NAIST-NTT System Description for Patent Translation Task at NTCIR-7

Mamoru Komachi† Masaaki Nagata‡ Yuji Matsumoto†

† Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{mamoru-k,matsu}@is.naist.jp
‡ NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
nagata.masaaki@lab.ntt.co.jp

Abstract

This paper proposes a semi-supervised approach to acquire domain specific translation knowledge from the collection of Wikipedia. The proposed method starts from a small number of seed translation pairs for each domain in a given corpus, and applies the regularized Laplacian to learn translation pairs relevant to the domain. This paper presents evaluation results using the NTCIR-7 Patent Translation Task.

Keywords: *Semi-supervised learning, Translation knowledge acquisition, Wikipedia mining.*

1. Introduction

Translation knowledge acquisition is one of the central research topics of machine translation. However, translation knowledge acquisition often depends on human annotation. Especially, annotation of technical terms requires domain knowledge. So far, reducing the cost of human annotation is one of the important problems for building machine translation systems.

To minimize the cost of hand-tagging resources, Wikipedia has been studied as a source of bilingual lexicon extraction. Wikipedia is a multilingual free online encyclopedia which is maintained by a community of volunteers. Currently, the English Wikipedia is the largest one with 2,297,611 articles, while the Japanese is the fifth place with 479,908 articles. More than 200,000 articles have both English and Japanese versions, and can easily be aligned by interlingual hyperlinks of Wikipedia. However, naïve extraction results in noisy bilingual lexicon, because Wikipedia has many ambiguous titles such as 1453 (English) pointing to 1453 年 [*year*] (Japanese). Also, domain adaptation of the extracted lexicon is a key issue since Wikipedia is a general-purpose encyclopedia.

Adafre and Rijke [1] proposed to use interlingual links in Wikipedia articles to obtain a bilingual lexicon. Their approach of using the interlingual links is straightforward. For each Wikipedia page in one language, they extracted interlingual hyperlinks as translations of the titles in other languages. Although they addressed the problem of ambiguous translations, they did not seem to disambiguate translation pairs since their aim was to find similar sentences instead of learning a high-quality bilingual lexicon.

Recently, Erdmann [5] showed that a link structure (combination of redirect page and link text information) can boost recall of bilingual lexicon extraction from Wikipedia. However, their method does not improve accuracy of bilingual lexicon extraction, and they did not evaluate the quality of the extracted results on machine translation system. The problem of selecting appropriate word sense still exists.

In this paper, we focus on extraction of a bilingual lexicon from Wikipedia. We propose a graph-based semi-supervised learning algorithm to refine a bilingual lexicon. Semi-supervised approaches have been adopted in many tasks such as word sense disambiguation [16, 11, 10] and named entity recognition [7, 4].

We followed the *one sense per domain* assumption described in [15] and extract the most likely translation pairs for each domain. We apply the recently proposed graph-theoretic algorithm, *regularized Laplacian*, to the task of finding the most relevant translation pairs to the domain at hand. Graph-based methods have attracted attention in NLP tasks recently, such as word sense disambiguation [9], knowledge acquisition [14] and language modeling [12].

Our work (1) extracts a bilingual lexicon from Wikipedia and measure its quality on machine translation task, and (2) refines a bilingual lexicon based on the graph structure of Wikipedia.

The rest of this paper is organized as follows. We pro-

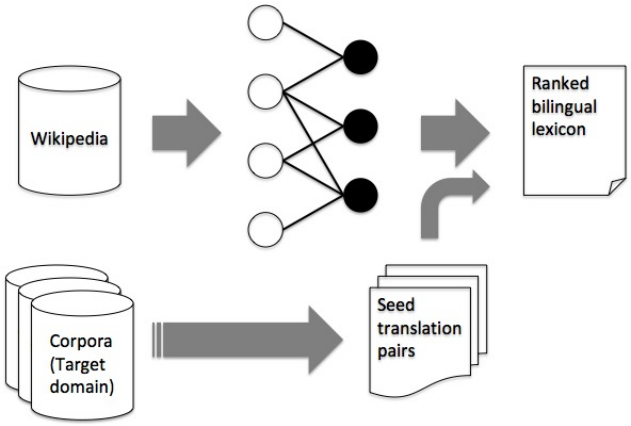

Figure 1. Overview of extraction of a bilingual lexicon of a target domain from Wikipedia

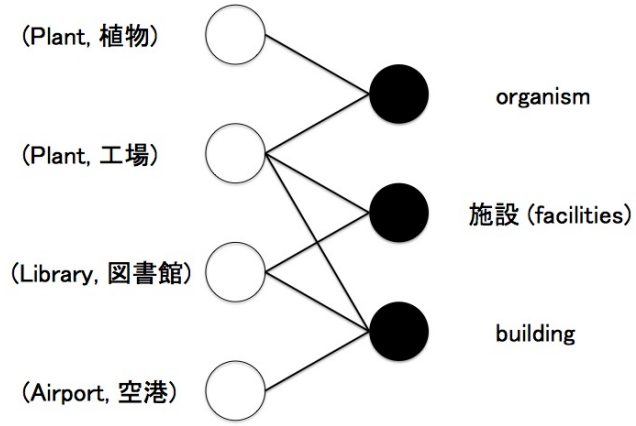

Figure 2. Bipartite graph from Wikipedia link structure. White nodes indicate translation pairs, whereas black nodes indicates patterns which co-occur with connected translation pairs.

pose a graph-based algorithm to disambiguate translation pairs in Section 2. It is used in link analysis community to find the most relevant node. In Section 3 we explain the experimental settings of our system for Patent Translation Task at NTCIR-7[6]. We evaluate and discuss experimental results in Section 4. Finally, we conclude our work and give future directions in Section 5.

2. Graph-based Algorithm to Acquire Domain Specific Bilingual Lexicon

We propose a semi-supervised approach to learn domain specific translation pairs from Wikipedia. Figure 1 depicts the overview of our extraction algorithm. The algorithm constructs a bipartite graph from Wikipedia and computes similarity between translation pairs over the graph. It requires only a small amount of translation pairs to disambiguate ambiguous translation pairs in a bilingual lexicon. It ranks a bilingual lexicon according to the similarity measure given seed translation pairs in a given domain.

2.1. Creating a Bipartite Graph from Wikipedia

First, we follow the steps described in [1] and extract a bilingual lexicon from Wikipedia. Wikipedia provides a vast number of named entities and technical terms. Some articles are associated with interlingual links. An interlingual link in Wikipedia is a link between two articles For example, [[en:Manga]] points to the English version of the article "Manga," which has an outgoing link to the Japanese version [[ja:漫画]] (*manga*). Translations of a page title (typically a noun phrase) are then given as the interlingual hyperlinks from that page. By taking the intersection of obtained list of translation pairs in both directions [1] one can obtain a large bilingual lexicon in reasonable quality.

There are cases where a word has more than one sense. In such a case, Wikipedia provides a special page called *disambiguation* to help disambiguate word senses. One of the main problem in machine translation is to select which word sense is appropriate for a given context. We assume the *one sense per domain* hypothesis [15] by exploiting the fact that the distribution of word senses is highly skewed depending on domains. According to the hypothesis, the task of selecting the right sense is then to select the most relevant sense to the given domain.

To calculate the relatedness of a translation pair to a given domain, we manually prepare seed translation pairs from the domain and measure similarity between a translation pair and the seeds. The similarity is computed over a bipartite graph created from interlingual links and abstracts of Wikipedia[2]

Figure 2 illustrates a bipartite graph constructed from Wikipedia. In this bipartite graph, related translation pairs tend to connect to similar set of patterns (e.g. (Library, 図書館) is more similar to (Plant, 工場) than (Plant, 植物) because they share two patterns, "施設" *facilities* (a term occurring in Japanese abstract) and "building" (a term occurring in English abstract)), and vice versa.

[1] Some page titles can not be translated back to their original ones (e.g. ambiguous words), and thus the two sets of translation pairs are not necessarily identical.

[2] Abstracts are automatically generated by taking the first paragraph of each page. The abstract file (abstract.xml) is distributed as part of dump files of Wikipedia database and can be downloaded at http://download.wikimedia.org/.

The steps for constructing a bipartite graph is defined as follows:

1. Add *translation pairs* (en,ja) as white nodes.
2. Add bag-of-content words (hereafter referred to as *patterns*) appearing in abstracts of both languages as black nodes. Note that a pattern may be either single English or Japanese word.
3. Add edges from translation pairs to co-occurring patterns.

The intuition behind the bipartite graph construction is that if two translation pairs share similar patterns they must be related. We will explain similarity measure on this graph in the next subsection.

2.2. A Graph-based Similarity Measure between Translation Pairs

Second, we estimate similarity between translation pairs in a bilingual lexicon by the regularized Laplacian [13, 3], which is used in link analysis community to calculate relatedness between nodes based on the graph Laplacian.

Let $|T|$ and $|P|$ be the numbers of translation pairs and patterns, respectively, and M be a pattern-translation pair matrix whose (p,t)-element $[M]_{pt}$ holds the number of co-occurrence of pattern p and translation pair t in Wikipedia. A symmetric matrix $A = M^T M$ holds the similarity between translation pairs. The matrix A is obtained from the bipartite graph M.

We then compute the graph Laplacian based on the similarity matrix A. Let G be a weighted undirected graph whose adjacency (weight) matrix is the symmetric matrix A. Let $\rho(A)$ denote the spectral radius of A. The (combinatorial) graph Laplacian L of a graph G is defined as follows:

$$L = D - A \tag{1}$$

where D is a diagonal matrix, in which the ith diagonal element $[D]_{ii}$ is given by

$$[D]_{ii} = \sum_j [A]_{ij}. \tag{2}$$

Here, $[A]_{ij}$ stands for the (i,j) element of A. The *regularized Laplacian kernel* R_β with diffusion factor $\beta(\rho(L) > \beta \geq 0)$ is defined as follows: [3]

$$R_\beta = \sum_{n=0}^{\infty} \beta^n (-L)^n = (I + \beta L)^{-1}. \tag{3}$$

[3] It has been reported that normalization of A improves performance in application [8], so we normalize L by $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$.

Using this kernel, the algorithm takes a seed vector $\mathbf{t_0}$ to compute a score vector of a translation pair $\mathbf{t}$:

$$\mathbf{t} = R_\beta \mathbf{t_0} \tag{4}$$

where $\mathbf{t_0}$ is a $|T|$-dimensional vector with 1 at the position of seed translation pairs, and 0 elsewhere. $\mathbf{t}$ can be regarded as a ranked list of translation pairs sorted by the score of the ith element of the vector.

The regularized Laplacian computes all the possible paths in a graph, and consequently can calculate influence between nodes in a long distance in the graph. Also, Equations (1) and (2) show that the negative Laplacian $-L$ can be regarded as a modification to the graph G with the weight of self-loops re-weighted to negative values. In this modified graph, if a translation pair co-occurs with a pattern which also co-occurs with a large number of other translation pairs, a self-loop of a node in the instance similarity graph induced by $M^T M$ will receive a higher negative weight. This results in down-weighting such translation pairs and patterns, and hence weakens the effect of high frequent translation pairs and patterns.

Seed translation pairs are expected to be the representative of the domain, and thus should co-occur with domain-specific patterns. To filfill this requirement, frequent but unambiguous translation pairs should be carefully selected.

3. Experiment

3.1. Corpus and Tools

We used the first Patent Parallel Corpus (PPC-1) for the experiment. We only used Parallel Sentence Data (PSD). The data is treated as a simple list of parallel sentences. No context and structural information which could be obtained from the Parallel Patent Data (PSD) is used.

The PSD of PPC-1 comes in four files: a training data file of about 1.8 million parallel sentences `train.txt` and three development data files of about a thousand parallel sentences {`dev`,`devtest`,`test`}`.txt`. We used `train.txt` for training, `dev.txt` for parameter tuning, and `test.txt` for testing during development.

We used an open source statistical machine translation system Moses[4] as a baseline system for NTCIR-7. We basically followed the instructions written at the homepage of WMT2008[5] to built the baseline system for their shared task. We build the language model by using the SRI Language Modeling Toolkit[6].

[4] http://www.statmt.org/moses/
[5] http://www.statmt.org/wmt08/
[6] http://www.speech.sri.com/projects/srilm/

3.2 Preprocessing and Evaluation

English sentences are tokenized and lowercased by using `tokenizer.perl` and `lowercase.perl`, the scripts provided by the WMT2008 organizers. As for Japanese sentences, its encoding is first converted from EUC-JP to UTF-8 and they are normalized under NFKC by using the Perl library. They are then word segmented by using the open source Japanese morphological analyzer MeCab[7].

In normalization form NFKC, Compatibility Decomposition and Canonical Composition is performed to unicode string. In Japanese, it roughly means double width alphabets and numbers are converted to single width, and single width Katakanas are converted to double width.

Before building the translation model, long sentences with more than 80 words are removed by using the script `clean-corpus-n.perl`. This reduces the number of training sentence pairs from 1,798,571 to 1,768,853. Both translation model and language model are made from the resulting bilingual sentences pairs.

For English outputs, detokenization is done by the script `detokenizer.perl`. Recaser is trained by using Moses from the English side of the training sentences as described in the WMT2008 baseline system. BLEU score is computed by the script `doc_bleu.rb` provided by the NTCIR-7 organizers.

3.3. Bilingual Lexicon Extraction

The use of a bilingual lexicon from Wikipedia described in Section 2 is straightforward. We add the extracted bilingual lexicon to the training corpus to learn the translation probability between translation pairs.[8]

A snapshot of Wikipedia was taken on 12 March 2008. The page titles are aligned by interlingual hyperlinks and 222,739 translation pairs are extracted in total. Non-Japanese nor English characters such as Arabic and Cyrillic are removed. Most of the formatting information which is not relevant for the current task are discarded. Eventually, 197,770 translation pairs are retained for the full Wikipedia bilingual lexicon.

We randomly split the bilingual lexicon into 8 sub lexicons (due to memory limits). 5 seeds are manually chosen for each sub lexicon (total $8 \times 5 = 40$ seeds). Sample seed translation pairs are displayed in Table 1.

After applying the regularized Laplacian kernel, the top 10%, 50% and 75% of the ranked list for each sub lexicon are collected. The intersection of the 8 collected lists is the 10%, 50% and 75% bilingual lexicons, respectively. Table 2 shows the number of translation pairs for each bilingual lexicon, along with several examples.[9]

Table 1. Sample seed translation pairs

(thermal spray, 溶射)
(epoxy, エポキシ樹脂)
(single crystal, 単結晶)
(laser cooling, レーザー冷却)
(centrifugal compressor, 遠心式圧縮機)

Table 3. BLEU score for Patent Translation Task at NTCIR-7

	single-ref		fmlrun-int	
	JE	EJ	JE	EJ
baseline	26.39	28.25	25.34*1	27.19*3
Wikipedia (10%)	—	27.47	—	—
Wikipedia (50%)	—	27.46	—	—
Wikipedia (75%)	—	27.42	—	—
Wikipedia (100%)	26.48	27.28	25.48*2	28.15*4

3.4. Results

Table 3 presents BLEU scores for each translation direction for Patent Translation Task at NTCIR-7. The results of adding Wikipedia as a bilingual lexicon is shown. Wikipedia (10,50,75%) compares the effect of the graph-based refinement of the bilingual lexicon. *Fmlrun-int* stands for intrinsic evaluation for the formal run, while *single-ref* uses the reference sentence distributed with the fmlrun-int dataset to compute BLEU score.

(*1) and (*2) correspond to the submitted results with GROUP-ID "NAIST-NTT" for RUN 1 and 2 on the TASK "JE," whereas (*3) and (*4) correspond to the submitted results with GROUP-ID "NAIST-NTT" for RUN 1 and 2 on the TASK "EJ," respectively.

4. Discussion

Table 3 demonstrates that the extracted bilingual lexicon slightly improves BLEU score (0.09 for fmlrun-int and 0.14 for official) in Japanese to English translation. However, adding the extracted bilingual lexicon constantly degrades

[7] http://mecab.sourceforge.net/

[8] One of the common ways of using a dictionary in GIZA++ is to include it as additional training data. See articles in Moses mailing list http://article.gmane.org/gname.comp.nlp.moses.user/921 for detail.

[9] The total number of words for each ranked lexicon does not necessarily proportional to the full Wikipedia since there are duplicates in the split sub lexicons.

Table 2. Samples of the extracted bilingual lexicon from Wikipedia. Coverage of unknown words in the test corpus is also shown in parenthesis.

Wikipedia	♯ of words	samples
10%	11,970 (1.9%)	(natural selection, 自然選択説), (scrabble, スクラブル) , (phase transition, 相転移), (diamond, ダイアモンド), (videocassette recorder, ビデオテープレコーダ)
50%	75,420 (7.7%)	(movement for multiparty democracy, 複数政党制民主主義運動), (fentanyl, フェンタニル) [an opioid analgestic], (sigma sagittarii, ヌンキ) [the second brightest star system in the constellation Sagittarius]), (shintaro abe, 安倍晋太郎) [the former prime minister of Japan], (nippon television, 日本テレビ放送網)
75%	113,277 (11.5%)	(pride final conflict 2003, pride grandprix 2003 決勝戦) [a mixed martial arts event held by PRIDE Fighting Championships], (uglyness, 醜), (palma il vecchio, パルマ・イル・ヴェッキオ) [an Italian painter], (jean gilles, ジャン・ジル) [a French composer; a French soldier], (amiloride, アミロライド) [a potassium-sparing diuretic]
100%	197,770 (13.5%)	(brilliant corners, ブリリアント・コーナーズ) [an album by a jazz musician], (charly mottet シャーリー・モテ) [a French former professional cyclist], (deep purple in rock, ディープ・パープル・イン・ロック) [an album by an English rock band], (june 2003, 「最近の出来事」2003 年 6 月) [navigational entry for events happened in June 2003], (moanin', モーニン) [a jazz album]
filtered	24,969	(1, 1 年) [year], (UTC+9, UTC+9) [Japanese side contains only alphanumeric characters], (Aera, AERA) [case-insensitive match] (大岡越前, 大岡越前) [garbage in English side], (image:himeji castle frontview.jpg, himeji castle frontview.jpg) [Wikipedia format navigational links], (user:eririnrinrin, eririnrinrin) [Wikipedia specific entries],

BLEU score for fmlrun-int dataset in English to Japanese direction, while it outperforms baseline in BLEU score by 1 for the official run. It is not clear why the reported results are not consistent with the results of fmlrun-int, and thus re-examination is needed to verify the efficiency of the proposed method.

By comparing Wikipedia (75,100%) and others, it is suggested that adding the whole bilingual lexicon extracted from Wikipedia may be too noisy to learn phrase alignments. One possibility is to extract only highly relevant terms to the domain (at the expense of coverage), and another possibility is to investigate better way to integrate a bilingual lexicon to phrase-based statistical machine translation.

5. Conclusion and Future Work

In this paper, we demonstrated that a large scale bilingual lexicon can be extracted from Wikipedia. The bilingual lexicon may be improved in its quality by a graph-based kernel. We have reported the results on NAIST-NTT system for Patent Translation Task at NTCIR-7.

Although adding a dictionary to a training corpus has the advantage of simplicity, it is not the best way to incorporate word sense disambiguation into machine translation system. Carpuat et al.[2] showed that reranking of the phrase table improves performance of statistical machine translation. It is one of the future work to integrate graph-based word sense disambiguation into statistical machine translation framework.

Acknowledgements

The first author is partially supported by the Japan Society for Promotion of Science (JSPS), Grant-in-Aid for JSPS Fellows.

References

[1] S. F. Adafre and M. de Rijke. Finding Similar Sentences across Multiple Languages in Wikipedia. In *Proceedings of EACL 2006 Workshop: Wikis and blogs and other dynamic text source*, 2006.

[2] M. Carpuat and D. Wu. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, 2007.

[3] P. Y. Chebotarev and E. V. Shamis. On proximity measures for graph vertices. *Automation and Remote Control*, 59(10):1443–1459, 1998.

[4] M. Collins and Y. Singer. Unsupervised Models for Named Entity Classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110, 1999.

[5] M. Erdmann. Extraction of Bilingual Terminology from the Link Structure of Wikipedia, 2008. Master's thesis, Osaka University, Osaka, Japan.

[6] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 2008.

[7] M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pages 539–545, 1992.

[8] R. Johnson and T. Zhang. On the Effectiveness of Laplacian Normalization for Graph Semi-supervised Learning. *Journal of Machine Learning Research*, 8:1489–1517, 2007.

[9] M. Komachi, T. Kudo, M. Shimbo, and Y. Matsumoto. Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1019, 2008.

[10] Z.-Y. Niu, D.-H. Ji, and C. L. Tan. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 395–402, 2005.

[11] T. P. Pham, H. T. Ng, and W. S. Lee. Word Sense Disambiguation with Semi-Supervised Learning. In *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI 2005)*, pages 1093–1098, 2005.

[12] H. Schütze and M. Walsh. A graph-theoretic model of lexical syntactic acquisition. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 916–925, 2008.

[13] A. J. Smola and R. I. Kondor. Kernels and Regularization of Graphs. In *Proceedings of the 16th Annual Conference on Learning Theory*, pages 144–158, 2003.

[14] P. P. Talukdar, J. Reisinger, M. Pasca, D. Ravichandran, R. Bhagat, and F. Pereira. Weakly-supervised acquisition of labeled class instances using graph random walks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 581–589, 2008.

[15] M. Thelen and E. Riloff. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 214–221, 2002.

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