Retrieving Bilingual Verb–noun Collocations by Integrating Cross-Language Category Hierarchies

Fumiyo Fukumoto  Yoshimi Suzuki  Kazuyuki Yamashita∗
Interdisciplinary Graduate School of Medicine and Engineering
Univ. of Yamanashi
{fukumoto, ysuzuki}@yamanashi.ac.jp

∗The Center for Educational Research
Faculty of Education and Human Sciences
Univ. of Yamanashi
dkazuyuki@yamanashi.ac.jp

Abstract

This paper presents a method of retrieving bilingual collocations of a verb and its objective noun from cross-lingual documents with similar contents. Relevant documents are obtained by integrating cross-language hierarchies. The results showed a 15.1% improvement over the baseline non-hierarchy model, and a 6.0% improvement over use of relevant documents retrieved from a single hierarchy. Moreover, we found that some of the retrieved collocations were domain-specific.

1 Introduction

A bilingual lexicon is important for cross-lingual NLP applications, such as CLIR, and multilingual topic tracking. Much of the previous work on finding bilingual lexicons has made use of comparable corpora, which exhibit various degrees of parallelism. Fung et al. (2004) described corpora ranging from noisy parallel, to comparable, and finally to very non-parallel. Obviously, the latter are easy to collect because very non-parallel corpora consist of sets of documents in two different languages from the same period of dates. However, a good solution is required to produce a higher quality of lexicon retrieval.

In this paper, we focus on English and Japanese bilingual verb–objective noun collocations which we call verb–noun collocations and retrieve them using very non-parallel corpora. The method first finds cross-lingual relevant document pairs with similar contents from non-parallel corpora, and then we estimate bilingual verb–noun collocations within these relevant documents. Relevant documents are defined here as pairs of English and Japanese documents that report identical or closely related contents, e.g., a pair of documents describing an aircraft crash and the ensuing investigation to compensate the victims’ families or any safety measures proposed as a result of the crash. In the task of retrieving cross-lingual relevant documents, it is crucial to identify an event as something occurs at some specific place and time associated with some specific action. One solution is to use a topic, i.e., category in the hierarchical structure, such as Internet directories. Although a topic is not an event, it can be a broader class of event. Therefore, it is helpful for retrieving relevant documents, and thus bilingual verb–noun collocations.

Consider the Reuters’96 and Mainichi newspaper documents shown in Figure 1. The documents report on the same event, “Russian space station collides with cargo craft,” were published within two days of each other, and have overlapping content. Moreover, as indicated by the double-headed arrows in the figure, there are a number of bilingual collocations. However, as shown in Figure 1, the Reuters document is classified into “Science and Technology,” while the Mainichi document is classified into “Space Navigation”. This is natural because categories in the hierarchical structures are defined by different human experts. Therefore, a hierarchy tends to have some bias in both defining hierarchical structure and classifying documents, and as a result some hierarchies written in one language are coarse-grained, while others written in other languages are fine-grained. Our attempt using the results of integrating different hierarchies for retrieving relevant documents was postulated to be able to solve this defect of the differences in
hierarchies, and to improve the efficiency and efficacy of retrieving collocations.

2 System Description

The method consists of three steps: integrating category hierarchies, retrieving cross-lingual relevant documents, and retrieving collocations from relevant documents.

2.1 Integrating Hierarchies

The method for integrating different category hierarchies does not simply merge two different hierarchies into a large hierarchy, but instead retrieves pairs of categories, where each category is relevant to each other. The procedure consists of two substeps: Cross-language text classification (CLTC) and estimating category correspondences.

2.1.1 Cross-language text classification

The corpora we used are the Reuters’96 and the RWCP of the Mainichi Japanese newspapers. In the CLTC task, we used English and Japanese data to train the Reuters’96 categorical hierarchy and the Mainichi UDC code hierarchy (Mainichi hierarchy), respectively. In the Reuters’96 hierarchy, the system was trained using labeled English documents, and classified translated labeled Japanese documents. Similarly, for Mainichi hierarchy, the system was trained using labeled Japanese documents, and classified translated labeled English documents. We used Japanese-English and English-Japanese MT software.

We used a learning model, Support Vector Machines (SVMs) (Vapnik, 1995), to classify documents, as SVMs have been shown to be effective for text classification. We used the “One-against-the-Rest” version of the SVMs at each level of a hierarchy. We classify test documents using a hierarchy by learning separate classifiers at each internal node of the hierarchy. We used a Boolean function $b(L_1) \& \cdots \& b(L_m)$, where $b(L_i)$ is a decision threshold value of the $i$-th hierarchical level. The process is repeated by greedily selecting subbranches until a leaf is reached.

We classified translated Mainichi documents with Mainichi category $m$ into Reuters categories using SVMs classifiers. Similarly, each translated Reuters document with category $r$ was classified into Mainichi categories. Figure 2 illustrates the classification of Reuters and Mainichi documents. A document with Mainichi category “$m1$” is classified into Reuters category “$r12$”, and a document with Reuters category “$r1$” is classified into Mainichi category “$m21$”. As a result, we obtained category pairs, e.g., (r12, m1), and (m21, r1), from the documents assigned to the categories in each hierarchy.

2.1.2 Estimating category correspondences

The assumption of category correspondences is that semantically similar categories, such as “Equity markets” and “Bond markets” exhibit similar statistical properties than dissimilar categories, such as “Equity markets” and “Sports”. We applied $\chi^2$ statistics to the results of CLTC. Let us take a look at the Reuters’96 hierarchy. Sup-
pose that the translated Mainichi document with Mainichi category \( m \in M \) (where \( M \) is a set of Mainichi categories) is assigned to Reuters category \( r \in R \) (\( R \) is a set of Reuters’96 categories). We can retrieve Reuters and Mainichi category pairs, and estimate category correspondences according to the \( \chi^2 \) statistics shown in Eq. (1).

\[
\chi^2(r, m) = \frac{f(r, m) - E(r, m)}{E(r, m)} = \frac{m_f - m_{r,m}}{m_{r,m}}
\]

where \( E(r, m) = S_r \times \frac{S_m}{S_R} \).

\( S_r = \sum_{k \in M} f(r, k) \), \( S_R = \sum_{r \in R} S_r \).

Here, the co-occurrence frequency of \( r \) and \( m \), \( f(r, m) \) is equal to the number of category \( m \) documents assigned to \( r \). Similar to the Reuters hierarchy, we can estimate category correspondences from Mainichi hierarchy, and extract a pair \((r, m)\) according to the \( \chi^2 \) value. We note that the similarity obtained by each hierarchy does not have a fixed range. Thus, we apply the normalization strategy shown in Eq. (2) to the results obtained by each hierarchy to bring the similarity value into the range \([0,1]\).

\[
\chi^2_{new}(r, m) = \frac{\chi^2_{old}(r, m) - \chi^2_{min}(r, m)}{\chi^2_{max}(r, m) - \chi^2_{min}(r, m)}.
\]

Let \( SP_r \) and \( SP_m \) are a set of pairs obtained by Reuters hierarchy and Mainichi hierarchy, respectively. We construct the set of \( r \) and \( m \) category pairs, \( SP_{(r,m)} = \{(r,m) \mid (r,m) \in SP_r \cap SP_m\} \), where each pair is sorted in descending order of \( \chi^2 \) value. For each pair of \( SP_{(r,m)} \), if the value of \( \chi^2 \) is higher than a lower bound \( L_\chi \), two categories, \( r \) and \( m \), are regarded as similar.\(^2\)

\[2\]

2.2 Retrieval of Relevant Documents

We used the results of category correspondences from the Reuters and Mainichi hierarchies to retrieve relevant documents. Recall that we used English and Japanese documents with quite different hierarchical structures. The task thus consists of two criteria: retrieving relevant documents based on English (we call this Int\_hi \& Eng) and in Japanese (Int\_hi \& Jap). Let \( d^r_i \) (\( 1 \leq i \leq s \)) be a Reuters document that is classified into the Reuters category \( r \). Let \( d^m_j \) (\( 1 \leq j \leq t \)) be a Mainichi document that belongs to the Mainichi category \( m \). Here, \( s \) and \( t \) are the number of documents classified into \( r \) and \( m \), respectively. Each Reuters document \( d^r_i \) is translated into a Japanese document \( d^{r\_mt}_i \) by an MT system. Each Mainichi document \( d^m_j \) is translated into an English document \( d^{m\_nt}_j \).

Retrieving relevant documents itself is quite simple. As illustrated in Figure 3, in “Int\_hi \& Eng” with a set of similar categories consisting of \( r \) and \( m \), for each Reuters and translated Mainichi document, we calculate BM25 similarities between them.

\[
BM25(d^r_i, d^{m\_nt}_j) = \sum_{w \in d^r_i} \theta_w \frac{(k_1 + 1)qtf}{k_3 + qtf} \frac{(k_3 + 1)qtf}{k_3 + qtf},
\]

where \( w \) is a word within \( d^{m\_nt}_j \), and \( \theta_w \) is the weight of \( w \), \( \theta_w = \log \left( \frac{N-n+0.5}{(n+0.5)} \right) \). \( N \) is the number of Reuters documents within the same category \( r \), and \( n \) is the number of documents which contains \( w \). \( K \) refers to \( k_1 ((1 - b) + b \frac{dl}{avdl}) \), \( k_1 \), \( b \), and \( k_3 \) are parameters and set to 1, 1, and 1,000, respectively. \( dl \) is the document length of \( d^r_i \) and \( avdl \) is the average document length in words. \( tf \) and \( qtf \) are the frequency of occurrence of \( w \) in \( d^r_i \) and \( d^{m\_nt}_j \), respectively. If the similarity value between them is higher than a lower bound \( L_\theta \), we regarded these as relevant documents. The procedure is applied to all documents belonging to the sets of similar categories. “Int\_hi \& Jap” is the same as “Int\_hi \& Eng” except for the use of \( d^{r\_mt}_i \) and \( d^m_j \) for comparison. We compared the performance of these tasks, and found that “Int\_hi \& Eng” was better than “Int\_hi \& Jap”. In section 3, we show results with “Int\_hi \& Eng” due to lack of space.
2.3 Acquisition of Bilingual Collocations

The final step is to estimate bilingual correspondences from relevant documents. All Japanese documents were parsed using the syntactic analyzer Cabocha (Kudo and Matsumoto, 2003). English documents were parsed with the syntactic analyzer Lin (1993). In both English and Japanese, we extracted all the dependency triplets(obj, n, v).

Here, n refers to a noun which is an object of a verb v in a sentence.\(^3\) Hereafter, we describe the Reuters English dependency triplet as vn_r, and that of Mainichi as vn_m. The method to retrieve bilingual correspondences consists of two sub-steps: document-based retrieval and sentence-based retrieval.

2.3.1 Document-based retrieval

We extract vn_r and vn_m pairs from the results of relevant documents:

\[
\{vn_r, vn_m\} \quad \text{s.t.} \quad 3d^1_r \ni vn_r, \quad 3d^m_m \ni vn_m
\]

BM25(d^1_r, d^m_m) \geq L_\theta. \quad (4)

Next, we estimate the bilingual correspondences according to the \(\chi^2(vn_r, vn_m)\) statistics shown in Eq. (1). In Eq. (1), we replace r by vn_r and m by vn_m. \(f(r, m)\) is replaced by \(f(vn_r, vn_m)\), i.e., the co-occurrence frequency of vn_r and vn_m.

2.3.2 Sentence-based retrieval

We note that bilingual correspondences obtained by document-based retrieval are not reliable. This is because many verb–noun collocations appear in a pair of relevant documents, as can be seen from Figure 1. Therefore, we applied sentence-based retrieval to the results obtained by document-based retrieval. First, we extract vn_r and vn_m pairs the \(\chi^2\) values of which are higher than 0. Next, for each vn_r and vn_m pair, we assign sentence-based similarity:

\[
S_{\text{sim}}(vn_r, vn_m) = \max_{S_{\text{sim}}(vn_r, vn_m)} \text{sim}(S_{\text{vn}_r}, S_{\text{vn}_m}). \quad (5)
\]

Here, \(Set_r\) and \(Set_m\) are a set of sentences that include vn_r and vn_m, respectively. The similarity between \(S_{\text{vn}_r}\) and \(S_{\text{vn}_m}\) is shown in Eq. (6).

\[
sim(S_{\text{vn}_r}, S_{\text{vn}_m}) = \frac{2 \cdot \text{co}(S_{\text{vn}_r} \cap S_{\text{vn}_m})}{|S_{\text{vn}_r}| + |S_{\text{vn}_m}| - 2 \cdot \text{co}(S_{\text{vn}_r} \cap S_{\text{vn}_m}) + 2} \quad (6)
\]

where \(|X|\) is the number of content words in a sentence \(X\), and \(\text{co}(S_{\text{vn}_r} \cap S_{\text{vn}_m})\) refers to the number of content words that appear in both \(S_{\text{vn}_r}\) and \(S_{\text{vn}_m}\). \(S_{\text{vn}_r}\) is a translation result of \(S_{\text{vn}_m}\). We retrieved vn_r and vn_m as a bilingual lexicon that satisfies:

\[
\{vn_r, vn_m\} = \arg\max_{\{vn_r, vn_m\} \in BP(vn_m)} S_{\text{sim}}(vn_r, vn_m), \quad (7)
\]

where \(BP(vn_m)\) is a set of bilingual verb–noun pairs, each of which includes vn_m on the Japanese side.

3 Experiments

3.1 Integrating hierarchies

3.1.1 Experimental setup

We used Reuters’96 and UDC code hierarchies. The Reuters’96 corpus from 20th Aug. 1996 to 19th Aug. 1997 consists of 806,791 documents organized into coarse-grained categories, i.e., 126 categories with a four-level hierarchy. The RWCP corpus labeled with UDC codes selected from 1994 Mainichi newspaper consists of 27,755 documents organized into a fine-grained categories, i.e., 9,951 categories with a seven-level hierarchy (RWCP., 1998). We used Japanese-English and English-Japanese MT software (Internet Honyaku-uno-Ousama for Linux, Ver.5, IBM Corp.) for CLTC. We divided both Reuters’96 (from 20th Aug. 1996 to 19th May 1997) and RWCP corpora into two equal sets: a training set to train SVM classifiers, and a test set for TC to generate pairs of similar categories. We divided the test set into two parts: the first was used to estimate thresholds, i.e., a decision threshold \(b\) used in CLTC, and lower bound \(L_{\chi^2}\); and the second was used to generate pairs of similar categories using the threshold. We chose \(b = 0\) for each level of a hierarchy. The lower bound \(L_{\chi^2}\) was .003.

We selected 109 categories from Reuters and 4,739 categories from Mainichi, which have at least five documents in each set. We used content words for both English and Japanese documents. We compared the results obtained by hierarchical approach to those obtained by the flat
For evaluation of category correspondences, we used F1-score (F1) which is a measure that balances precision (Prec) and recall (Rec). Let Cor be a set of correct category pairs. The precise definitions of the precision and recall of the task are given below:

\[
\text{Prec} = \frac{\left\{ (r, m) \mid (r, m) \in \text{Cor}, \chi^2(r, m) \geq L_{\chi^2} \right\}}{\left\{ (r, m) \mid \chi^2(r, m) \geq L_{\chi^2} \right\}}
\]

\[
\text{Rec} = \frac{\left\{ (r, m) \mid (r, m) \in \text{Cor}, \chi^2(r, m) \geq L_{\chi^2} \right\}}{\left\{ (r, m) \mid (r, m) \in \text{Cor} \right\}}
\]

### 3.1.2 Results

Table 1 shows F1 of category correspondences with \(L_{\chi^2} = .003\). “Mai & Reu” shows the results obtained by our method. “Mai” and “Reu” show the results using only one hierarchy. For example, “Mai” shows the results in which both Mainichi and translated Reuters documents are classified into categories with Mainichi hierarchy, and estimated category correspondences.

Integrating hierarchies is more effective than only a single hierarchy. Moreover, we found advantages in the F1 for the hierarchical approach (“Hierarchy” in Table 1) in comparison with a baseline flat approach (“Flat”). We note that the result of “Mai” was worse than that of “Reu” in both approaches. One reason is that the accuracy of TC. The micro-average F1 of TC for Reuters hierarchy was .815, while that of Mainichi was .673, as Mainichi hierarchy consists of many categories, and the number of training data for each category were smaller than those of Reuters. The results obtained by our method depend on the performance of TC. Therefore, it will be necessary to examine some semi-supervised learning techniques to improve classification accuracy.

### 3.2 Relevant document retrieval

#### 3.2.1 Experimental setup

The training data for choosing the lower bound \(L_\theta\) used in the relevant document retrieval is Reuters and RWCP from 13th to 21st Jun. 1997. The difference in dates between them is less than \(\pm 3\) days. For example, when the date of the RWCP document is 18th Jun., the corresponding Reuters date is from 15th to 21st Jun. We chose \(L_\theta\) that maximized the average F1 among them.

Table 2 shows the test data, i.e., the total number of collected documents and the number of related documents collected manually for the evaluation. We implemented the following approaches including related work, and compared these results with those obtained by our methods, Int\_hi & Eng.

1. **No\_hierarchy**: Categories with each hierarchy are not used in the approach. The approach is the same as the method reported by Collier et al. (1998) except for term weights and similarities. We calculate similarities between Reuters and translated Mainichi documents, where the difference in dates is less than \(\pm 3\) days. (No\_hi & Eng).

2. **Hierarchy**: The approach uses only Reuters hierarchy (we call this Reu Hierarchy). Reuters documents and translated Mainichi documents are classified into categories with Reuters hierarchy. We calculate BM25 between Reuters and Mainichi documents within the same category. The procedure is applied for all categories of the hierarchies.

The judgment of relevant documents was the same as our method: if the value of similarity between two documents is higher than a lower bound \(L_\theta\), we regarded them as relevant documents.

#### 3.2.2 Results

The retrieval results are shown in Table 3 and Figure 4. Table 3 shows best performance of each method against \(L_\theta\). As can be seen clearly from Table 3 and Figure 4, the results with integrating hierarchies improved overall performance.

---

4 The classification was determined to be correct if the two human judges agreed on the evaluation.

5 The classification was determined by two human.
Table 3: Retrieval performance

| Approach            | Prec | Rec  | F1-score | $L_\theta$ |
|---------------------|------|------|----------|------------|
| No hi & Eng         | .417 | .322 | .363     | 40         |
| Reu Hierarchy       | .356 | .544 | .430     | 20         |
| Int hi & Eng        | .839 | .585 | .689     | 20         |

Table 4: # of documents vs $L_\theta$

| Approach            | Lower Bound $L_\theta$ | Pairs       | Eng | Jap |
|---------------------|-------------------------|-------------|-----|-----|
| No hi & Eng         | 100                     | 3,042,166   | 428,042 | 70,080 |
|                     | 80                      | 3,042,166   | 428,042 | 70,080 |
|                     | 60                      | 3,042,166   | 428,042 | 70,080 |
|                     | 40                      | 3,042,166   | 428,042 | 70,080 |
|                     | 20                      | 3,042,166   | 428,042 | 70,080 |
| Reu Hierarchy       | 100                     | 27,181,243  | 43,0181 | 99,452 |
|                     | 80                      | 27,181,243  | 43,0181 | 99,452 |
|                     | 60                      | 27,181,243  | 43,0181 | 99,452 |
|                     | 40                      | 27,181,243  | 43,0181 | 99,452 |
|                     | 20                      | 27,181,243  | 43,0181 | 99,452 |
| Int hi & Eng        | 100                     | 81,904,243  | 45,965 | 654,787 |
|                     | 80                      | 81,904,243  | 45,965 | 654,787 |
|                     | 60                      | 81,904,243  | 45,965 | 654,787 |
|                     | 40                      | 81,904,243  | 45,965 | 654,787 |
|                     | 20                      | 81,904,243  | 45,965 | 654,787 |

3.3 Bilingual Verb–noun Collocations

Finally, we report the results of bilingual verb–noun collocations.

3.3.1 Experimental setup

The data for relevant document retrieval was the Reuters and Mainichi corpora from the same period, i.e., 20th Aug. 1996 to 19th Aug. 1997. The total number of Reuters documents was 806,791, and that of Mainichi was 119,822. As the number of Reuters documents was far greater than that of Mainichi documents, we estimated collocations from the results of cross-lingually retrieving relevant English documents with Japanese query documents. The difference in dates between them was less than ±3 days. Table 5 shows retrieved relevant documents that showed best performance of each method against $L_\theta$. From these data, we extracted bilingual verb–noun collocations.

3.3.2 Results

Table 6 shows the numbers of English and Japanese monolingual verb–noun collocations, those of candidate collocations against which bilingual correspondences were estimated, and those of correct collocations. “D & S” of candidate collocations indicates the number of collocations when we applied both document- and sentence-based retrieval. “Doc” indicates the number of collocations when we applied only document-based retrieval. “D & S” and “Doc” of correct collocations show the number of correct collocations in the topmost 1,000 according to sentence similarity and the $\chi^2$ statistics, respectively. As shown in
Table 6, the results obtained by integrating hierarchies showed a 15.1% (32.8 - 17.7) improvement over the baseline non-hierarchy model, and a 6.0% (32.8 - 26.8) improvement over use of a single hierarchy. We manually compared those 328 bilingual collocations with an existing bilingual lexicon where 78 of them (23.8%) were not included in it. Moreover, 168 of 328 (51.2%) were not correctly translated by Japanese-English MT software. These observations clearly support the usefulness of the method.

It is very important to compare the column “rate” for the numbers of candidate collocations with that for the numbers of correct collocations. In all approaches, sentence-based retrieval was effective in removing useless collocations, especially in our method, about 1.5% of the size obtained by “Doc” was retrieved, while about 4.6(328/72) times the number of correct collocations were obtained in the topmost 1,000 collocations. These observations showed that sentence-based retrieval contributes to a marked reduction in the number of useless collocations without a decrease in accuracy.

The last column in Table 6 shows the results using Inverse Rank Score (IRS), which is a measure of system performance by considering the rank of correct bilingual collocations within the candidate collocations. It is the sum of the inverse rank of each matching collocations, e.g., correct collocations by manual evaluation matches at ranks 2 and 4 give an IRS of \(\frac{1}{2} + \frac{1}{4} = 0.75\). With at most 1,000 collocations, the maximum IRS score is 7.485, and the higher the IRS value, the better the system performance. As shown in Table 6, the performance by integrating hierarchies was much better than that of the non-hierarchical approach, and slightly better than those obtained by a single hierarchy. However, correct retrieved collocations were different from each other. Table 7 lists examples of bilingual collocations obtained by a single hierarchy and integrating hierarchies. The category is “Sport”.\(^8\) \(x,y\) of category pair in Table 7 refer to Reuters and Mainichi category correspondences. Examples in Table 7 denote only English verb-noun collocations.

It is interesting to note that 12 of 154 collocations, such as “earn medal” and “block shot” obtained by integrating hierarchies were also obtained by a single hierarchy approach. However, other collocations such as “get strikeout” and “make birdie” which were obtained in a particular category (Sport, Baseball) and (Sport, Golf), did not appear in either of the results using a single hierarchy or a non hierarchical approach. These observations again clearly support the usefulness of our method.

### 4 Previous Work

Much of the previous work on finding bilingual lexicons used comparable corpora. One attempt involved directly retrieving bilingual lexicons from corpora. One approach focused on extracting word translations (Gaussier et al., 2004). The techniques were based on the idea that semantically similar words appear in similar contexts. Unlike parallel corpora, the position of a word in a document is useless for translation into the other language. In these techniques, the frequency of words in the monolingual document is calculated and their contextual similarity is measured across languages. Another approach focused on sentence extraction (Fung and Cheung, 2004). One limitation of all these methods is that they need to control the experimental evaluation to avoid estimation of every bilingual lexicon appearing in comparable corpora.

The alternative consists of two steps: first, cross-lingual relevant documents are retrieved from comparable corpora, then bilingual term correspondences within these relevant documents are estimated. Thus, the accuracy depends on the performance of relevant documents retrieval. Much of the previous work in finding relevant documents used MT systems or existing bilingual lexicons to translate one language into another. Document pairs are then retrieved using some measure of document similarity. Another approach to retrieving relevant documents involves the collection of relevant document URLs from the WWW (Resnik and Smith, 2003). Utsuro et al. (2003) proposed a method for acquiring bilingual lexicons that involved retrieval of relevant English and Japanese documents from news sites on the WWW. Our work is also applicable to retrieval of relevant documents on the web because it estimates every bilingual lexicon only appearing in

\(^8\)We used an existing bilingual lexicon, Eijiro on the Web, 1.91 million words, (http://www.alc.co.jp) for evaluation. If collocations were not included, the estimation was determined by two human judges.

\(^7\)The number of words in the Japanese-English dictionary (Internet Honyaku-no-Ousama for Linux, Ver.5, IBM Corp.) was about 250,000.

\(^8\)We obtained 98 category pairs in the Sport category.
Table 6: Numbers of monolingual and bilingual verb–noun collocations

| Approach & \((L_\theta)\) | Category pair | Monolingual patterns | Candidate collocations | # of collocations | # of Correct collocations (top 1,000) | Inverse rank score (top 1,000) |
|--------------------------|---------------|-----------------------|------------------------|------------------|---------------------------------------|-------------------------------|
|                          |               | Jap / Eng             | D & S / Doc            |                  | D & S / Doc                           | D & S / Doc                   |
| NoHi & Eng (40)          |               | 25,163 / 44,762       | 25,163 / 6,976,214     | 25,163           | 6,976,214                             | 1.35                         |
| Reu Hierarchy (20)       |               | 10,576 / 37,022       | 10,576 / 1,272,102     | 10,576           | 1,272,102                             | 2.24                         |
| IntHi & Eng (20)         |               | 8,347 / 21,524        | 8,347 / 560,472        | 8,347            | 560,472                               | 2.33                         |

Table 7: Examples of bilingual verb–noun collocations

| Approach & \((L_\theta)\) | Category or category pair | No of collocations | # of correct collocations(%) | Examples (English) |
|--------------------------|----------------------------|--------------------|-------------------------------|--------------------|
| Reu Hierarchy (20)       | (Sport, Baseball)         | 110 / 8,838        | 24(21.8)                      | create chance, *earn medal*, feel pressure, block shot, establish record, take chance |
|                          | (Sport, Relay)            | 177 / 3,418        | 18(10.2)                      | get strikeout, leave base, throw pitch, lead ranking, run km, win athletic |
|                          | (Sport, Tennis)           | 115 / 2,636        | 32(27.8)                      | lose prize, money, play exhibition game, make birdie, have birdie, hole putt, miss putt |
|                          | (Sport, Golf)             | 131 / 2,654        | 28(21.4)                      | block shot, score defender, give free kick |
|                          | (Sport, Soccer)           | 86 / 1,317         | 34(39.5)                      | |
|                          | (Sport, Sumo)             | 75 / 775           | 2(2.7)                        | lead sumo, set championship |
|                          | (Sport, Ski jump)         | 68 / 661           | 10(14.7)                      | postpone downhill, *earn medal* |
|                          | (Sport, Football)         | 37 / 461           | 6(16.2)                       | play football, lease football, stadium |

Munteanu and Marcu (2006) proposed a method for extracting parallel sub-sentential fragments from very non-parallel bilingual corpora. The method is based on the fact that very non-parallel corpora has none or few good sentence pairs, while existing methods for exploiting comparable corpora look for parallel data at the sentence level. Their methodology is the first aimed at detecting sub-sentential correspondences, while they have not reported that the method is also applicable for large amount of data with good performance, especially in the case of large-scale evaluation such as that presented in this paper.

5 Conclusion

We have developed an approach to bilingual verb–noun collocations from non-parallel corpora. The results showed the effectiveness of the method. Future work will include: (i) applying the method to retrieve other types of collocations (Smadja, 1993), and (ii) evaluating the method using Internet directories.

References

Collier, N., H. Hirakawa, and A. Kumano. 1998. Machine Translation vs. Dictionary Term Translation - a Comparison for English-Japanese News Article Alignment. In Proc. of 36th ACL and 17th COLING., pages 263–267.

Fung, P. and P. Cheung. 2004. Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In Proc. of EMNLP2004., pages 57–63.

Gaussier, E., H-M. Renders, I. Matveeva, C. Goutte, and H. Déjean. 2004. A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In Proc. of 42nd ACL, pages 527–534.

Kudo, T. and Y. Matsumoto. 2003. Fast Methods for Kernel-based Text Analysis. In Proc. of 41th ACL, pages 24–31.

Lin, D. 1993. Principle-based Parsing without Overgeneration. In Proc. of 31st ACL, pages 112–120.

Munteanu, D. S. and D. Marcu. 2006. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In Proc. of 21st COLING and 44th ACL, pages 81–88.

Resnik, P. and N. A. Smith. 2003. The Web as a Parallel Corpus. Computational Linguistics., 29(3):349–380.

RWCP. 1998. Rwc Text Database. In Real World Computing Partnership.

Smadja, F. 1993. Retrieving Collocations from Text: Xtract. Computational Linguistics., 19(1):243–178.

Utsuro, T., T. Horiuchi, T. Hamamoto, K. Hino, and T. Nakayama. 2003. Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora. In Proc. of 10th EACL, pages 355–362.

Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer.