are aligned often in parallel sentences. $\alpha$ is a smoothing parameter. We use $\alpha = 1$ in this paper. If a word $c$ has no correspondence in $T_{J,E}$, we regard that it corresponds to a null word. $C_\varphi(c)$ and $C(c)$ are the numbers of parallel sentences where $c$ has no correspondence and parallel sentences containing $c$, respectively. $LLR_{c,\varphi}(c, \varphi \mid T_{J,E})$ has a high value when $c$ is not usually aligned. Consequently, Score$_L$ has a high value when the word alignments in $T_{E,J}$ are often aligned and the isolated words in $T_{J,E}$ are usually not aligned.

For example, in Figure 1, Score$_L$ is calculated as $\text{Score}_L(T_{J,E}) = LLR_{j,e}(j_1,e_1|T_{J,E}) + LLR_{j,e}(j_1,e_2|T_{J,E}) + LLR_{j,e}(j_3,e_2|T_{J,E}) + LLR_{c,\varphi}(j_2,\varphi|T_{J,E}) + LLR_{c,\varphi}(e_3,\varphi|T_{J,E})$.

### 3.4 C-value

The C-value (Frantzi et al., 1996) has been used to measure the validity of monolingual term candidates. C-value of term $T$ is defined as:

$$C\text{-}value(T) = (|T|-1)\left(n(T) - \frac{t(T)}{c(T)}\right) \quad (1)$$

where $|T|$ is the number of words in $T$, $n(T)$ is the number of occurrences of $T$ in the monolingual corpus, $t(T)$ is the number of total occurrences of the terms containing $T$ as a substring, and $c(T)$ is the number of the terms containing $T$. Note that if $T$ consists of one word, its C-value is 0.

We use the C-value to measure the validity of bilingual term candidates. We first calculate the C-value of all term candidates in Japanese and English. Next, we rank the term candidates in each language in decreasing order of C-values. As a result, $J$ and $E$ in term candidate $T_{J,E}$ are assigned ranks $R(J)$ and $R(E)$ in each language. Finally, Score$_C$ is defined as $\text{Score}_C(T_{J,E}) = \frac{R(J)+R(E)}{2}$.

### 3.5 Bilingual term counting

We compare two methods for counting the number of occurrences of term $T$. The first involves counting the number of occurrences of term $T$ without regarding where $T$ occurs. The second involves count the number of occurrences of term $T$ only when it occurs alone, i.e., we do not count the number of occurrences of term $T$ when it occurs as a substring of a longer term.

We apply these counting methods to Score$_F$ and Score$_L$ because the C-value has already incorpolated a modified counting method that considers nested-collocations. We use Score$_{F1}$, Score$_{F2}$, Score$_{L1}$ and Score$_{L2}$ to denote Score$_F$ and Score$_L$ with the first and second counting methods.

### 3.6 Combination of measures

As shown in the experiments below, Score$_{F2}$ and Score$_{L2}$ are better than Score$_{F1}$ and Score$_{L1}$. In addition, Score$_{F2}$, Score$_{L2}$, and Score$_C$ extract different bilingual term candidates. Thus, we combine these measures using their ranks. That is, if we let $R(\text{Score}_{F2}(T_{J,E}))$, $R(\text{Score}_{L2}(T_{J,E}))$, and $R(\text{Score}_C(T_{J,E}))$ be the ranks of $T_{J,E}$ according to these measures, we then define Score$_{FLC}$ as $\text{Score}_{FLC}(T_{J,E}) = \frac{R(\text{Score}_{F2}(T_{J,E}))+R(\text{Score}_{L2}(T_{J,E}))+R(\text{Score}_C(T_{J,E}))}{3}$

## 4 Experiments

We extracted bilingual term candidates from a Japanese-English parallel corpus consisting of documents related to apparel products. The parallel corpus consisted of about 60,000 sentences, with 821,310 Japanese words, and 891,120 English words. The number of bilingual term candidates extracted from the phrase-table was 22,543 pairs.

### 4.1 Evaluation of translation quality

We manually evaluated 100 bilingual term candidates that were randomly selected from the top 1,000 candidates for each statistical measure.

| | F1 | L1 | C | F2 | L2 | FLC |
|---|---|---|---|---|---|---|
| A | 43 | 77 | 78 | 71 | 79 | 87 |
| A’ | 25 | 5 | 6 | 18 | 4 | 2 |
| B | 24 | 18 | 14 | 8 | 17 | 11 |
| C | 8 | 0 | 2 | 3 | 0 | 0 |

Table 1: Evaluation of translation quality

Table 1 shows the evaluation results of the translation quality of the bilingual terms. In this table, rows “A”, “A”’, “B” and “C” indicate that the bilingual terms are “correct”, “correct depending on contexts”, “partly correct”, and “incorrect”, respectively. Columns “F1”, “F2”, “L1”, “L2”, “C”, and “FLC” mean Score$_{F1}$, Score$_{F2}$, Score$_{L1}$, Score$_{L2}$, Score$_{C}$, and Score$_{FLC}$, respectively. Table 2 shows examples of the bilingual term extracted by Score$_{F2}$, Score$_{L2}$, and Score$_{C}$. Table 1 shows that Score$_{F1}$ is inferior to the other methods. The number of As was 43 and was statistically significantly ($p < 0.01$) less than those obtained by other measures based on the two-sided proportional test. While the differences among other measures were not statistically significant, Score$_{FLC}$ had the best translation quality and the differences with Score$_{L2}$ were almost statistically significant ($p = 0.066$).

Form this table, we concluded that we can use unsupervised methods to extract bilingual terms with high accuracy.

### 4.2 Comparison of measures

As shown in Table 1, Score$_{FLC}$ was the best of all measures. We attribute this to the diversity of bilingual terms extracted by each of the measures, as described below, and the bilingual term having each characteristic has high accuracy.

We investigated the correlation between the rankings of the candidates extracted by each measure.

| | L1 | C | F2 | L2 |
|---|---|---|---|---|
| F1 | 0.332 | 0.075 | 0.704 | 0.288 |
| L1 | | 0.313 | 0.377 | 0.854 |
| C | | | 0.128 | 0.312 |
| F2 | | | | 0.422 |

Table 3: Rank correlation coefficients between rankings of each measure.

| F2 ∩ L2 | F2 ∩ C | L2 ∩ C | F2 ∩ L2 ∩ C |
|---|---|---|---|
| 369 | 190 | 423 | 187 |

Table 4: Number of common bilingual terms in ranking of each measure (Top 1,000).

Table 3 shows the Kendall’s rank correlation coefficients between these measures. As shown in the table, correlations among these measures were not high except for (F1, F2) and (L1, L2). This indicates that these measures extracted different bilingual term candidates. Indeed, as shown in Table 4, there are few common bilingual terms in the top 1000 of each measure’s ranking. The 187 bilingual terms in F2∩L2∩C should have characteristics from each measure and high translation quality.

Next, we investigated the characteristics of the extracted bilingual terms by each measure. Figures 2(a) and 2(b) show the average number of words and occurrences of bilingual terms for every 1000 ranks. The distribution of English words (not shown) was similar to those of Japanese words.

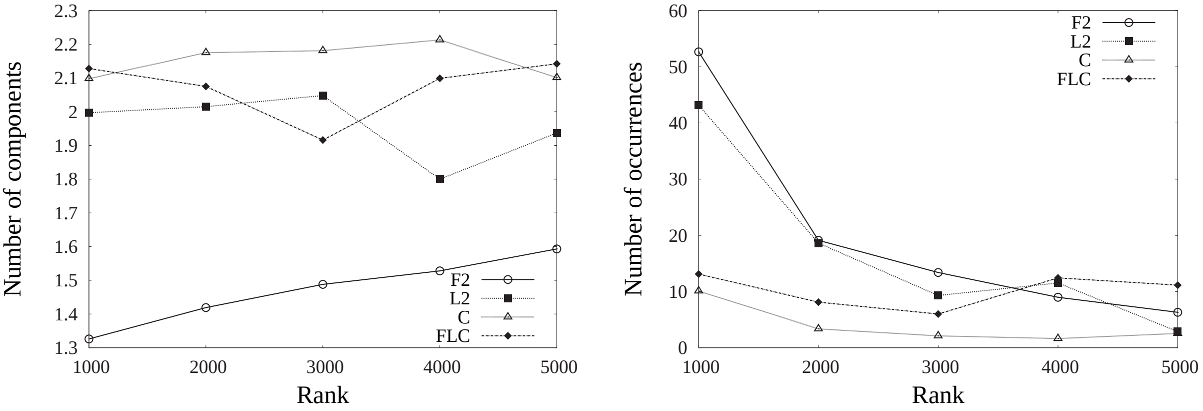
Figure 2 shows that Score$_{F2}$ extracted bilingual terms that have few words and high occurrence frequency.

Score$_{L2}$ also extracted bilingual terms that have high occurrence, but ones with many words. This is because Score$_{L2}$ is the sum of the log-likelihood ratios of the words in the corresponding term. That is, bilingual terms that have many words with established word alignments are ranked highly in Score$_{L2}$ ranking.

In Score$_{C}$, the bilingual terms that have few occurrences and many words were extracted. As already mentioned in Section 3.4, the C-value assigns 0 to the term candidates consisting of one word. This is the reason why the bilingual terms consisting of more than one word are highly ranked in Score$_{C}$ ranking. Although C-value does not consider the bilingual relationship of the candidate terms, it ex-

| | F2 | L2 | C |
|---|---|---|---|
| A | **daiya**⇔diamond<br>**orizinaru botan**⇔original buttons<br>**wanpi-su**⇔one-piece dress | **daun jaketto**⇔down jacket<br>**kata osi reza-**⇔embossed leather<br>**kotton zi**⇔cotton fabric | **kitake nagame**⇔long length<br>**ga-ze sozai**⇔gauze material<br>**guren tyekku**⇔glen plaid |
| A’ | **siagari**⇔finish<br>**pointo**⇔accent<br>**gara**⇔patterns | **kisetu kan**⇔seasonal look<br>**iro zukai**⇔coloring<br>**koukyuu kan**⇔high quality touch | **kobana gara**⇔floral pattern<br>**pasu ke-su**⇔card case<br>**sozai kan**⇔unique look |
| B | **uesuto bubun** (waist part)<br>⇔waist<br>**iro** (color)<br>⇔different colors | **konbou sozai** (blend material)<br>⇔blend<br>**akusento** (accent)<br>⇔nice accent | **iro oti** (faded color)<br>⇔faded look<br>**kinou sei** (functionality)<br>⇔terms of functionality |
| C | **sodeguti** (cuff)<br>⇔hem<br>**hadazawari** (feel)<br>⇔comfortable | | **siruetto bodi-** (body silhouette)<br>⇔item features<br>**mo-do kan** (fashion sense)<br>⇔new model |

Table 2: Examples of the extracted bilingual term and their evaluation.



(a) Number of Japanese words in a term

(b) Number of occurrences

Figure 2: Characteristics of the extracted bilingual term by each measures (Top 5,000).

tracted precise translation candidates as shown in Table 1. This indicates that C-value can filter out noisy bilingual term candidates from the phrase-table.

$\text{Score}_{FLC}$, $\text{Score}_{F2}$, $\text{Score}_{L2}$, and $\text{Score}_{C}$ filter each other the noisy bilingual term extracted by each score. The characteristic of $\text{Score}_{FLC}$ indicated a tendency similar to $\text{Score}_{C}$ in Figure 2. From this, the $\text{Score}_{C}$’s residual noise was filtered by $\text{Score}_{F2}$ and $\text{Score}_{L2}$. If we want to extract the bilingual term having the $\text{Score}_{C}$’s characteristic, we can extract the bilingual term that is more accurate than $\text{Score}_{C}$ by using $\text{Score}_{FLC}$.

| | Bilingual term | F1 | F2 |
|---|---|---|---|
| incorrect | **ringu**⇔coloring | 35 | 21,676 |
| correct | **suri-bu**⇔sleeve | 749 | 5,433 |

Table 5: Bilingual term extracted by $\text{Score}_{F1}$ whereas $\text{Score}'_{F2}$ did not extract, and its rank (The Japanese term candidates are highlighted in bold).

### 4.3 Effectiveness of substring consideration

Table 5 shows examples of the extracted bilingual terms. The first example is wrongly extracted by $\text{Score}_{F1}$ but is correctly extracted by $\text{Score}_{F2}$ as **kara- ringu** and *coloring*. $\text{Score}_{F1}$ wrongly extracted ringu (ring) instead of **kara- ringu** because **ringu** is a substring of **kara- ringu**. This means that the co-occurrence frequency of **ringu** and *coloring* is larger than that of **kara- ringu** and *coloring*. As a result, it extracted the wrong bilingual term. This shows that the counting method considering the substrings are useful for eliminating the bilingual term which tends to occur as the substrings of other bilingual terms.

The second example represents the correct bilingual terms that were not extracted by the Score$_{F2}$ due to their low frequency of co-occurrence. The **suri-bu**⇔*sleeve* pair had a strong tendency to be substrings in our experiments. In fact, the Japanese term candidates included 38 candidates containing **suri-bu** as a substring, while the English term candidates included 49 candidates containing *sleeve* as a substring. As a result, the co-occurrence frequency used in Score$_{F2}$ became lower than that in Score$_{F1}$.

## 5 Conclusion

In this study, we compared three statistical measures for extracting bilingual terms from the phrasetable built from a parallel corpus. Each measure extracts different bilingual term candidates. Specifically, each method differs in the number of words extracted and the occurrences of bilingual terms. Consequently, the combination of these measures ranks valid bilingual terms highly.

## References

Katerina T. Frantzi and Sophia Ananiadou. 1996. Extracting Nested Collocations. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*. pp.41–46

Xiaorong Fan, Nobuyuki Shimizu and Hiroi Nakagawa. 2009. AutomaticExtraction of Bilingual Terms From A Chinese-Japanese Parallel Corpus. In *Proceedings of the 3rd International Universal Communication Symposium (IUCS)*. pp.41–45

Hans Hjelm. 2007. Identifying Cross Language Term Equivalents Using Statistical Machine Translation and Distributional Association Measures. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*. pp.97–104.

Masaki Itagaki, Takako Aikawa and Xiaodong He. 2007. Automatic Validation of Terminology Translation Consistency with Statistical Method. In *Proceedings of the Machine Translation Summit XI (MT summit)*, pp.269–274.

J Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp.967–975.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics(ACL)*, pp.177–180.

Lianhau Lee, Aiti Aw, Min Zhang, and Haizhou Li. 2010. EM-based Hybrid Model for Bilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics(COLING)*, pp.639–646

Lieve Macken, Els Lefever, and Veronique Hoste. 2008. Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus. In *Proceedings of the 22nd International Conference on Computational Linguistics(COLING)*, pp.529–536

Xavier Robitaille, Yasuhiro Sasaki, Masatsugu Tonoike, Satoshi Sato, and Takehito Utsuro. 2006. Compiling French-Japanese Terminologies from the Web. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp.225–232.

Masatsugu Tonoike, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2005. Translation Estimation for Technical Terms using Corpus collected from the Web. In *Proceedings of the Pacific Association for Computational Linguistics (PACLING)*, pp.325–331.