

Extracting Translation Pairs from Comparable Corpora through Graph-based Label Propagation

Akihiro Tamura^{†,††}, Taro Watanabe[†], Eiichiro Sumita[†],
Hiroya Takamura^{††} and Manabu Okumura^{††}

This paper proposes a novel method for bilingual lexicon extraction from comparable corpora using graph-based label propagation. A previous study found that performance drastically decreases when the coverage of a seed lexicon is small. We address this problem by using indirect relations with bilingual seeds together with direct relations, in which each word is represented by a distribution of lexical seeds. The seed distributions are propagated over a graph that represents relations among words. Translation pairs are extracted by identifying word pairs with high similarities in the seed distributions. We propose two types of graphs: (1) a co-occurrence graph, representing co-occurrence relations between words; and (2) a similarity graph, representing context similarities between words. Evaluations on comparable corpora of English and Japanese patent documents show that our proposed graph propagation method outperforms conventional methods. Further, the similarity graph improved performance by clustering synonyms into the same translation.

Key Words: *Bilingual lexicon, Comparable corpora, Label propagation*

1 Introduction

Bilingual lexicons are important resources for bilingual tasks, such as machine translation (MT) and cross-language information retrieval (CLIR). Therefore, the automatic construction of bilingual lexicons from corpora is an important issue that has attracted many researchers. As a solution, a previous study proposed the extraction of bilingual lexicons from comparable corpora in which documents were not direct translations but shared a topic or domain¹. The use of comparable corpora is motivated by the availability of large parallel corpora only for a few language pairs and limited domains.

Most previous methods are based on assumption (I), namely that a word and its translation tend to appear in similar contexts across languages (Rapp 1999). On the basis of this assumption, many methods calculate word similarity using context and then extract word translation pairs that have high context similarities. These methods are called context-similarity-based methods.

[†] Multilingual Translation Laboratory, National Institute of Information and Communications Technology

^{††} Precision and Intelligence Laboratory, Tokyo Institute of Technology

¹ Although Vulić et al. (2011) regarded document-aligned texts such as texts on Wikipedia as comparable corpora, we do not limit comparable corpora to these types of texts.

Context similarities are usually computed using a seed bilingual lexicon (e.g., a general bilingual dictionary) by projecting contexts expressed in two different languages into the same space. In the projection, information not represented by the seed lexicon is discarded. Therefore, context-similarity-based methods cannot find accurate translation pairs if a small seed lexicon is used.

Some previous methods attempted to alleviate this problem of limited seed lexicon size (Koehn and Knight 2002; Morin and Prochasson 2011; Hazem, Morin, and Saldarriaga 2011), while other approaches did not require any seed lexicon (Rapp 1995; Fung 1995; Haghighi, Liang, Berg-Kirkpatrick, and Klein 2008; Ismail and Manandhar 2010; Daumé III and Jagarlamudi 2011). However, these approaches suffer from high computational cost (Rapp 1995), sensitivity to parameters (Hazem et al. 2011), low accuracy (Fung 1995; Ismail and Manandhar 2010), and ineffectiveness for language pairs using different characters (Koehn and Knight 2002; Haghighi et al. 2008; Daumé III and Jagarlamudi 2011).

Given the above problems, we propose a novel method that uses a graph-based label propagation (LP) technique (Zhu and Ghahramani 2002). Our proposed method is based on assumption (II), which is derived by recursively applying assumption (I) to the “contexts”; assumption (II) states that a word and its translation tend to have similar co-occurrence (direct and indirect) relations with all bilingual seeds across languages.

On the basis of assumption (II), we propose the following three-step approach: (1) construct a graph for each language with each edge indicating a direct co-occurrence relation; (2) represent every word as a seed translation distribution by iteratively propagating lexical seeds in each graph; and (3) find two words in different languages with a high similarity with respect to seed distributions. By propagating all seeds on the graph, indirect co-occurrence relations, which have been neglected in previous methods, are also considered when computing bilingual relations. In addition to the co-occurrence-based graph construction, we propose a similarity graph, which also considers context similarities between words.

The main contributions of this paper are as follows:

- We propose a bilingual lexicon extraction method that captures co-occurrence relations with all seeds, including indirect relations, by using graph-based LP. In our experiments, we confirm that the proposed method outperforms conventional context-similarity-based methods (Rapp 1999; Andrade, Nasukawa, and Tsujii 2010) and works well even if the coverage of a seed lexicon is small.
- We propose a similarity graph that represents context similarities between words. In our experiments, we confirm that the similarity graph is more effective than the co-occurrence graph.

2 Context-Similarity-Based Extraction

Extracting bilingual lexicon from comparable corpora was pioneered in Rapp (1995) and Fung (1995). The popular similarity-based methods consist of the following steps: (1) model the contexts; (2) calculate the context similarities; and (3) find the translation pairs.

Step 1. Modeling contexts: The context of each word is generally modeled by a vector in which each dimension corresponds to a context word and has a value indicating occurrence correlation. Various definitions for context have been used, such as distance-based context (e.g., in a sentence (Laroche and Langlais 2010), paragraph (Fung and McKeown 1997), or predefined window (Rapp 1999; Andrade et al. 2010)) and syntax-based context (e.g., predecessors and successors in dependency trees (Garera, Callison-Burch, and Yarowsky 2009) or a certain dependency position (Otero and Campos 2008)). Some treated context words equally regardless of their positions (Fung and Yee 1998), while others treated words separately for each position (Rapp 1999). Similarly, various correlation measures have been used, such as the log likelihood ratio (Rapp 1999; Chiao and Zweigenbaum 2002), tf-idf (Fung and Yee 1998), pointwise mutual information (PMI) (Andrade et al. 2010), and context heterogeneity (Fung 1995).

Other approaches abound. Shao and Ng (2004) represented contexts using language models. Andrade et al. (2010) used a set of words with a positive association as a context. Andrade et al. (2011a) used dependency relations instead of context words. Ismail and Manandhar (2010) used only in-domain words in contexts. Pekar et al. (2006) constructed smoothed context vectors for rare words. Laws et al. (2010) used graphs in which vertices correspond to words and edges indicate three types of syntactic relations such as adjectival modification.

Step 2. Calculating context similarities: Contexts expressed in two languages are projected into the same space. In general, previous methods used a seed bilingual lexicon for this projection. After that, similarities are calculated on the basis of the projected context vectors using various measures, such as the city-block metric (Rapp 1999), the cosine similarity (Fung and Yee 1998), the weighted Jaccard index (Hazem et al. 2011), the Jensen-Shannon divergence (Pekar et al. 2006), the number of overlapping context words (Andrade et al. 2010), SimRank (Laws et al. 2010), and the Euclidean distance (Fung 1995).

Andrade et al. (2011b) performed a linear transformation of context vectors in accordance with the notion that importance varies by context positions. Gaussier et al. (2004) projected context vectors via latent classes to capture synonymy and polysemy in a seed lexicon. Fišer et al. (2011) and Kaji (2005) calculated two-way similarities.

Step 3. Finding translation pairs: A pair of words is treated as a translation pair when their

query word. The tables above indicate the correlation strength between the occurrences of a query word and its context words. From the tables shown, a context vector for the Japanese word “ピラニア (piranha)” has the dimensions “ $F_{アマゾン} = 0.8$, $F_{ジャングル} = 0.6$, $F_{淡水} = 0.5$,” where “ $F_x = a$ ” indicates that the value of the dimension corresponding to the word x is a . Similarly, context vectors for the English words “piranha” and “anaconda” have the dimensions “ $F_{Amazon} = 0.8$, $F_{jungle} = 0.6$, $F_{freshwater} = 0.5$ ” and “ $F_{Amazon} = 0.8$, $F_{jungle} = 0.6$,” respectively. These context vectors are projected into the same space by using the seed lexicon for calculating similarities between the context vectors. After the context vector for “ピラニア (piranha)” is projected, the information on co-occurrence with “淡水 (freshwater)” disappears, because the seed lexicon does not include “淡水 (freshwater).” The same problem occurs with “piranha.” Hence, after the projection, all context vectors for “ピラニア (piranha),” “piranha,” and “anaconda” become the same with the dimensions “ $F_{アマゾン-Amazon} = 0.8$, $F_{ジャングル-jungle} = 0.6$.” Therefore, both the similarity of “ピラニア (piranha)” and “piranha” and that of “ピラニア (piranha)” and “anaconda” are 1, and the pair “ピラニア (piranha)” and “anaconda” could be wrongly identified as a translation pair.

An additional previous study focused on the problem of limited seed lexicon size. Morin and Prochasson (2011) complemented the seed lexicon with a bilingual lexicon extracted from parallel sentences. Koehn and Knight (2002) used identically spelled words in two languages as the seed lexicon. However, this method would not be applicable to language pairs using different characters, such as English and Japanese. Hazem et al. (2011) exploited k -nearest words for a query word, which is very sensitive to the parameter k .

Some previous studies did not require any seed lexicon. Rapp (1995) proposed a computationally demanding matrix permutation method that maximizes similarity between co-occurrence matrices in two languages. Ismail and Manandhar (2010) introduced a similarity measure between two words in different languages without requiring any seed lexicon. Fung (1995) used context heterogeneity vectors in which each dimension is independent of the language types. However, their performances are worse than those of conventional methods that use a small seed lexicon. Haghighi et al. (2008) and Daumé III and Jagarlamudi (2011) proposed a generative model based on probabilistic canonical correlation analysis in which words are represented by context and orthographic features². However, their experiments showed that orthographic features are important for their effectiveness, which means they provide low performance for language pairs

² In Haghighi et al. (2008) and Daumé III and Jagarlamudi (2011), indirect relations with seeds are considered topologically, but our method utilizes degrees of indirect correlations with seeds.

using different characters.

3 Label Propagation Based Extraction

As described in Section 2, the performance of the previous study suffers significant degradation when used with a small seed lexicon. This problem can be resolved by incorporating indirect relations with all seeds when identifying translation pairs. For example, in Figure 1, “ピラニア (piranha)” has a certain degree of association with the seed “魚—fish” through “淡水 (freshwater)” in both Japanese and English, although “ピラニア (piranha)” and “魚 (fish)” do not co-occur in the same contexts. Moreover, “anaconda” has very little association with the seed “魚—fish” in English. Therefore, the indirect relation with the seed “魚—fish” enables to discriminate between “piranha” and “anaconda” and could be an important clue for identifying a correct translation pair.

To utilize indirect relations, we introduce assumption (II), namely that a word and its translation tend to have similar co-occurrence (direct and indirect) relations with all bilingual seeds across languages³. On the basis of assumption (II), we propose to identify a word pair as a translation pair when its co-occurrence (direct and indirect) relations with all seeds are similar.

To obtain co-occurrence relations with all seeds, including indirect relations, we focus on a graph-based LP technique (Zhu and Ghahramani 2002). In general, LP transfers labels from labeled to unlabeled data points. Through this process, all vertices have soft labels that can be interpreted as label distributions. We apply LP to bilingual lexicon extraction by treating each lexical seed as a label and representing each word as a vertex in a graph with each edge encoding a direct co-occurrence relation. Lexical seeds are propagated as labels and seed distributions are obtained for each word. From the seed distributions, we identify translation pairs.

In summary, our proposed method consists of the following three steps, as presented in Figure 2: (1) graph construction for each language; (2) seed propagation in each graph; and (3) translation pair extraction. More formally, our algorithm extracts translation pairs T from two monolingual corpora D^e in language e and D^f in language f using a seed lexicon S . Step 1-1 shows graph construction on language e , where a graph G^e with a set of edges E^e , vertices V^e , and edge weights W^e is constructed from D^e . In Step 1-2, another graph G^f is constructed for language f . Next, seed distributions Q^e for all vertices V^e of G^e are computed by propagating

³ Assumption (I) indicates that direct co-occurrence relations between a word and its context words are preserved across different languages. Therefore, assumption (II) is derived by recursively applying assumption (I) to the “context words.”

Algorithm Bilingual Lexicon Extraction

Require: comparable corpora D^e and D^f ,
a seed lexicon S consists of S^e and S^f

Ensure: Output translation pairs T

1-1: $G^e = \{E^e, V^e, W^e\} \leftarrow \text{construct-graph}(D^e)$

1-2: $G^f = \{E^f, V^f, W^f\} \leftarrow \text{construct-graph}(D^f)$

2-1: $\tilde{G}^e = \{E^e, V^e, W^e, Q^e\} \leftarrow \text{propagate-seed}(G^e, S^e)$

2-2: $\tilde{G}^f = \{E^f, V^f, W^f, Q^f\} \leftarrow \text{propagate-seed}(G^f, S^f)$

3: $T \leftarrow \text{extract-translation}(Q^e, Q^f, S)$

Fig. 2 Algorithm of Proposed Method

language e words in the seed lexicon $S^e = \{v^e | \langle v^f, v^e \rangle \in S\}$ on the graph G^e in Step 2-1. Similarly, Step 2-2 propagates language f words in the seed lexicon $S^f = \{v^f | \langle v^f, v^e \rangle \in S\}$ on the graph G^f . Finally, Step 3 extracts translation pairs T based on Q^e and Q^f paired by S .

3.1 Graph Construction

Here, we construct a graph representing the association between words for each language. Each graph is undirected because such associations do not require direction. Graphs are constructed via the following steps.

Step 1. Vertex assignment extracts words from each corpus and assigns a vertex to each extracted word. Let $V = \{v_1, \dots, v_n\}$ be a set of vertices.

Step 2. Edge weight calculation calculates the association strength between two words as the edge weight. Let E and W be a set of edges and weights, respectively, where $e_{ij} \in E$ links v_i and v_j and $w_{ij} \in W$ is the weight of e_{ij} . Note that $|E| = |W|$.

Step 3. Edge pruning excludes edges whose weights are below a predefined threshold to reduce the computational cost of seed propagation.

We propose two types of graphs that differ in the association measure used in Step 2, namely a co-occurrence graph and a similarity graph⁴.

3.1.1 Co-occurrence Graph

A co-occurrence graph directly encodes assumption (II). Each edge in the graph indicates the correlation strength between occurrences of the two linked words. An example is shown in Figure 1.

⁴ We can combine the association measures used in co-occurrence and similarity graphs. We leave this combination approach for future work.

In the **edge weight calculation** step, co-occurrence frequencies are first computed for each word pair in the same context. Next, the correlation strength is estimated. There are various definitions of context and correlation measures that can be used (e.g., the approaches used for modeling contexts in context-similarity-based methods). In this paper, we use words in a predefined window (window size of 10) as the context and PMI as the correlation measure. More specifically,

$$w_{ij} = PMI(v_i, v_j) = \log \frac{p(v_i, v_j)}{p(v_i) \cdot p(v_j)}, \quad (1)$$

where $p(v_i)$ (or $p(v_j)$) is the probability that v_i (or v_j) occurs in a context, and $p(v_i, v_j)$ is the probability that v_i and v_j co-occur within the same context. We estimate $PMI(v_i, v_j)$ using the Bayesian method proposed by Andrade et al. (2010). Finally, edges with a negative association, $PMI(v_i, v_j) \leq 0$, are pruned in the **edge pruning** step.

3.1.2 Similarity Graph

Co-occurrence graphs are very sensitive to accidental relations caused by infrequent co-occurrences. Thus, we propose a similarity graph in which context similarities are employed as edge weights instead of simple co-occurrence-based correlations. Since context similarities are computed by the global correlation among words that co-occur, a similarity graph is less subject to accidental co-occurrences. The use of a similarity graph is inspired by assumption (III), namely that a word and its translation tend to have similar context similarities with all bilingual seeds across languages⁵.

In the **edge weight calculation** step, we first construct a correlation vector representing co-occurrence relations for each word. The correlation vectors are constructed in the same way as the context vectors used in context-similarity-based methods (Section 2); here context words are words in a predefined window (window size of 4), the association measure is PMI, and context words are treated separately for each position. A correlation vector for each position is computed separately, and then concatenated into a single vector within the window. Second, we calculate similarities between correlation vectors. While there are numerous similarity measures that can be used, we opt to use the cosine similarity:

$$w_{ij} = \text{Cos}(\vec{f}_i, \vec{f}_j) = \frac{\vec{f}_i \cdot \vec{f}_j}{|\vec{f}_i| |\vec{f}_j|}, \quad (2)$$

⁵ This assumption is justified because context similarities are based on co-occurrence relations that are preserved across different languages.

where $\vec{f}_i$ (or $\vec{f}_j$) is the correlation vector of v_i (or v_j). Finally, in the **edge pruning** step, we preserve the edges with the top 100 weights for each vertex.

3.2 Seed Propagation

LP is a graph-based technique that transfers labels from labeled to unlabeled data to infer labels for unlabeled data. This general technique is primarily used when there is a scarcity of labeled data but abundant unlabeled data. LP has been successfully applied to common natural language processing tasks, such as word sense disambiguation (Niu, Ji, and Tan 2005; Alexandrescu and Kirchhoff 2007), part-of-speech tagging (Das and Petrov 2011), and multi-class lexicon acquisition (Alexandrescu and Kirchhoff 2007).

LP generates a label distribution for each vertex in which the weights of all labels add up to 1 by optimizing the following objective function (Zhu and Ghahramani 2002):

$$\begin{aligned}
 C(q) &= \sum_{v_i \in V \setminus V_S, v_j \in N(v_i)} w_{ij} \cdot \|q_i - q_j\| \\
 &= \sum_{v_i \in V \setminus V_S, v_j \in N(v_i)} w_{ij} \cdot \sum_z (q_i(z) - q_j(z))^2 \\
 \text{s.t. } &\sum_z q_i(z) = 1 \ \forall v_i \text{ and } q_i(z) \geq 0 \ \forall v_i, z \text{ and } q_i = u \ \forall v_i \in V_S,
 \end{aligned} \tag{3}$$

where $V_S \subseteq V$ is the subset of all vertices and corresponds to lexical seeds, $N(v_i)$ is the set of vertices linking to v_i , $q_i (i = 1 \cdots |V|)$ is the seed distribution for v_i , $q_i(z)$ is the weight of a label (i.e., a lexical seed) z in q_i , and u is a uniform distribution. The objective function acts as a smoothness regularizer that encourages the similarity in seed distributions of close vertices. We adopt the LP technique to obtain relations with all bilingual seeds, including indirect relations, by treating each lexical seed on the same language as a label.

In solving the optimization problem, we adopt an iterative update-based method described in Zhu and Ghahramani (2002)⁶. First, each lexical seed is assigned to a label. Then, the labels are iteratively propagated in the graph described in Section 3.1. After the completion of such propagations, a label distribution for each vertex is generated.

Specifically, the seed distribution for each word v_i is initialized as follows:

$$q_i^0(z) = \begin{cases} 1 & \text{if } v_i \in V_S \text{ and } z = v_i \\ 0 & \text{if } v_i \in V_S \text{ and } z \neq v_i \\ u(z) & \text{otherwise} \end{cases}, \tag{4}$$

⁶ We cannot adopt a closed-form solution that can be computed using matrix methods described in Zhu, Ghahramani, and Lafferty (2003), because it requires the inversion of a matrix of order $|V|$.

where q_i^k ($i = 1 \cdots |V|$) is the seed distribution for v_i after k propagations and $q_i^k(z)$ is the weight of a label z in q_i^k .

After initialization, we iteratively propagate the seeds through weighted edges. In each propagation, seeds are probabilistically propagated from linked vertices under the condition that larger edge weights allow seeds to propagate more easily. Thus, the closer the vertices are, the more likely they are to have similar seed distributions. In Figure 1, the balloons attached to vertices in the graphs show examples of the seed distributions generated by these propagations. For example, the English word “piranha” has a seed distribution with weights of seeds “Amazon,” “jungle,” and “fish” as 0.5, 0.3, and 0.2, respectively. Specifically, each seed distribution is updated as follows:

$$q_i^k(z) = \begin{cases} q_i^0(z) & \text{if } v_i \in V_S \\ \frac{\sum_{v_j \in N(v_i)} w_{ij} \cdot q_j^{k-1}(z)}{\sum_{v_j \in N(v_i)} w_{ij}} & \text{otherwise} \end{cases} \quad (5)$$

We ran this procedure for 10 iterations in our experiments.

Seed distributions for words in the Japanese graph in Figure 1 are computed as follows. Note that, hereafter, the first, second, and third dimension of a seed distribution correspond to lexical seeds “アマゾン,” “ジャングル,” and “魚,” respectively. First, seed distributions are initialized using Equation (4). The seed distributions for seeds “アマゾン,” “ジャングル,” and “魚” are initialized as $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, respectively. Seed distributions for “淡水” and “ピラニア,” which are not seeds, are set as uniform distribution $(0.\dot{3}, 0.\dot{3}, 0.\dot{3})$. Second, seed information is propagated through weighted edges by using Equation (5). The seed distributions for seeds “アマゾン,” “ジャングル,” and “魚” remain at the initial distribution throughout the seed propagation. The seed distribution for “淡水” is updated by transferring seed information from nearby vertices, “魚” and “ピラニア,” as follows:

$$\begin{aligned} q_{\text{淡水}}^1 &= \frac{w_{\text{淡水}\cdot\text{魚}} \times q_{\text{魚}}^0 + w_{\text{淡水}\cdot\text{ピラニア}} \times q_{\text{ピラニア}}^0}{w_{\text{淡水}\cdot\text{魚}} + w_{\text{淡水}\cdot\text{ピラニア}}} \\ &= (0.128, 0.128, 0.744). \end{aligned}$$

Similarly, the seed distribution for “ピラニア” is updated as follows:

$$\begin{aligned} q_{\text{ピラニア}}^1 &= \frac{w_{\text{ピラニア}\cdot\text{淡水}} \times q_{\text{淡水}}^0 + w_{\text{ピラニア}\cdot\text{ジャングル}} \times q_{\text{ジャングル}}^0 + w_{\text{ピラニア}\cdot\text{アマゾン}} \times q_{\text{アマゾン}}^0}{w_{\text{ピラニア}\cdot\text{淡水}} + w_{\text{ピラニア}\cdot\text{ジャングル}} + w_{\text{ピラニア}\cdot\text{アマゾン}}} \\ &= (0.509, 0.403, 0.088). \end{aligned}$$

After the first iteration, the distributions are updated as follows:

$$q_{淡水}^2 = \frac{w_{淡水・魚} \times q_{魚}^1 + w_{淡水・ピラニア} \times q_{ピラニア}^1}{w_{淡水・魚} + w_{淡水・ピラニア}}$$

$$q_{ピラニア}^2 = \frac{w_{ピラニア・淡水} \times q_{淡水}^1 + w_{ピラニア・ジャングル} \times q_{ジャングル}^1 + w_{ピラニア・アマゾン} \times q_{アマゾン}^1}{w_{ピラニア・淡水} + w_{ピラニア・ジャングル} + w_{ピラニア・アマゾン}}.$$

These procedures are iteratively performed.

3.3 Translation Pair Extraction

After LP, we treat a pair of words in different languages with similar seed distributions as a translation pair. Seed distribution can be regarded as a vector in which each dimension corresponds to each lexical seed and has an updated weight through LP. A similarity between seed distributions can therefore be calculated in a way similar to that in the context-similarity-based method. In this paper, we use the cosine similarity as the similarity measure, defined as follows:

$$Cos(q_x^f, q_y^e) = \frac{\sum_{\langle v_k^f, v_l^e \rangle \in S} q_x^f(v_k^f) \cdot q_y^e(v_l^e)}{\sqrt{\sum_{\langle v_k^f, v_l^e \rangle \in S} (q_x^f(v_k^f))^2} \sqrt{\sum_{\langle v_k^f, v_l^e \rangle \in S} (q_y^e(v_l^e))^2}}, \quad (6)$$

where q_x^f (or q_y^e) is the seed distribution for a word x (or y) in the source language (or target language) and S is the seed lexicon consisting of the word pair $\langle v_k^f, v_l^e \rangle$.

Since a seed lexicon includes one-to-many or many-to-one translation pairs, certain dimensions may be counted more than once in Equation (6). This means that each seed distribution in the **translation pair extraction** step cannot be considered as a probability distribution, although the seeds are probabilistically propagated. Therefore, we do not use a measure of the difference between two probability distributions (e.g., the Kullback-Leibler divergence).

In Figure 3, the similarity of the seed distributions for “ピラニア” and “anaconda” are computed on the basis of Equation (6). The numerator of Equation (6) is calculated as follows:

$$\begin{aligned} & \sum_{(v_i^f, v_i^e) \in S} q_{ピラニア}(v_i^f) \cdot q_{anaconda}(v_i^e) \\ &= q_{ピラニア}(アマゾン) \times q_{anaconda}(Amazon) + q_{ピラニア}(ジャングル) \times q_{anaconda}(jungle) \\ & \quad + q_{ピラニア}(魚) \times q_{anaconda}(fish) + q_{ピラニア}(魚類) \times q_{anaconda}(fish) \end{aligned}$$

Note that the dimension “fish” of seed distribution for “anaconda” is counted twice. In short,

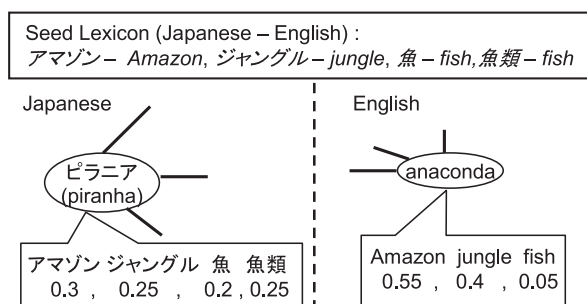


Fig. 3 Example of a Seed Lexicon and Seed Distributions

the similarity is calculated as follows:

$$\begin{aligned}
 \text{Cos}(q_{\text{ピラニア}}, q_{\text{anaconda}}) &= \frac{0.3 \times 0.55 + 0.25 \times 0.4 + 0.2 \times 0.05 + 0.25 \times 0.05}{\sqrt{0.3^2 + 0.25^2 + 0.2^2 + 0.25^2} \sqrt{0.55^2 + 0.4^2 + 0.05^2 + 0.05^2}} \\
 &= 0.83
 \end{aligned}$$

4 Experiment

4.1 Experimental Data

For our experiments, we used English and Japanese patent documents published between 1993 and 2005 by the US Patent & Trademark Office and the Japanese Patent Office, respectively. We selected the documents belonging to the *physics* domain from each monolingual corpus based on the International Patent Classification (IPC) code⁷ because it has the most documents among all domains. The comparable corpora consisted of approximately 1.5 million Japanese documents and 438,000 English documents. Table 1 shows the details of our experimental data.

Note that these documents are not aligned; however, some data includes the NTCIR parallel data used in the NTCIR-8 patent translation task for training, development, and testing purposes (Fujii, Utiyama, Yamamoto, Utsuro, Ehara, Echizen-ya, and Shimohata 2010). The majority of the “parallel” data comes from the document-aligned data that originates from Japanese patents. However, these patents are considerably lesser than the entire experimental data; only 2.0% (=29,554/1,479,831) of Japanese documents and 5.7% (=25,148/438,227) of English documents are employed in the NTCIR parallel data. Further, such documents are not completely

⁷ SECTION G of IPC code indicates the *physics* domain.

parallel at the document level, in part because the documents are automatically aligned. Therefore, our experiments are designed for bilingual lexicon mining from very comparable corpora.

We employed seed lexicons from the following two sources: (1) the EDR bilingual dictionary (EDR 1990) and (2) automatic word alignments generated by running GIZA++ (Och and Ney 2003) with the NTCIR parallel data consisting of 3,190,654 parallel sentences. From each source, we extracted pairs of nouns appearing in our corpus. From (2), we excluded word pairs in which the average of two-way translation probabilities was below 0.5. Lexicons from (1) and (2) accounted for 27,353 and 2,853 pairs, respectively, and the two sets were not exclusive. To measure the impact of seed lexicon size, we prepared the following two seed lexicons: Lex_L , a large seed lexicon, which is a union of all extracted word pairs; and Lex_S , a small seed lexicon, which is a union of a random sampling of one-tenth of the pairs from (1) and one-tenth of the pairs from (2). Table 2 shows the size of each seed lexicon. Note that our seed lexicons included one-to-many or many-to-one translation pairs. As pointed out by Morin and Prochasson (2011), it is effective to augment a seed lexicon with the lexical entries automatically extracted from in-domain parallel sentences, although the augmented seed lexicon may contain incorrect translation pairs. Our preliminary examination justified their findings, and we used the merged seed

Table 1 Data Set Employed for Our Experiments

		Japanese	English
documents	in <i>parallel</i>	29,554	25,148
	not in <i>parallel</i>	1,450,277	413,079
	total	1,479,831	438,227
content words	in <i>EDR/parallel</i>	81,006	588,097
	not in <i>EDR/parallel</i>	1,030,296	3,511,728
	total	1,111,302	4,099,825
nouns and unknown words	in <i>EDR/parallel</i>	65,207	571,145
	not in <i>EDR/parallel</i>	1,012,267	3,499,444
	total	1,077,474	4,070,589

EDR: EDR bilingual dictionary, *parallel*: NTCIR parallel data

Table 2 Size of Seed Lexicons

	Pair	Japanese Word	English Word
Lex_S	2,742	2,566	2,326
Lex_L	28,053	18,587	12,893

lexicon, (1) and (2), throughout our experiments⁸.

The Japanese texts were segmented and part-of-speech-tagged by ChaSen⁹. The English texts were tokenized and part-of-speech-tagged by TreeTagger (Schmid 1994). Next, function words were removed since they have little semantic information and spuriously co-occur with many words. Thus, our data set contained approximately 1.1 million unique Japanese content words and 4.1 million unique English content words, as shown in Table 1. Note that the number of unique English content words is four times the Japanese words, even though there are lesser English documents. This is because the English words contain words in tables or mathematical formulae; however, this is not the case for Japanese words because the data format differs for English and Japanese.

Both the EDR bilingual dictionary and the NTCIR parallel data accounted for only approximately 7.3% ($=81,006/1,111,302$) of the Japanese words and 1.4% ($=588,097/4,099,825$) of the English words in our experimental data. This is due to the presence of many technical terms and neologisms in patent documents. Therefore, the patent translation task requires bilingual lexicon extraction from nonparallel data.

In our experiments, we concentrated on content words identified as either a noun or an unknown¹⁰ by ChaSen, which accounted for 1,077,474 and 4,070,589 Japanese and English words, respectively. We further excluded those covered by either the EDR bilingual dictionary or NTCIR parallel data, which resulted in 1,012,267 and 3,499,444 Japanese and English words, respectively (Table 1). This is because we aim to complement existing bilingual dictionaries or parallel data. We randomly selected 1,000 words from the 1,012,267 Japanese words as our test data. Note that our experimental data may not include translation equivalents of the Japanese test words.

4.2 Competing Methods

We evaluated two types of our LP-based methods (*Cooc* and *Sim*) against two baselines (*Rapp* and *Andrade*). When constructing graphs for LP, *Cooc* employs co-occurrence graphs, whereas *Sim* uses similarity graphs (as described in Section 3). *Rapp* is a typical context-similarity-based method described in Section 2 (Rapp 1999). Context words are words in a window (window size of 10) and are treated separately for each position. Associations with context words are computed

⁸ Top 1 and 20 accuracies of *Cooc* using both seeds (1) and (2) are 9.2% and 28.3%, respectively. On the other hand, those using only seed (1) are 7.9% and 23.4%, respectively.

⁹ IPAdic Version 2.7.0 (<http://chasen.naist.jp/snapshot/ipadic/ipadic/doc/ipadic-ja.pdf>) was used as the dictionary. ChaSen is available for download from <http://chasen-legacy.sourceforge.jp/>.

¹⁰ We focused on performance for nouns and unknowns because they occupy majority of the words not covered by either the EDR bilingual dictionary or NTCIR parallel data. We would like to investigate other word types as our future work.

using the log likelihood ratio (Dunning 1993). The similarity measure between context vectors is the city-block metric, otherwise known as the Manhattan distance. The Manhattan distance between two vectors ($\vec{x} = (x_1, \dots, x_n)$, $\vec{y} = (y_1, \dots, y_n)$) is computed as follows:

$$\text{Manhattan}(\vec{x}, \vec{y}) = \sum_{i=1}^n |x_i - y_i|. \quad (7)$$

Andrade is a sophisticated context-similarity-based method (Andrade et al. 2010). Context is a set of words with a positive association in a window (window size of 10). The association is calculated using PMI estimated by a Bayesian method. A similarity between contexts is estimated on the basis of the number of overlapping words, as follows:

$$\text{Sim}(x, y) = -\log(P(\text{matches} = m)), \quad (8)$$

where x is a query word, y is a candidate, m is the number of overlapping words between context words of x and y , and $P(\text{matches} = m)$ is the probability that the candidate was selected by chance. $P(\text{matches} = m)$ is calculated as follows:

$$P(\text{matches} = m) = \frac{qC_m \cdot w - qC_{c-m}}{wC_c}, \quad (9)$$

where q (or c) is the number of context words of x (or y) and w is the total number of seeds.

4.3 Experimental Results

Figure 4 plots the performance of each method using Lex_S or Lex_L . Hereafter, $\text{Method}(L)$ (or $\text{Method}(S)$) denotes the *Method* using Lex_L (or Lex_S). We measured the performance on bilingual lexicon extraction as Top N ($1 \leq N \leq 20$) accuracy (Acc_N), which is the number of test words whose Top N translation candidates contain the correct translation equivalent over the total number of test words. We manually¹¹ evaluated whether the translation candidates contained a correct translation equivalent. We did not use recall because the appearance of the translation equivalents of a test word in the corpus was unknown. The Top N accuracies for $N = 1$ and 20 are summarized in Table 3.

As indicated in Table 3, our LP-based methods outperform baselines both when using Lex_S and Lex_L . The improvements are statistically significant in the sign test with a 1% significance level. Results show that capturing relations with all seeds including indirect relations, is effective. Figure 4 also indicates the merit of our methods in all Top N accuracy curves.

¹¹ We could not evaluate using existing dictionaries because most test data are technical terms and neologisms that are not included in the dictionaries.

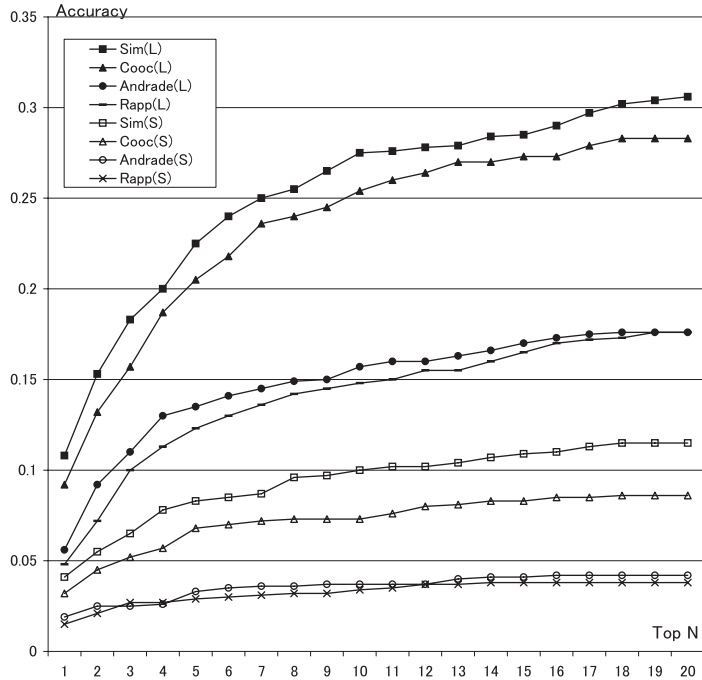


Fig. 4 Top N Accuracy Graph

Table 3 Performance on Bilingual Lexicon Extraction

	Lex_S		Lex_L	
	Acc_1	Acc_{20}	Acc_1	Acc_{20}
<i>Rapp</i>	1.5%	3.8%	4.8%	17.6%
<i>Andrade</i>	1.9%	4.2%	5.6%	17.6%
<i>Cooc</i>	3.2%	8.6%	9.2%	28.3%
<i>Sim</i>	4.1%	11.5%	10.8%	30.6%

The accuracies of the baselines are worse than previous studies: 14% Acc_1 and 46% Acc_{10} (Andrade et al. 2010) and 72% Acc_1 (Rapp 1999). This discrepancy exists because previous studies evaluated only test words whose translation equivalents existed in the experimental data, which was not always true in our experiments. Moreover, previous studies evaluated only high-frequency words, such as common nouns (Rapp 1999) and words with a document frequency of at least 50 (Andrade et al. 2010). Our test data, on the other hand, included many low-frequency words. It is generally true that the translation of high-frequency words is much easier than that of low-frequency words. We discuss the impact of test word frequencies in detail in Section 5.3.

Table 3 and Figure 4 also show that *Sim* outperforms *Cooc* both when using Lex_S and Lex_L .

The improvements of Acc_{20} are statistically significant in the sign test with a 5% significance level.

5 Discussion

5.1 Effect of Indirect Relations with Seeds

Table 4 shows a list of the top 5 translation candidates for the Japanese word “躁鬱病 (manic-depression)” for each method, where the values in parentheses indicate the ranks of the correct translations. Table 5 shows the top 5 lexical seeds that characterize the query word, where the values in the parentheses indicate weights. Table 4 shows cases in which $Cooc(L)$ found the correct translation equivalent but $Andrade(L)$ did not. Table 5 shows that $Cooc(L)$ utilized more seeds closely tied to the query word (e.g., “神経症 (neurosis),” “不眠症 (insomnia)”), which did not occur in the context of the query word in the experimental data. Results show that indirectly related seeds are also important clues, and our proposed method can utilize them.

5.2 Impact of Seed Lexicon Size

Table 3 shows that a reduction of seed lexicon size degrades performance. This is natural for the baseline methods because Lex_S cannot translate most context words that are necessary for word characterization. Consider $Andrade(L)$ and $Andrade(S)$ in the above example. Table 5 shows that $Andrade(S)$ uses fewer relevant seeds with the query word and expresses the query word by seeds with less association. For example, “精神病 (psychosis)” cannot be used in

Table 4 Translation Candidates for 躁鬱病 (manic-depression)

	$Sim(L)$ (2)	$Cooc(L)$ (5)	$Andrade(L)$ (181)
1	psychosis	polynephropathy	disease
2	manic-depression	neuroleptic	bowel
3	epilepsy	iridocyclitis	disorder
4	insomnia	Tic	symptom
5	dementia	manic-depression	sclerosis
	$Sim(S)$ (974)	$Cooc(S)$ (1652)	$Andrade(S)$ (1747)
1	ulceration	dyslnesia	bulimia
2	ulcer	encephalomyelopathy	spasticity
3	naphthol	ganglionic	Parkinson
4	dementia	corticobasal	Asymmetric
5	gastritis	praecox	anorexia

Table 5 Seeds with the Highest Weights

躁鬱病				
	<i>Cooc(L)</i>	<i>Andrade(L)</i>	<i>Cooc(S)</i>	<i>Andrade(S)</i>
1	睡眠薬 (0.12)	睡眠薬 (7.6)	痴呆 (0.016)	後天 (5.0)
	narcotic	narcotic	dementia	posteriori
2	精神病 (0.11)	老年 (6.3)	継子 (0.014)	痴呆 (3.7)
	psychosis	old	alien, stepchild	dementia
3	神経症 (0.08)	精神病 (6.3)	後天 (0.012)	潰瘍 (3.2)
	neurosis	psychosis	posteriori	ulcer
4	ホルモン (0.05)	気管支炎 (5.6)	陽性 (0.012)	ピリオド (2.9)
	hormone	bronchitis	electropositivity	period
5	不眠症 (0.04)	後天 (5.0)	潰瘍 (0.011)	重度 (2.5)
	insomnia	posteriori	ulcer	seriousness
manic-depression				
	<i>Cooc(L)</i>	<i>Andrade(L)</i>	<i>Cooc(S)</i>	<i>Andrade(S)</i>
1	illness (0.15)	illness (8.6)	ganja (0.012)	galop (7.0)
2	neurosis (0.11)	psychotherapeutics (7.0)	carbanilide (0.011)	madness (5.4)
3	seizure (0.07)	galop (7.0)	paludism (0.011)	libido (5.2)
4	psychosis (0.06)	psychosis (6.8)	resignation (0.010)	vittiligo (4.6)
5	insomnia (0.04)	somnambulism (6.7)	galop (0.009)	dementia (4.3)

Andrade(S) because *Lex_S* does not have that seed. Therefore, to find correct translation pairs is more difficult for *Andrade(S)*.

The proposed methods also share the same tendency, although each word is expressed by all seeds in the seed lexicon. Consider *Cooc(L)* and *Cooc(S)* in the above example. Table 5 shows that *Cooc(S)* expresses the query word with a smooth seed distribution, which is difficult to discriminate from others. This is because *Lex_S* did not have relevant seeds for the query word. Further, this explains why *Cooc(S)* cannot find the correct translation equivalent. On the other hand, *Cooc(L)* characterizes “躁鬱病” and “manic-depression” by strongly relevant seeds (e.g., “精神病 (psychosis),” “神経症 (neurosis)”), and then finds the correct translation equivalent.

To examine the robustness of seed lexicon size, we calculated the reduction rate of *Acc₂₀* via the following expression: $(Acc_{20} \text{ with } Lex_L - Acc_{20} \text{ with } Lex_S) / Acc_{20} \text{ with } Lex_L$. The reduction rates of *Rapp*, *Andrade*, *Cooc*, and *Sim* are 78.4%, 76.1%, 69.6%, and 62.4%, respectively. The

difference between degradation in *Cooc* and *Andrade* is statistically significant in the sign test with a 1% significance level. These results indicate that the proposed methods are more robust to seed lexicon size than the baselines, because the proposed methods can utilize seeds with indirect relations, while the baselines utilize only seeds in the context.

To verify our claim, we examined the number of test words that occurred without seeds in the context. There were 570 such words in $Rapp(S)$, 387 in $Rapp(L)$, 572 in $Andrade(S)$, and 388 in $Andrade(L)$. The baselines cannot find their translation equivalents. Such words occur even if Lex_L is used, and this number increases when Lex_S is used. Our proposed methods were able to utilize all seeds to find equivalents for such words. Therefore, our proposed methods work well even if the coverage of a seed lexicon is low.

5.3 Impact of Word Frequencies

Our test data includes many low-frequency words that are not covered by the EDR bilingual dictionary or NTCIR parallel data. More specifically, 624 words in the test data appear below 50 times. Hereafter, these words are called low-frequency words, while the other words are called high-frequency words. Table 6 shows Acc_N using Lex_L for the 624 low- and 376 high-frequency words. It also shows that performance for low-frequency words is much worse than that of high-frequency words. This is because the translation of high-frequency words utilizes abundant and reliable context information, whereas context information for low-frequency words is statistically unreliable.

In our proposed methods, edges linking rare words were sometimes generated on the basis of accidental co-occurrences; in such cases, unrelated seed information was transferred through these edges. Therefore, even our LP-based methods, especially *Cooc*, could not identify the correct translation equivalents for rare words. *Sim* alleviated the problem by using a similarity graph in which edges were generated on the basis of global correlation among words, as indicated by Table 6. It also suggests that the top 20 translation candidates for high-frequency words may

Table 6 Comparison of Performance for High- and Low-Frequency Words

	Low Freq.		High Freq.	
	Acc_1	Acc_{20}	Acc_1	Acc_{20}
$Rapp(L)$	3.3%	12.8%	7.2%	25.6%
$Andrade(L)$	3.8%	12.3%	8.6%	26.3%
$Cooc(L)$	6.4%	20.9%	13.9%	40.7%
$Sim(L)$	8.2%	23.7%	15.0%	42.0%

contribute to bilingual tasks such as MT and CLIR, although the overall performance was still low.

5.4 Effect of Similarity Graphs

We examined Acc_N for synonyms of lexical seeds in Japanese. The Acc_1 and Acc_{20} of $Sim(L)$ are 15.6% and 56.3%, respectively, while those of $Cooc(L)$ are 9.4% and 37.5%, respectively. These results show that similarity graphs are effective for clustering synonyms into the same translation equivalents. For example, $Sim(L)$ extracted the correct translation pair of the English word “iodine” and the Japanese word “イオデイン,” a synonym of the lexical seed “ヨウ素 (iodine)” in Japanese. This occurred because synonyms tend to be linked in the similarity graph and have similar seed distributions. On the other hand, in the co-occurrence graph, synonyms tend to be indirectly linked through mutual context words so that their seed distributions are far away from each another.

Further, there are many loanwords in patent documents that are spelled differently from document to document. For example, the loanword for the English word “user” is often written as “ユーザ,” but it is sometimes written as “ユーザー,” with an additional prolonged sound mark. Therefore, the similarity graph is particularly effective for the experimental data.

5.5 Impact of Seed Propagation

We also examined the impact of seed propagation. Table 7 shows the performance of our proposed methods in which the number of LP iterations varied by $k = 0, 5$, and 10 . Note that $Cooc(k = 0)$ or $Sim(k = 0)$ does not perform seed propagation. In particular, $Cooc(k = 0)$ is a conventional context-similarity-based method and is different from *Andrade* in the similarity measure between contexts: $Cooc(k = 0)$ uses the cosine similarity (Equation (6)) and *Andrade*

Table 7 Impact of Seed Propagation

	Lex_S		Lex_L	
	Acc_1	Acc_{20}	Acc_1	Acc_{20}
$Cooc(k = 0)$	1.5%	3.9%	5.2%	17.4%
$Cooc(k = 5)$	2.7%	7.3%	8.5%	26.9%
$Cooc(k = 10)$	3.2%	8.6%	9.2%	28.3%
$Sim(k = 0)$	2.3%	5.2%	6.6%	19.5%
$Sim(k = 5)$	3.6%	9.1%	8.9%	27.7%
$Sim(k = 10)$	4.1%	11.5%	10.8%	30.6%
<i>Andrade</i>	1.9%	4.2%	5.6%	17.6%

uses the measure based on the number of overlapping words (Equation (8)).

In general, we observe that more iterations improve the performance both when using Lex_S and Lex_L . In particular, at the fifth iteration, the improvements are statistically significant in the sign test with a 1% significance level. Further iterations, such as the tenth iteration, improve performance by propagating seed information to nodes that are far away from seeds in the co-occurrence or similarity graph.

5.6 Error Analysis

Besides errors caused by infrequently occurring words, our test data included words for which translation equivalents could not be found inherently. The first type includes words whose equivalents do not exist in the English corpus. Our test data included 133 such words (12 high- and 121 low-frequency words). This is an unavoidable problem for methods based on comparable corpora.

The second type includes words for which English equivalents are compound words. The Japanese morphological analyzer tends to group a compound word into a single word, while the English text analyzer does not group collocations of words divided by the delimiter *space*. For example, the single Japanese word “掌紋” is equivalent to “palm pattern” or “palm print,” which consists of two words. This case was considered as an error even though the proposed methods found the word “palm” equivalent to “掌紋.” Our test data included 142 such cases (20 high- and 122 low-frequency words). A solution to this multiword expression problem is to include compound words in the vocabulary and construct co-occurrence or similarity graphs with significantly more vertices. Due to the computational cost, we leave this for future work.

Table 8 shows the results of recalculation after removing these two types of error words from our test data. In the table, “High Freq.” corresponds to a standard experimental setting, which evaluated only high-frequency words whose translation equivalents existed in the experimental data.

The primary reason for other errors is word sense ambiguity, which is different in every language. For example, the Japanese word “右” means “right” and “conservatism” in English. Our proposed methods merged different senses by propagating seeds through these polysemous words in only one language. This caused translation pairs to have wrong seed distributions, and therefore our proposed methods could not identify correct translation pairs. We leave this word sense disambiguation problem for future work.

Table 8 Performance under Ideal Setting

	Acc_1 (Low Freq., High Freq.)	Acc_{20} (Low Freq., High Freq.)
<i>Rapp</i> (<i>L</i>)	6.6% (5.5%, 7.8%)	24.3% (21.0%, 27.9%)
<i>Andrade</i> (<i>L</i>)	7.7% (6.3%, 9.3%)	24.3% (20.2%, 28.8%)
<i>Cooc</i> (<i>L</i>)	12.7% (10.5%, 15.1%)	39.0% (34.1%, 44.5%)
<i>Sim</i> (<i>L</i>)	14.9% (13.4%, 16.6%)	42.2% (38.8%, 45.9%)

6 Related Studies

Besides the comparable corpus approach discussed in Section 2, many alternatives have been proposed for bilingual lexicon extraction. The first alternative includes finding translation pairs in parallel corpora (Wu and Xia 1994; Fung and Church 1994; Och and Ney 2003). However, large parallel corpora are available only for a few language pairs and limited domains. Moreover, even the large parallel corpora are relatively small in comparison with comparable corpora.

The second alternative includes exploiting the web. Lu et al. (2004) extracted translation pairs by mining web anchor texts and link structures. As an alternative, mixed-language webpages are analyzed by first retrieving texts that include both source and target languages from the web via search engines or simple rules, and then extracting translation pairs from the mixed-language texts utilizing various clues. Zhang and Vines (2004) used co-occurrence statistics; Cheng et al. (2004) used co-occurrences and context similarity information; and Huang et al. (2005) used phonetic, semantic, and frequency-distance features. Lin et al. (2008) proposed a method for extracting parenthetically translated terms, where a word alignment algorithm is used for establishing the correspondences between in- and pre-parenthesis words. However, these methods cannot find translation pairs when they are not linked through link structures or when they do not co-occur in the same text.

Transliteration is a completely different method of bilingual lexicon acquisition in which a word in one language is converted into another language using phonetic equivalence (Knight and Graehl 1998; Karimi, Scholer, and Turpin 2011). Although machine transliteration works particularly well for proper names and loanwords, it cannot be employed for phonetically dissimilar translations.

All abovementioned methods may extract translation pairs more precisely than our comparable corpus approach when their underlying assumptions are satisfied. Since their methods are orthogonal to our LP-based method, we might improve the performance by augmenting a seed lexicon with translation pairs extracted using the above methods.

7 Conclusion

In this paper, we proposed a novel bilingual lexicon extraction method using LP for alleviating the problem of limited seed lexicon size. Our proposed method captured relations with all seeds, including indirect relations, by propagating seed information. Further, in addition to co-occurrence graphs, we proposed our method using similarity graphs in the propagation process. Our experiments showed that the proposed method outperforms conventional context-similarity-based methods (Rapp 1999; Andrade et al. 2010) and the similarity graphs improve the performance by clustering synonyms into the same translation.

We plan to investigate many open problems in our future work. One such problem is word sense disambiguation and the translation of compound words as described in Daille and Morin (2005) and Morin et al. (2007). In addition, indirect relations have been used in other tasks, such as paraphrase acquisition from bilingual parallel corpora (Kok and Brockett 2010). We will utilize their random walk approach and other graph-based techniques, such as modified adsorption (Talukdar and Crammer 2009) to generate seed distributions. We also plan an end-to-end evaluation, for instance, by employing the extracted bilingual lexicon in an MT system.

Acknowledgement

This paper is based on the conference paper by Tamura, Watanabe, and Sumita (2012), extended with a more complete description of our proposed method and the conventional methods and an additional analysis of the experimental results. We thank anonymous reviewers for their helpful suggestions and comments on the first version of this paper.

Reference

- Alexandrescu, A. and Kirchhoff, K. (2007). “Data-Driven Graph Construction for Semi-Supervised Graph-Based Learning in NLP.” In *Proceedings of NAACL-HLT 2007*, pp. 204–211.

- Andrade, D., Matsuzaki, T., and Tsujii, J. (2011a). “Effective Use of Dependency Structure for Bilingual Lexicon Creation.” In *Proceedings of CICLing 2011*, pp. 80–92.
- Andrade, D., Matsuzaki, T., and Tsujii, J. (2011b). “Learning the Optimal Use of Dependency-parsing Information for Finding Translations with Comparable Corpora.” In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pp. 10–18.
- Andrade, D., Nasukawa, T., and Tsujii, J. (2010). “Robust Measurement and Comparison of Context Similarity for Finding Translation Pairs.” In *Proceedings of COLING 2010*, pp. 19–27.
- Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., and Chien, L.-F. (2004). “Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval.” In *Proceedings of SIGIR 2004*, pp. 146–153.
- Chiao, Y.-C. and Zweigenbaum, P. (2002). “Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora.” In *Proceedings of COLING 2002*, pp. 1208–1212.
- Daille, B. and Morin, E. (2005). “French-English Terminology Extraction from Comparable Corpora.” In *Proceedings of IJCNLP 2005*, pp. 707–718.
- Das, D. and Petrov, S. (2011). “Unsupervised Part-of-Speech Tagging with Bilingual Graph-Based Projections.” In *Proceedings of ACL-HLT 2011*, pp. 600–609.
- Daumé III, H. and Jagarlamudi, J. (2011). “Domain Adaptation for Machine Translation by Mining Unseen Words.” In *Proceedings of ACL-HLT2011*, pp. 407–412.
- Déjean, H., Gaussier, É., and Sadat, F. (2002). “An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction.” In *Proceedings of COLING 2002*, pp. 1–7.
- Dunning, T. (1993). “Accurate Methods for the Statistics of Surprise and Coincidence.” *COMPUTATIONAL LINGUISTICS*, **19** (1), pp. 61–74.
- EDR (1990). “Bilingual Dictionary.” Tech. rep., TR-029, Japan Electronic Dictionary Research Institute, Tokyo.
- Fišer, D., Ljubešić, N., Vintar, Š., and Pollak, S. (2011). “Building and Using Comparable Corpora for Domain-Specific Bilingual Lexicon Extraction.” In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pp. 19–26.
- Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H., and Shimohata, S. (2010). “Overview of the Patent Translation Task at the NTCIR-8 Workshop.” In *Proceedings of the 8th NTCIR Workshop*, pp. 371–376.
- Fung, P. (1995). “Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus.” In *Proceedings of the 3rd Annual Workshop on Very Large Corpora*, pp. 173–183.
- Fung, P. and Church, K. W. (1994). “K-vec: A New Approach for Aligning Parallel Texts.” In

- Proceedings of COLING 1994*, pp. 1096–1102.
- Fung, P. and McKeown, K. (1997). “Finding Terminology Translations from Non-parallel Corpora.” In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pp. 192–202.
- Fung, P. and Yee, L. Y. (1998). “An IR Approach for Translating New Words from Nonparallel, Comparable Texts.” In *Proceedings of COLING-ACL 1998*, pp. 414–420.
- Garera, N., Callison-Burch, C., and Yarowsky, D. (2009). “Improving Translation Lexicon Induction from Monolingual Corpora via Dependency Contexts and Part-of-Speech Equivalences.” In *Proceedings of CoNLL 2009*, pp. 129–137.
- Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004). “A Geometric View on Bilingual Lexicon Extraction from Comparable Corpora.” In *Proceedings of ACL 2004*, pp. 526–533.
- Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D. (2008). “Learning Bilingual Lexicons from Monolingual Corpora.” In *Proceedings of ACL-08: HLT*, pp. 771–779.
- Hazem, A., Morin, E., and Saldarriaga, S. P. (2011). “Bilingual Lexicon Extraction from Comparable Corpora as Metasearch.” In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pp. 35–43.
- Huang, F., Zhang, Y., and Vogel, S. (2005). “Mining Key Phrase Translations from Web Corpora.” In *Proceedings of HLT-EMNLP 2005*, pp. 483–490.
- Ismail, A. and Manandhar, S. (2010). “Bilingual Lexicon Extraction from Comparable Corpora using In-domain Terms.” In *Proceedings of COLING 2010*, pp. 481–489.
- Kaji, H. (2005). “Extracting Translation Equivalents from Bilingual Comparable Corpora.” *IEICE transactions on information and systems*, **E88-D** (2), pp. 313–323.
- Karimi, S., Scholer, F., and Turpin, A. (2011). “Machine Transliteration Survey.” *ACM Computing Surveys*, **43** (3), pp. 1–46.
- Knight, K. and Graehl, J. (1998). “Machine Transliteration.” *Computational Linguistics*, **24** (4), pp. 599–612.
- Koehn, P. and Knight, K. (2002). “Learning a Translation Lexicon from Monolingual Corpora.” In *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, pp. 9–16.
- Kok, S. and Brockett, C. (2010). “Hitting the Right Paraphrases in Good Time.” In *Proceedings of HLT-NAACL 2010*, pp. 145–153.
- Laroche, A. and Langlais, P. (2010). “Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora.” In *Proceedings of COLING 2010*, pp. 617–625.
- Laws, F., Michelbacher, L., Dorow, B., Scheible, C., Heid, U., and Schütze, H. (2010). “A

- Linguistically Grounded Graph Model for Bilingual Lexicon Extraction.” In *Proceedings of COLING 2010*, pp. 614–622.
- Lin, D., Zhao, S., Durme, B. V., and Pasca, M. (2008). “Mining Parenthetical Translations from the Web by Word Alignment.” In *Proceedings of ACL-08: HLT*, pp. 994–1002.
- Lu, W.-H., Chien, L.-F., and Lee, H.-J. (2004). “Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach.” *ACM Transactions on Information Systems*, **22** (2), pp. 242–269.
- Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). “Bilingual Terminology Mining—Using Brain, not brawn comparable corpora.” In *Proceedings of ACL 2007*, pp. 664–671.
- Morin, E. and Prochasson, E. (2011). “Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora.” In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pp. 27–34.
- Niu, Z.-Y., Ji, D.-H., and Tan, C. L. (2005). “Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning.” In *Proceedings of ACL 2005*, pp. 395–402.
- Och, F. J. and Ney, H. (2003). “A Systematic Comparison of Various Statistical Alignment Models.” *Computational Linguistics*, **29** (1), pp. 19–51.
- Otero, P. G. and Campos, J. R. P. (2008). “Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary.” In *Proceedings of CICLing 2008*, pp. 423–433.
- Pekar, V., Mitkov, R., Blagoev, D., and Mulloni, A. (2006). “Finding Translations for Low-Frequency Words in Comparable Corpora.” *Machine Translation*, **20** (4), pp. 247–266.
- Prochasson, E. and Fung, P. (2011). “Rare Word Translation Extraction from Aligned Comparable Documents.” In *Proceedings of ACL-HLT 2011*, pp. 1327–1335.
- Rapp, R. (1995). “Identifying Word Translations in Non-Parallel Texts.” In *Proceedings of ACL 1995*, pp. 320–322.
- Rapp, R. (1999). “Automatic Identification of Word Translations from Unrelated English and German Corpora.” In *Proceedings of ACL 1999*, pp. 519–526.
- Schmid, H. (1994). “Probabilistic Part-of-Speech Tagging Using Decision Trees.” In *Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49.
- Shao, L. and Ng, H. T. (2004). “Mining New Word Translations from Comparable Corpora.” In *Proceedings of COLING 2004*, pp. 618–624.
- Talukdar, P. P. and Crammer, K. (2009). “New Regularized Algorithms for Transductive Learning.” In *Proceedings of ECML-PKDD 2009*, pp. 442–457.
- Tamura, A., Watanabe, T., and Sumita, E. (2012). “Bilingual Lexicon Extraction from Compara-

- ble Corpora Using Label Propagation.” In *Proceedings of EMNLP-CoNLL 2012*, pp. 24–36.
- Vulić, I., Smet, W. D., and Moens, M.-F. (2011). “Identifying Word Translations from Comparable Corpora Using Latent Topic Models.” In *Proceedings of ACL-HLT 2011*, pp. 479–484.
- Wu, D. and Xia, X. (1994). “Learning an English-Chinese Lexicon from a Parallel Corpus.” In *Proceedings of AMTA 1994*, pp. 206–213.
- Zhang, Y. and Vines, P. (2004). “Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval.” In *Proceedings of SIGIR 2004*, pp. 162–169.
- Zhu, X. and Ghahramani, Z. (2002). “Learning from Labeled and Unlabeled Data with Label Propagation.” Tech. rep., CMU-CALD-02-107.
- Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). “Semi-supervised Learning using Gaussian Fields and Harmonic Functions.” In *Proceedings of ICML 2003*, pp. 912–919.

Akihiro Tamura: received his B.E. and M.E. from the Tokyo Institute of Technology in 2005 and 2007, respectively. He is currently a researcher at the National Institute of Information and Communications Technology. His current research interest is natural language processing. He is a member of the Information Processing Society of Japan and the Association for Computational Linguistics.

Taro Watanabe: received his B.E. and M.E. degrees in Information Science from Kyoto University in 1994 and 1997, respectively, and obtained his M.S. degree in Language and Information Technologies from the School of Computer Science, Carnegie Mellon University in 2000. In 2004, he earned his Ph.D. in Informatics from Kyoto University. After serving as a researcher at ATR and NTT, Dr. Watanabe is currently a senior researcher at the National Institute of Information and Communications Technology. His research interests include natural language processing, machine learning and statistical machine translation.

Eiichiro Sumita: received his M.S. in Computer Science from the University of Electro-Communications in 1982 and his Ph.D. in Engineering from Kyoto University in 1999. Dr. Sumita is the Director of the NICT Multilingual Translation Laboratory. His research interests include machine translation and e-Learning.

Hiroya Takamura: received his B.E. and M.E. from the University of Tokyo in 1997 and 2000, respectively (in 1999, he was a research student at Technische Universitaet von Wien). He received his Dr. Eng. from Nara Institute of Science and Technology in 2003. He was an assistant professor from 2003 to 2010 and is currently an associate professor at the Tokyo Institute of Technology. His current research interest is computational linguistics. He is a member of the Information Processing Society of Japan and the Association for Computational Linguistics.

Manabu Okumura: received his B.E., M.E., and Dr. Eng. from the Tokyo Institute of Technology in 1984, 1986, and 1989, respectively. He was an assistant professor at the Department of Computer Science, Tokyo Institute of Technology from 1989 to 1992, and an associate professor at the School of Information Science, Japan Advanced Institute of Science and Technology from 1992 to 2000. He was also a visiting associate professor at the Department of Computer Science, University of Toronto from 1997 to 1998. He is currently an associate professor at Precision and Intelligence Laboratory, Tokyo Institute of Technology. His current research interests include natural language processing, especially automatic text summarization, sentiment analysis, and text mining. He is a member of the Information Processing Society of Japan, the American Association for Artificial Intelligence, the Association for Computational Linguistics, and the Japanese Cognitive Science Society.

(Received October 9, 2012)

(Revised December 21, 2012)

(Accepted February 14, 2013)