# Is (*President*, 大統領) a Correct Sense Pair?
## *Linking and Creating Bilingual Sense Correspondences*

Fumiyo Fukumoto, Yoshimi Suzuki and Attaporn Wangpoonsarp
*Graduate Faculty of Interdisciplinary Research, University of Yamanashi, Japan*

Keywords: Bilingual Sense Correspondence, Domain Specific Senses.

Abstract: This paper presents a method of linking and creating bilingual sense correspondences between English and Japanese noun word dictionaries. Locally, we extracted bilingual noun words using sentence-based similarity. Globally, for each monolingual dictionary, we identified domain-specific senses using a textual corpus with category information. We incorporated these, *i.e.*, we assigned a sense to each noun word of the extracted bilingual words keeping domain (category) consistency. Evaluation on the WordNet 3.0 and EDR Japanese dictionaries using Reuters and Mainichi Japanese newspaper corpora showed 23.1% improvement of bilingual noun word extraction over the baseline with local data view only. Moreover, we found that the extracted bilingual noun senses can be used as a lexical resource for the machine translation.

# 1 INTRODUCTION

There has long been a great deal of interest in retrieval of bilingual lexicons with the availability of a number of large-scale corpora. The extracted bilingual lexicon is widely used for cross-lingual NLP applications, such as machine translation (MT), cross-lingual information retrieval (CLIR), multilingual topic tracking, text classification and summarization. Much of the previous work on finding bilingual lexicons has focused on *word* correspondence, rather than *meaning*. Interpretation of whether the extracted bilingual lexicon is correct relies on human intervention. For example, in most cases, we can accept a pair of English word “president” and Japanese word “大統領” as a bilingual lexicon, while the noun “president” has at least six different senses in WordNet, and “大統領” has two senses in the EDR dictionary.

In this paper, we propose a method for bilingual noun word *sense* correspondence. We focused on manually constructed monolingual dictionaries in different languages because dictionaries such as WordNet (Miller et al., 1990; Miller, 1995), ACQUILEX (Briscoe, 1991), COMLEX (Grishman et al., 1994), and EDR Japanese dictionaries[1] are fine-grained and they are successfully utilized not only for NLP but also for NLP applications. We used them to make up the deficit of corpus statistics obtained from comparable corpora. We identified bilingual noun word senses based on local and global data views. Here, local data view indicates relevance between English and Japanese sentences. We extracted bilingual noun words by applying sentence-based similarity.

Global data view refers to domain/category information. The assumption behind this is that predominant sense of a word can depend on the domain or source of a document (Gale et al., 1992; McCarthy et al., 2007). We identified domain-specific senses using a corpus with category information. This method first identifies each sense of a word in the dictionary to its corresponding category. For each category, we created a graph, where each node refers to each sense of a word, and edges between nodes indicate similarity between two nodes. We applied a Markov Random Walk (MRW) model to the graph and ranked scores for each sense. Finally, we incorporated the local and global information, *i.e.*, we assigned a domain-specific sense to each noun word of the extracted bilingual words maintaining domain/category consistency.

The rest of the paper is organized as follows. The next section describes our approach, especially local and global data views. Section 3 reports some experiments with a discussion of evaluation. Finally, we summarize existing related work and conclude in Section 5.

[1] www2.nict.go.jp/ipp/EDR/ENG/indexTop.html

Fukumoto F., Suzuki Y. and Wangpoonsarp A.
Is (President,åď'ğçţśéăŸ) a Correct Sense Pair? - Linking and Creating Bilingual Sense Correspondences.
DOI: 10.5220/0006489900390048
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD 2017)*, pages 39-48
ISBN: 978-989-758-272-1

# 2 SYSTEM DESIGN

The method consists of three steps: (1) bilingual noun word extraction, (2) identification of domain-specific senses, and (3) bilingual sense correspondence. Hereafter, we describe bilingual noun words as BN words and bilingual noun word senses as BNSs.

## 2.1 Local Data View

The first step is to extract BN words from the corpora. This process consists of two sub-steps: (i) retrieval of relevant documents, and (ii) BN word extraction based on sentence level similarity.

### 2.1.1 Retrieval of Relevant Documents

We used Reuters'96 and the Mainichi Japanese newspaper documents. Let $d_i^J$ $(1 \leq i \leq s)$ and $d_j^E$ $(1 \leq j \leq t)$ be a Mainichi document and a Reuters document, respectively. Each Reuters document $d_j^E$ is translated into a Japanese document $d_j^{E\_mt}$ using English-Japanese MT software. We calculated similarity between two documents using BM25 (Robertson and Walker, 1994), which is widely used in information retrieval studies[2]. BM25 is given by:

$$\mathrm{BM25}(d_i^J, d_j^{E\_mt}) = \sum_{w \in d_j^{E\_mt}} w^{(1)} \frac{(k_1+1)tf_i}{K+tf_i} \frac{(k_3+1)qtf_j}{k_3+qtf_j}, \quad (1)$$

where $w$ is a word within $d_j^{E\_mt}$; $w^{(1)} = \log \frac{(t-n+0.5)}{(n+0.5)}$ is the weight of $w$; $t$ is the number of Mainichi documents; $n$ is the number of documents containing $w$; $K$ refers to $k_1((1-b)+b\frac{dl_i}{avdl})$; $k_1$, $b$, and $k_3$ are parameters set to 1, 1, and 1,000, respectively; $dl_i$ is the document length of $d_i^J$; $avdl$ is the average document length in words; and $tf_i$ and $qtf_j$ are the frequencies of occurrence of the $w$ $(\in d_j^{E\_mt})$ in $d_i^J$, and $d_j^{E\_mt}$, respectively. If the similarity value between $d_i^J$ and $d_j^{E\_mt}$ is higher than the lower bound $L_\theta$, these are regarded as relevant documents, and a document pair is created.

### 2.1.2 Bilingual Word Extraction

We assume that BN word correspondences obtained using relevant documents are unreliable as many noun words appear in a pair of relevant documents. Therefore, we applied sentence level retrieval. First, we applied simple $\chi^2$ statistics to the extracted pairs of relevant documents and extracted Mainichi noun word $w_J$ and Reuters noun word $w_E$ pairs with $\chi^2$ values greater than zero. The $\chi^2$ statistic measures the lack of independence between $w_J$ and $w_E$, and can be compared to the $\chi^2$ distribution with one degree of freedom to judge extremeness (Yang and Pedersen, 1997). It is defined by:

Table 1: Variables of $\chi^2$ statistics.

| | $w_E$ | $\neg w_E$ |
|---|---|---|
| $w_J$ | $f(w_J, w_E) = a$ | $f(w_J, \neg\, w_E) = b$ |
| $\neg w_J$ | $f(\neg\, w_J, w_E) = c$ | $f(\neg\, w_J, \neg\, w_E) = d$ |

$$\chi^2(w_J, w_E) \quad = \quad \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}. \quad (2)$$

$a$, $b$, $c$, and $d$ in Eq. (2) are shown in Table 1. $f(x,y)$ in Table 1 refers to the co-occurrence frequencies of $x$ and $y$ on the Japanese and English sides, respectively. Next, for each $w_J$ and $w_E$ pair, we apply sentence level similarity given by:

$$S\_sim(w_J, w_E) = \max_{S\_w_J \in Set_J, S\_w_E \in Set_E} sim(S\_w_J, S\_w_E),$$

where

$$sim(S\_w_J, S\_w_E) = \frac{\mid S\_w_J \cap S^{mt}\_w_E \mid}{\mid S\_w_J \mid + \mid S^{mt}\_w_E \mid -2\times \mid S\_w_J \cap S^{mt}\_w_E \mid +2}. \quad (3)$$

$Set_J$ and $Set_E$ are sets of sentences that include $w_J$ and $w_E$, respectively; $|X|$ is the number of noun words in a sentence $X$; $\mid S\_w_J \cap S^{mt}\_w_E \mid$ refers to the number of noun words that appear in both $S\_w_J$ and $S^{mt}\_w_E$; and $S^{mt}\_w_E$ is the translation result of $S\_w_E$. The larger value of $sim(S\_w_J, S\_w_E)$ indicates the more similar these two sentences $S\_w_J$ and $S\_w_E$. As shown in Eq. (4), we retrieved $w_J$ and $w_E$ as the BN word such that the similarity between $w_J$ and $w'_E \in BP(w_J)$ are the largest value. Here, $BP(w_J)$ is a set of BN word pairs, each of which includes $w_J$ on the Japanese side.

$$\langle w_J, w_E \rangle \quad = \quad \mathop{\arg\max}_{\langle w_J, w'_E \rangle \in BP(w_J)} S\_sim(w_J, w'_E). \quad (4)$$

## 2.2 Global Data View

The second step is to identify domain-specific senses for each noun word. Globally, we used bilingual category/domain correspondences, Reuters and Mainichi categories. The process is applied independently to the English (WordNet) and Japanese (EDR) dictionaries. We used the MRW model to ranking the senses. For each Reuters category $c_E$ and Mainichi category $c_J$ such as "sports" and "economy" assigned to the Reuters (Mainichi) documents, we created a Graph $G$

[2] http://trec.nist.gov/

= ($S$, $E$) that reflects the relationships between senses in a set $S$. $S$ refers to a set of noun word senses in the Reuters (Mainichi) documents assigned to the category $c_E$ ($c_J$). Each sense $s_i \in S$ is represented by a vector. Each dimension of a vector corresponds to each word appearing in $syn \cup gloss$, where $syn$ indicates a synset and $gloss$ refers to a gloss text in a dictionary. Each element of a dimension is a frequency count of the word in $syn \cup gloss$. $E$ is a set of edges, which is a subset of $S * S$. Each edge $e_{ij} \in E$ is associated with an affinity weight $aw(i \rightarrow j)$. We used two affinity types between $i$ and $j$. Directed relation is a hyponym relation, and undirected relation is a synonym relation. The weight is computed using the standard cosine measure between two senses. The transition probability from $s_i$ to $s_j$ is then defined by normalizing the corresponding affinity weight:

$$p(i \rightarrow j) \;=\; \begin{cases} \dfrac{aw(i \rightarrow j)}{\sum_{k=1}^{|S|} aw(i \rightarrow k)}, \text{ if } \Sigma aw \neq 0 \\ 0 \text{ , otherwise.} \end{cases} \quad (5)$$

We used the row-normalized matrix $U = (U_{ij})_{|S|*|S|}$ to describe $G$ with each sense corresponding to the transition probability, where $U_{ij} = p(i \rightarrow j)$. To make $U$ a stochastic matrix, the rows with all zero elements are replaced by a smoothing vector with all elements set to $\frac{1}{|S|}$. The matrix form of the saliency score $Score(s_i)$ can be formulated in a recursive form as in the MRW model: $\vec{\lambda} = \mu U^T \vec{\lambda} + \frac{(1-\mu)}{|S|}\ \vec{e}$, where $\vec{\lambda}$ = $[Score(s_i)]_{|S|*1}$ is a vector of saliency score for the senses. $\vec{e}$ is a column vector with all elements equal to 1. $\mu$ is a damping factor. We set $\mu$ to 0.85, as in the PageRank (Brin and Pagee, 1998). The final asymmetric transition matrix $M$ is given by:

$$M \;=\; \mu U^T + \frac{(1-\mu)}{\mid S \mid}\vec{e}\vec{e}^T \qquad (6)$$

Each sense score in a specific category is obtained by the principal eigenvector of the new transition matrix $M$. We applied the algorithm for each category.[3] For implementation, we used the Eigen library.[4] We chose a vector with the largest eigenvalues. We normalized a vector, and obtained rank scores of senses.

## 2.3 Bilingual Sense Correspondence

The final step is to retrieve BNSs using both results obtained by local and global data views. Let $\langle w_J, w_E \rangle \in Set_{\{w_J, w_E\}}$ be a pair of nouns obtained by the BN word extraction procedure where $w_J$ has the number of $m$ senses $w_{J\_s_i}$ ($1 \leq i \leq m$) and $w_E$ has the $n$ senses $w_{E\_s_j}$ ($1 \leq j \leq n$), respectively. We retrieved BNS pair $w_{J\_s_i}$ and $w_{E\_s_j}$:

$$\begin{aligned} \langle w_{J\_s_i}, w_{E\_s_j} \rangle_{(c_J, c_E)} \text{ s.t. } & w_{J\_s_i} \in Set_{c_J}, \\ & w_{E\_s_j} \in Set_{c_E}, \\ \langle w_J, w_E \rangle & \text{ satisfies RNN.} \end{aligned} \qquad (7)$$

We recall that the corpora used to identify domain-specific senses are Reuters and Mainichi documents, each of which has different categories. Therefore, we estimated the category correspondence according to the $\chi^2(c_J, c_E)$ statistics shown in Eq. (2). In Eq. (2), we simply replaced $w_J$ by $c_J$ and $w_E$ by $c_E$ as each document of relevant document pairs has category information. The subscript ($c_J$,$c_E$) of $\langle w_{J\_s_i}, w_{E\_s_j} \rangle_{(c_J, c_E)}$ in formula (7) refers to the result of category correspondence between Reuters and Mainichi obtained by Eq. (2). $Set_{c_J}$ and $Set_{c_E}$ show ranked sense lists for $c_J$, and $c_E$, respectively, obtained by the MRW model. RNN is the so-called Reciprocal Nearest Neighbors in that two noun senses are each other's most similar noun (Hindle, 1990).

# 3 EXPERIMENTS

We evaluated each of the three procedures in the experiments, (1) BN extraction by using the results of relevant documents retrieval, (2) retrieving domain specific senses, and (3) BNS correspondences.

## 3.1 Local Data View

In this subsection, we report the results of BN extraction.

### 3.1.1 Experimental Setup

We used Reuters'96 and Mainichi Japanese newspaper corpora from 20 August 1996 to 19 August 1997. The Reuters'96 corpus consists of 806,792 documents organized into three types of coarse-grained categories, *i.e.*, topic, industry, and region. Each consists of 126, 870, and 366 categories. The Mainichi corpus consists of 119,051 documents organized into 16 categories. The difference in dates between Reuters and Mainichi documents is less than ±3 days, *e.g.*, when the date of the Reuters document is 27 August, the corresponding Mainichi data is from 24 to 30 August. We set a small date difference because if some

[3] The principal eigenvector is obtained by the power method and inverse iteration method.

[4] http://eigen.tuxfamily.org/index.php?title=Main_Page

Table 2: Bilingual noun extraction.

| | Pairs | Eng | Jap | # of bilingual nouns | # of nouns (top 1,000) | IRS (top 1,000) |
|---|---|---|---|---|---|---|
| Docs | 196,368 | 133,854 | 20,882 | 172,895 | 162 | 2.32 |
| Docs & Sent | | | | 115,918 | 329 | 4.83 |

event occurred at some specific time period, the press of all countries report it on one of these days. We used English-Japanese MT software (Internet Honyakuno-Ousama for Linux, Ver.5, IBM Corp.). Each Reuters document was translated into a Japanese document, and BM25 was applied.

The training data for choosing the lower bound $L_\theta$ were Reuters 20 August 1996 and Mainichi from 17 to 23 August 1996. The total number of English and Japanese documents collected were 2,586 and 2,137, respectively, and the number of relevant documents collected manually for evaluation was 157. The classification was determined to be correct if the two human judges agreed on the evaluation. We used the F-score for evaluation of relevant document retrieval, which is a measure that balances precision (Prec) and recall (Rec). Let *cSet* be a set of correct document pairs. The definitions of Prec and Rec are given by:

$$\text{Prec} = \frac{\mid \{(d_J, d_E) \mid (d_J, d_E) \in cSet, \text{BM25}(d_J, d_E) \geq L_\theta\} \mid}{\mid \{(d_J, d_E) \mid \text{BM25}(d_J, d_E) \geq L_\theta \mid}$$

$$\text{Rec} = \frac{\mid \{(d_J, d_E) \mid (d_J, d_E) \in cSet, \text{BM25}(d_J, d_E) \geq L_\theta\} \mid}{\mid \{(d_J, d_E) \mid (d_J, d_E) \in cSet \mid}$$

The best performance of F-score was 0.564 when the $L_\theta$ value was 150. We used the value ($L_\theta$ = 150) to extract BN word pairs. We used one year of Reuters and Mainichi documents except for the training data that were used to estimate $L_\theta$ values. The difference in dates between them was less than ±3 days.

#### 3.1.2 Results

The results of BN words extraction are shown in Table 2. We compared the results obtained by our method, "Docs & Sent" with the results obtained by only applying $\chi^2$ statistics to the results of relevant documents, "Docs" to examine how the sentence-based retrieval influences the performance. "Pairs" in Table 2 shows the number of bilingual document pairs. "Eng" and "Jap" show the numbers of Reuters and Mainichi documents within the pairs satisfying the similarity lower bound $L_\theta$ = 150, respectively. "# of nouns" shows the number of correct BN in the topmost 1,000 according to the $\chi^2$ statistics (Docs) and sentence-based similarity (Docs & Sent).

As shown in Table 2, sentence-based retrieval contributed to a reduction in the number of useless BN words without a decrease in accuracy, as about 67% (115,918/172,895) of the size obtained by "Docs" was retrieved, while about 2.03 (329/162) times the number of correct BN words were obtained in the topmost 1,000 nouns. "IRS" (Inverse Rank Score) is a measure of system performance by considering the rank of correct BN words within the candidate nouns; it is the sum of the inverse rank of each matching noun, *e.g.*, correct BN words by manual evaluation matches at ranks 2 and 4 give an IRS of $\frac{1}{2} + \frac{1}{4} = 0.75$. A higher IRS value indicates better system performance. Table 2 shows that sentence-based retrieval also contributes to ranking performance compared with the results obtained by applying $\chi^2$ statistics only.

### 3.2 The Global Data View

Secondly, we report the results of retrieving domain specific senses from the WordNet and EDR dictionaries.

#### 3.2.1 WordNet 3.0

We assigned the Reuters categories to each sense of words in WordNet. We selected 38 Reuters categories, each of which is assigned to more than 80,000 documents. For each category, we collected noun words with frequencies ≥ 5 within a document from the one-year Reuters corpus. There are no existing sense-tagged data for Reuters categories that could be used for evaluation. Therefore, we selected a limited number of words and evaluated these words qualitatively. To do this, we used SFC resources (Magnini and Cavaglia, 2000), which annotate WordNet 2.0 synsets with domain labels. We selected 4 categories that are easy to manually identify corresponding Reuters and SFC categories. Statistics of data and the results are shown in Table 3. "Words" shows the number of different words with frequency ≥ 5 within a document assigned to the category shown in "Reu," and "Sense" shows their total number of senses. "SFC" refers to the number of senses appearing in the SFC resource. "DSS" (Domain-Specific Senses) is the number of senses assigned by our method. "DSS of the # of Correct S" shows the number of correct senses in the topmost 1,000, and "D&S" refers to the number of correct senses appearing in both DSS and SFC.

Table 3: The results of sense assignments (WordNet).

| Reu / SFC | Words | Senses | SFC | DSS | # of Correct S | | IRS |
|---|---|---|---|---|---|---|---|
| | | | | | DSS | D&S | |
| Economics / Economy | 10,284 | 32,550 | 1,032 | 6,355 | 623 | 120 | 2.31 |
| Sports / Sports | 7,437 | 26,478 | 339 | 4,457 | 573 | 158 | 2.01 |
| War / Military | 7,681 | 26,831 | 913 | 4,366 | 510 | 68 | 1.41 |
| Politics / Politics | 10,349 | 32,536 | 793 | 5,981 | 612 | 103 | 2.21 |
| Average | 8,938 | 29,598 | 769 | 5,290 | 580 | 112 | 1.99 |

Table 4: DSS against the FS heuristic (WordNet).

| Cat | Sense | DSS | FS |
|---|---|---|---|
| Economics | 5.6 | 0.73 | 0.68 |
| Sports | 4.5 | 0.71 | 0.69 |
| War | 4.9 | 0.52 | 0.30 |
| Politics | 4.2 | 0.75 | 0.68 |
| Average | 4.8 | 0.68 | 0.59 |

As shown in Table 3, the numbers of correct senses obtained by "DSS" and "D&S" did not exactly match. This is not surprising because our method used the Reuters corpus, while the SFC resource consists of 96% of the WordNet synsets, each of which is manually annotated using 115 different SFC. The IRS depends on the category, and the average IRS was 1.99. It is interesting to note that some senses of words that were obtained correctly by our approach did not appear in the SFC resource. For example, the words "shot" and "strike" in the Sport category, and "liberty" and "military_hospital" in the war/military category were obtained by our approach but did not appear in the SFC. This is because we used WordNet 3.0, while SFC was based on WordNet 2.0. These observations clearly support the usefulness of our automated method.

In the WSD task, a first sense (FS) heuristic is often applied because of its power and lack of a requirement for expensive hand-annotated data sets (Cotton et al., 1998; McCarthy et al., 2007). We compared the results obtained by DSS to those obtained by the FS heuristic. For each category, we randomly selected 10 words from the senses assigned by DSS, and selected 10 sentences from the documents belonging to each corresponding category. As a result, we tested 100 sentences for each category. Table 4 shows the results. "Sense" refers to the number of average senses per word. The average precision (Avg) of our method was 0.68, while the result obtained by the FS was 0.59.

### 3.2.2 EDR Dictionary

We assigned categories from Mainichi Japanese documents to each sense of words in the EDR Japanese word dictionary[5]. We selected 11 out of 16 categories, each of which had a sufficient number of documents. All documents were tagged by the morphological analyzer Chasen (Matsumoto et al., 2000), and noun words were extracted. We choose three categories for which it is easy to manually create correct data. Table 5 shows the statistics of the data and the results of assignment. "Cor" shows the number of senses assigned by our approach correctly in the topmost 1,000 senses. The average IRS obtained by the EDR was 4.93 and Reuters was 1.99. This is reasonable as the assignment task using EDR is easy compared to WordNet, *i.e.*, the average number of senses per word of the former is 1.93, while that of the latter is 3.31.

Table 5: The results of sense assignments (EDR).

| Cat | Words | Senses | Cor | IRS |
|---|---|---|---|---|
| Economics | 15,906 | 30,869 | 740 | 6.39 |
| Sports | 17,556 | 33,595 | 559 | 2.77 |
| International | 13,906 | 27,239 | 451 | 5.63 |
| Average | 15,789 | 30,568 | 583 | 4.93 |

Table 6: DSS against the FS heuristic (EDR).

| Cat | Sense | DSS | FS |
|---|---|---|---|
| Economics | 2.793 | 0.79 | 0.60 |
| Sports | 2.873 | 0.65 | 0.52 |
| International | 2.873 | 0.68 | 0.58 |
| Average | 2.846 | 0.70 | 0.57 |

In the WSD task, we randomly selected 30 words from the senses assigned by DSS. For each word, we selected 10 sentences from the documents belonging to each corresponding category. The FS in the EDR is determined based on the EDR corpus. Table 6 shows the results. As can be seen in Table 6, DSS was also better than the FS heuristics in Japanese data. The overall performance for the FS (0.57) was not better, similar to the case for the English data (0.59), while the number of senses per word in the Japanese resource was smaller than that in WordNet. There were many senses that did not occur in the EDR corpus, *i.e.*,

[5] www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html?

Table 7: The results of bilingual sense correspondences.

| Approach | Candidates | | # of Correct ext.(1,000) | | RNN (1,071) | | IRS | |
|---|---|---|---|---|---|---|---|---|
| | BN | BNS | BN | BNS | BN | BNS | BN | BNS |
| Local | 171,895 | — | 329 | — | 431 | — | 2.32 | — |
| Local & Global | 171,895 | 115,918 | 437 | 312 | 679 | 580 | 4.10 | 4.35 |

Table 8: Examples of bilingual noun senses.

| Cat pair<br>(Mai, Reu) | # of Pairs | | # of Correct<br>senses(%) | Examples<br>Sense_id (gloss text) | |
|---|---|---|---|---|---|
| | Cand | RNN | | | |
| (International,<br>Government) | 269 | 10 | 9 (90.0) | EDR:<br>WordNet: | ゲリラ_01 (an irregular group of soldiers given to sneak attacks)<br>guerrilla_01 (a member of an irregular armed force) |
| (Economics,<br>Economics) | 4,964 | 58 | 44 (75.9) | EDR:<br>WordNet: | 株_09 (stock, share)<br>stock_01 (the capital raised by a corporation through the issue of shares) |
| (Sports,<br>Sports) | 6,560 | 68 | 48 (70.6) | EDR:<br>WordNet: | パー_02 (in golf, the average number of strokes for playing around a course)<br>par_01 (the number of strokes set for each hole on a golf course) |
| (Local news,<br>Crime) | 2,903 | 34 | 23 (67.6) | EDR:<br>WordNet: | 殺人_02 (the act of killing someone)<br>killer_01 (someone who causes the death of a person) |

62,460 nouns appeared in both EDR and Mainichi newspapers (from 1991 to 2000), 164,761 senses in all. Of these, 114,267 senses did not appear in the EDR corpus. This also demonstrates that automatic DSS works well compared to the frequency-based FS heuristics.

## 3.3 Bilingual Sense Correspondence

Finally, we evaluated the performance of BNS correspondences.

### 3.3.1 Experimental Setup

The data for BNS correspondence were the Reuters and Mainichi corpora from the same period, *i.e.*, 20 August 1996 to 19 August 1997. The total numbers of documents were 806,791, and 119,822, respectively. Locally, we extracted BN words using sentence-based similarity. We retrieved cross-lingually relevant Japanese documents with English documents. The difference in dates between them was less than ±3 days. We used these documents to extract BN words. Globally, we assigned domain-specific senses for each of the 38 categories for WordNet 3.0 and 11 categories for EDR. We estimated category correspondences, and retrieved BNSs according to their correspondences. We obtained 92 category correspondences in all with $\chi^2$ values larger than zero. From these data, we extracted BNSs.

### 3.3.2 Results

Table 7 shows the results of BNS correspondences. “Local” indicates the results using only sentence-based similarity, and “Local & Global” shows the results obtained by our method. “# of Correct ext.” refers to the number of correct extractions within the topmost 1,000, and “RNN” shows the number of correct extractions by applying RNN. “BN” refers to the number of BN words, and “BNS” is their senses.

As can be clearly seen in Table 7, the results with integration of local and global data views improved overall performance of BN word extraction compared to local data only because 10.8% (437−329)/1,000 improvement within the topmost 1,000, and 23.1% (679−431)/1,071 improvement with RNN. We obtained 312 BNSs within the topmost 1,000, although bilingual sense correspondence is a difficult task. Moreover, RNN is effective for BNS correspondences. We obtained a total of 1,071 BNSs, which 580 were correct, while that without RNN was 312 within the topmost 1,000.

Table 8 lists examples of BNSs for each category correspondence. ($x$, $y$) of the category pair refers to the Mainichi and Reuters category correspondence. “Cand” denotes the number of extracted word pairs, and “RNN” shows word pairs obtained by RNN. “Sense_id” shows the sense and its order of appearance in each dictionary, WordNet and EDR. Table 8

Table 9: Constituents of bilingual sense of words.

| | |
|---|---|
| Eng | authority, board, budget, business, case coast, company deficit, finance, group, issue money, opposition people, power, president seat, space, state, trial |
| Jap | グループ, ビジネス, 宇宙, 沖, 会長, 海上 株, 幹部, 機関, 議席, 金, 国民, 最終, 財政 資金, 事業, 事件, 社長, 主義, 人, 政治, 政府 赤字, 大統領, 団体, 当局, 場合, 法廷, 本部 問題, 予算, 力 |

Table 10: Examples of bilingual senses.

| Cat pair<br>(Mai, Reu) | Sense pair<br>(Mai, Reu) |
|---|---|
| (top news, industrial) | (権威_01, authority_01) |
| (international, international) | (当局_02, authority_05) |
| (top news, election) | (場合_02, case_01) |
| (top news, violence) | (事件_01, case_02) |
| (top news, funding) | (赤字_01, deficit_01) |
| (top news, politics) | (不足_02, deficit_02) |

shows that the first three senses of words, “guerrilla,” “stock,” and “par” corresponded correctly. However, “the act of killing someone” in the EDR was incorrectly identified by “killer” (person) in WordNet. Our approach for identifying BNSs is based on term-based corpus statistics. It will be necessary to investigate other types of lexicon, such as verb and subjective/objective noun collocations, for further improvement. Our method for category correspondence is very simple as it is based on the $\chi^2$ statistics. Much of the previous work on text categorization indicated that hierarchical structure (*e.g., MeSH, Yahoo!, LookSmart*) improves the accuracy of text categorization, and it is worth attempting to use a hierarchical structure to determine corresponding categories with high accuracy.

Finally, we compared the results of bilingual sense correspondences with a machine readable bilingual dictionary and two English-Japanese MT systems. We randomly choose 80 BNSs from 580 correct bilingual senses obtained by RNN. There were 40 different category pairs, and as shown in Table 9, the number of different English and Japanese words were 20, and 32, respectively. Table 10 shows examples of BNSs and its corresponding category pairs.

We choose titles of documents as a translation test data. Because they are short sentences, and hard to be affected by a syntactic analyzer in MT systems which enable to approximate a fair comparison. For each of the 80 BNS pairs, we created translation test data.

For example, in the word, “authority” of BNS pair (権威_01, authority_01) in Table 10, we randomly selected 10 titles sentences including “authority” from the Reuters documents assigned to the category “industrial”. We translated these test data by using two MT systems[6], and compared the translation result of “authority”. We also compared the results obtained by our method with English-Japanese dictionary, Eijirou on the WEB[7]. In the English-Japanese dictionary, we used the first-sense heuristic (choosing the first sense of a word). The comparative results are shown in Table 11.

Table 11: Comparison against a dictionary and MT systems.

| | # of Correct senses (%) |
|---|---|
| Eijirou | 329 / 800 (41.2) |
| Honyakuno-Ousama | 521 / 800 (65.1) |
| SYSTRANet | 550 / 800 (68.8) |
| RNN | 562 / 800 (70.3) |

As we can see from Table 11 that the results obtained by RNN was 70.3%, and it was better to the results obtained by bilingual dictionary (41.2%), and slightly better to the results obtained by SYSTRANet (68.8%). Moreover, it works well compared to the Honyakuno-Ousama (MT) that we used in the process of BN word extraction as local data view. It can be observed from these results that the extracted BNSs can be used as a lexical resource for MT system.

# 4 RELATED WORK

Our approach to link bilingual word senses in dictionaries can be regarded as a type of ontology alignment. The earliest such attempts were Chimaera (McGuinness et al., 2000) and PROMPT (Noy and Musen, 2000). Suchanek *et al.* developed an ontology alignment system called PARIS which relies on instance overlap-based cues to align instances, categories, and relations from two knowledge bases. They reported that the method is effective, although it remains insufficient to align ontologies that share few or no data entries in common. Wijaya *et al.* focused on the problem, and presented a method of aligning ontologies called PIDGIN that employs a very large natural language text as interlingua and graph-based self-supervised learning (Wijaya et al., 2013). The use of corpora is similar to our method, although the target of integration is quite different, *i.e.*, PIDGIN aimed at relation and category alignment, while our method aligned noun word senses.

In the context of bilingual lexicon extraction, much of the previous work used comparable corpora. One attempt involved directly retrieving bilin-

[6] We used Internet Honyakuno-Ousama, and SYSTRANet English-Japanese MT software. www.systranet.com/translate

[7] www.alc.co.jp

gual lexicons from corpora (Fung and Cheung, 2004; Gaussier et al., 2004). The alternative approach consists of two steps: first, cross-lingual relevant documents are retrieved from comparable corpora, and then bilingual term correspondences within these relevant documents are estimated. Much of the previous work in finding relevant documents used MT systems or existing bilingual lexicons to translate one language into another (Utsuro et al., 2003). Document pairs are then retrieved based on document similarity. Another approach to retrieving relevant documents involves the collection of relevant document URLs from the WWW (Resnik and Smith, 2003). Munteanu *et al.* proposed a method for extracting parallel sub-sentential fragments from very non-parallel bilingual corpora (Munteanu and Marcu, 2006). All of these methods successfully extracted bilingual lexicons but they ignored the meanings of words. One attempt to deal with the meaning of words is Kusner *et al.*'s method (Kusner et al., 2015). They presented the Word Mover's Distance (WMD) between text documents by utilizing word2vec embeddings (Mikolov et al., 2013a; Mikolov et al., 2013b). Word2vec learns a vector representation for each word using a neural network architecture consisting of three layers, *i.e.* an input layer, a projecting layer, and an output layer to predict nearby words.

There also has been a lot of work where bilingual word vectors are induced using parallel corpora (Brown et al., 1993; Haghighi et al., 2008; Dyer et al., 2013; Kocisky et al., 2014). Dyer *et al.* presented an alignment model called FASTALIGN model which uses an alignment distribution defined by a single parameter that measures how close the alignment is to the diagonal (Dyer et al., 2013). Blunsem *et al.* extended Dyer *et al.*'s model for learning bilingual word representations. They marginalize out the alignments which enable to capture more bilingual semantic context (Kocisky et al., 2014). Gouws *et al.* proposed a simple and computationally efficient model called BioBOWA (Bilingual Bag-of-Words without Alignments) for learning bilingual distributed representations of words which can scale to large monolingual datasets, and does not require word-aligned parallel training data (Gouws et al., 2015). The method requires monolingual data which trains directly, and extracts a bilingual signal from a smaller set of raw-text sentence-aligned data. They evaluated the induced cross-lingual embeddings on the two tasks, *i.e.* document classification and lexical translation task, and the results showed that the method outperforms current state-of-the-art methods, especially it contributes to reduce computational cost.

In the context of domain-specific senses of a word, Magnini *et al.* presented a lexical resource where WordNet 2.0 synsets were annotated with Subject Field Codes (SFC) by a procedure that exploits the WordNet structure (Magnini and Cavaglia, 2000; Bentivogli et al., 2004). 96% of the WordNet synsets of the noun hierarchy could have been annotated using 115 different SFC, while identification of the domain labels for senses required a considerable amount of hand-labeling. McCarthy *et al.* presented a method to find domain-specific predominant noun senses automatically using a thesaurus acquired from raw textual corpora and the WordNet similarity package (McCarthy et al., 2007). They tested two domains, "Sports" and "Finance."
indent To our knowledge, there have been only a few previous work on bilingual sense extraction (Plous and Ji, 2003). One approach is word translation disambiguation presented by Li *et al.* (Li and Li, 2004). Their method is based on one-to-many sense mapping. They used a machine learning technique that repeatedly constructs classifiers in the two languages in parallel. They reported that the approach significantly outperformed existing methods using two nouns (Pedersen, 2000), and seven of the twelve English words studied in WSD research by (Yarowsky, 1995). Their method requires a small number of sense-tagged training data in both of the two languages, while our method requires documents assigned to categories and a dictionary with gloss text which unfortunately hinders a direct and fair comparison between their system and ours.

Our method has three novel aspects. First, we propose a method to integrate different data views to improve the quality of bilingual sense correspondence. Second, from the perspective of existing knowledge-based integration, we propose a method for corresponding senses between two monolingual dictionaries via many-to-many sense mapping. Finally, from the perspective of robustness, the method is automated, and requires only documents assigned to domains/categories and a dictionary with gloss text. It can be applied easily to a new domain, sense inventory, or different languages given sufficient documents.

# 5 CONCLUSION

We have developed an approach for linking and creating bilingual sense correspondences by combining local and global data views. Future work will include: (i) investigating other types of lexicon for further improvement, (ii) extending the method to use hierarchical structures of categories for category correspondence, (iii) retrieving other parts of speech word senses, and (iv) evaluation of the method using dictionaries other than English WordNet and Japanese EDR.

## ACKNOWLEDGEMENTS

The authors would like to thank Bernardo Magnini for providing SFC resources, and the anonymous reviewers for their comments and suggestions. This work was supported by the Telecommunications Advancement Foundation, and Support Center for Advanced Telecommunications Technology Research, Foundation.

## REFERENCES

Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising the wordnet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 101–108.

Brin, S. and Pagee, L. (1998). Lexical Issues in Natural Language Processing. In *Proc. of the 7th International Conference on World Wide Web 7*, pages 107–117.

Briscoe, E. J. (1991). Lexical Issues in Natural Language Processing. In *Natural Language and Speech. Proceedings of the Symposium on Natural Language and Speech*, pages 39–68.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. 19(2):263–311.

Cotton, S., Edmonds, P., Kilgarriff, A., and Palmer, M. (1998). SENSEVAL-2, http://www.sle.sharp.co.uk/senseval2/.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proc. of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Fung, P. and Cheung, P. (2004). Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proc. of 2004 Conference on Empirical Methods in Natural Language Processing*, pages 57–63.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One Sense Per Discourse. In *Proc. of Speech and Natural Language Workshop 1992*, pages 233–237.

Gaussier, E., Renders, H.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004). A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 527–534.

Gouws, S., Bengio, Y., and Corrado, G. (2015). BilBOWA:Fast Bilingual Distributed Representations without Word Alignments. In *Proc. of the 32nd International Conference on Machine Learning*, pages 748–756.

Grishman, R., MacLeod, C., and Meyers, A. (1994). COMLEX Syntax: Building a Computational Lexicon. In *Proc. of the 15th International Conference on Computational Linguistics*, pages 268–272.

Haghighi, A., Liang, P., Kirkpatrick, T. B., and Klein, D. (2008). Learning Bilingual Lexicons from Monolingual Corpora. In *Proc. of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 771–779.

Hindle, D. (1990). Noun Classification from Predicate-Argument Structures. In *Proc. of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275.

Kocisky, T., Hermann, K. M., and Blunsom, P. (2014). Learning Bilingual Word Representations by Marginalizing Alignments. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 224–229.

Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. (2015). From Word Embeddings to Document Distances. In *Proc. of the 32nd International Conference on Machine Learning*, pages 957–966.

Li, H. and Li, C. (2004). Word Translation Disambiguation Using Bilingual Bootstrapping. In *Computational Linguistics*, pages 1–22.

Magnini, B. and Cavaglia, G. (2000). Integrating Subject Field Codes into WordNet. In *Proc. of LREC-2000*.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, Y., Takaoka, K., and Asahara, M. (2000). Japanese Morphological Analysis System ChaSen Version 2.2.1. In *NAIST Technical Report NAIST*.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2007). Unsupervised Acquisition of Predominant Word Senses. In *Computational Linguistics*, volume 33(4), pages 553–590.

McGuinness, D. L., Fikes, R., Rice, J., and Wilder, S. (2000). An Environment for Merging and Testing Large Ontologies. In *Proc. of the Conference on Principles of Knowledge Representation and Reasoning*, pages 483–493.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and Their Compositionality. In *Proc. of NIPS*, pages 3111–3119.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41.

Miller, G. A., Beckwith, R. T., Fellbaum, C. D., Gross, D., and Miller, K. J. (1990). WordNet: An on-line Lexical Database. In *International Journal of Lexicography*, volume 3(4), pages 235–244.

Munteanu, D. S. and Marcu, D. (2006). Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora.

In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88.

Noy, N. F. and Musen, M. A. (2000). Prompt: Algorithm and Tool for Automated Ontology Merging and Alignment. In *Proc. of the 17th National Conference on Artificial Intelligence*, pages 450–455.

Pedersen, T. (2000). A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. In *Proc. of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 197–207.

Plous, S. and Ji, H. (2003). A Model for Matching Semantic Maps between Language (French/English, English/French). In *Computational Linguistics*, pages 155–178.

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

Robertson, S. E. and Walker, S. (1994). Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241.

Utsuro, T., Horiuchi, T., Hamamoto, T., Hino, K., and Nakayama, T. (2003). Effect of Cross-Language IR in Bilingual Lexicon Acquisition from Comparable Corpora. In *Proc. of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 355–362.

Wijaya, D., Talukdar, P. P., and Mitchell, T. (2013). PIDGIN: Ontology Alignment using Web Text as Interlingua. In *Proc. of ACM Conference on Information and Knowledge Management*, pages 589–598.

Yang, Y. and Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proc. of the 14th International Conference on Machine Learning*, pages 412–420.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.