Mining New Translation Lexicons from Wikipedia

Jong-Hoon Oh, Daisuke Kawahara, Kiyotaka Uchimoto, and Hitoshi Isahara
Computational Linguistics Group,
National Institute of Information and Communications Technology (NICT)
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{rovellia, dk, uchimoto, isahara}@nict.go.jp

Abstract

In this paper, we propose a method of mining new translation lexicons from Wikipedia. To do this, first a network is constructed with nodes representing Wikipedia articles and links representing the dependencies between the articles. Then we learn a model of translation lexicons by investigating co-occurrence patterns of the existing translation lexicons in a Wikipedia network. Finally, new translation lexicons can be found by estimating the cross-lingual similarities between the nodes in different languages. Experiments show that our method effectively finds new translation lexicons.

1 Introduction

We use Wikipedia as the resource of translation lexicons. Wikipedia is an open and multilingual encyclopedia that provides a huge number of articles describing various concepts, from general to domain-specific[1]. Using Wikipedia as the translation lexicon resource has several advantages over manually aligned bilingual corpora and Web corpora. First, Wikipedia is a multilingual language resource updated every day, and there are generally several versions of the same article written in different languages that are linked to each other. It is thus easy to find up-to-date versions of the same article written in different languages. Second, Wikipedia is well-structured: the articles have a rich structure and hypertext links to related Wikipedia articles. It is thus easy to extract translation knowledge from Wikipedia articles.

If the different language versions of a Wikipedia article are linked to each other, the translation lexicons can be acquired by using these links. Unfortunately, few Wikipedia articles have such a link. For example, only 5% of the English Wikipedia articles are linked to the corresponding Japanese article[2]. To overcome this problem, we have developed a method of mining new translation lexicons from English and Japanese versions of Wikipedia. We first create a network of the English and Japanese Wikipedia articles that represents the dependency between the articles. The network is composed of nodes representing the articles and links representing their dependencies. We learn new translation lexicons by using existing bilingual translation lexicons and their corresponding nodes in the network as the basis of the new translation lexicons. We investigate the co-occurrence patterns of these existing lexicons in the network to learn a translation lexicon model. We then estimated cross-lingual similarities between Wikipedia articles (in different languages) based on the learned model to find new translation lexicons. To our knowledge, ours is the first attempt to learn a translation lexicon model from Wikipedia.

[1] As of November 2007, Wikipedia had approximately 9.1 million articles in 252 languages. English, German, French, Polish, and Japanese versions accounted for more than 4.2 million articles.

[2] As of May 2007, 0.17 million of the 3.17 million English articles have their Japanese version; while 0.17 million of the 0.49 million Japanese articles have their English version in Wikipedia.

2 Wikipedia

We used the English and Japanese ASCII text versions of Wikipedia[3]. The English version contains about 3.17 million articles and 1.96 million redirection pages while the Japanese version contains about 0.49 million articles and 0.2 million redirection pages. Table 1 lists terms and their definitions related to the Wikipedia structure.

Term	Definition
EN($\mathcal{A}$)	*Entity name* of $\mathcal{A}$
VN($\mathcal{A}$)	*Variants* of EN($\mathcal{A}$)
Text($\mathcal{A}$)	*Text body* of $\mathcal{A}$
Term($\mathcal{A}$)	*Terms* in Text($\mathcal{A}$)
Context($\mathcal{A}$)	*Context* of EN($\mathcal{A}$) in Wikipedia
OLink($\mathcal{A}$)	A set of EN($\mathcal{B}$), where $\mathcal{A} \rightarrow \mathcal{B}$
ILink($\mathcal{A}$)	A set of EN($\mathcal{B}$), where $\mathcal{A} \leftarrow \mathcal{B}$
BE($\mathcal{A}$,$\mathcal{B}$)	(EN($\mathcal{A}$), EN($\mathcal{B}$)), where $\mathcal{A} \bowtie \mathcal{B}$
BA($\mathcal{A}$,$\mathcal{B}$)	(Text($\mathcal{A}$), Text($\mathcal{B}$)), where $\mathcal{A} \bowtie \mathcal{B}$
WikiBLs	A set of BE($\mathcal{A}$,$\mathcal{B}$) in Wikipedia
WikiBAs	A set of BA($\mathcal{A}$,$\mathcal{B}$) in Wikipedia
WikiPSs	Parallel sentences automatically extracted from WikiBAs

Table 1: Terms and their definition used in this paper. $\mathcal{A}$ and $\mathcal{B}$ represent Wikipedia articles.

We morphologically analyzed Text($\mathcal{A}$) in order to get Context($\mathcal{A}$) and Term($\mathcal{A}$)[4]. We extracted all nouns and entity names in Text($\mathcal{A}$) as Term($\mathcal{A}$) and those collocated with EN($\mathcal{A}$) within context size k in all Wikipedia articles as Context($\mathcal{A}$). References to other Wikipedia articles in Text($\mathcal{A}$) are marked using "[[" and "]]," which correspond

[3] We used the English and Japanese versions of Wikipedia as of May 2007, which are available as database dumps at `http://download.wikipedia.org`.

[4] We used TreeTagger [1] and JUMAN [2] as the English and Japanese morphological analyzers, respectively

to a hyperlink in the online version of Wikipedia. From the viewpoint of language versions of Wikipedia, we classify the references or hyperlinks to other Wikipedia articles into "**IntraWiki links**" (links within a single language version: notated as $\mathcal{A} \rightarrow \mathcal{B}$ meaning $\mathcal{A}$ has references to $\mathcal{B}$) and "**InterWiki links**" (links across different language versions: notated as $\mathcal{A} \bowtie \mathcal{B}$ meaning $\mathcal{A}$ and $\mathcal{B}$ are linked to each other via an InterWiki link). We classify IntraWiki links into **outgoing links** and **incoming links** on the basis of their direction. The InterWiki links are particularly useful because they enable us to extract translation lexicons and to find bilingual corpora. By analyzing the InterWiki links, we can construct **WikiBLs** (Wikipedia bilingual lexicons), and **WikiBAs** (Wikipedia bilingual article texts)[5]. Parallel sentences form the basis of many cross-lingual research in natural language processing. Because WikiBAs are article-aligned rather than sentence-aligned, we extract parallel sentences from the WikiBAs by using a sentence alignment algorithm [3]. We call the parallel sentences extracted from WikiBAs **WikiPSs**.

Using the information given in Table 1, we can construct English and Japanese Wikipedia networks. In the networks, we represent each Wikipedia article ($\mathcal{A}$) as a node and connect the nodes with IntraWiki and Inter links. Henceforth, we use **EWikiNet**, **JWikiNet**, and **WikiNet** to represent the English Wikipedia network, the Japanese Wikipedia network, and the network comprising EWikiNet and JWikiNet. The WikiNet as well as the WikiBLs, WikiPSs, and WikiBAs are used as basic knowledge to learn a translation lexicon model.

3 Mining Translation Lexicons

Given an English term corresponding to an EWikiNet node, our task is to find the JWikiNet node that is the most cross-lingually similar to the EWikiNet node. Thus, we can regard our task as learning a discriminative model that can serve as a cross-lingual similarity function between EWikiNet and JWikiNet nodes. Let us assume that we already have $J_i = \{j_{i1}, j_{i2}, \cdots, j_{iN}\}$ — a set of candidate JWikiNet nodes for each EWikiNet node e_i —, and that $j_{ik} \in J_i$ is a translation of EN(e_i). For any training example of e_i, we can represent positive training samples as j_{ik} and negative training samples as $j_{il} \in J_i$, where $l \neq k$. We can then learn the discriminative model that best fits these training examples. Next, we use the learned model to find new translation lexicons by identifying new InterWiki relationships between EWikiNet and JWikiNet nodes.

3.1 Candidate Extraction

We find $J_i = \{j_{i1}, j_{i2}, \cdots, j_{iN}\}$ from e_i by using cross-lingual content similarities (Term($\mathcal{A}$) and Context($\mathcal{A}$)) and link similarities (OLink($\mathcal{A}$) and ILink($\mathcal{A}$)) between EWikiNet and JWikiNet nodes. Let $X(\mathcal{A})$ be a set of {Term($\mathcal{A}$), Context($\mathcal{A}$), OLink($\mathcal{A}$), ILink($\mathcal{A}$)}. We can regard $x(j_k) \in X(j_k)$ as a Japanese document and $x(e_i) \in X(e_i)$ as an English query. Thus, our candidate extraction problem can be converted into a cross language information retrieval problem: how to retrieve the most relevant JWikiNet nodes for a given EWikiNet node from the different viewpoints of $x(\mathcal{A}) \in X(\mathcal{A})$. Let $D_{x(j_k)}$ be a Japanese document corresponding to $x(j_k) \in X(j_k)$ of JWikiNet node j_k, $Q_{x(e_i)}$ be an English query corresponding $x(e_i) \in X(e_i)$ of EWikiNet node e_i, and $TQ_{x(e_i)}$ be a Japanese version of $Q_{x(e_i)}$ translated using a translation dictionary. First, we construct indexes for $D_{x(j_k)}$ of every JWikiNet node j_k. For a given EWikiNet node e_i, four English queries, $Q_{x(e_i)}$, are generated and translated into Japanese ($TQ_{x(e_i)}$) by using the WikiBLs. Then, $TQ_{x(e_i)}$ retrieves the top n JWikiNet nodes by using the Okapi BM-25 weighting scheme, which is given in Eq. (1) [4]. We can get N candidate JWikiNet nodes for e_i by combining the top n JWikiNet nodes ($J_i(Term)$, $J_i(Context)$, $J_i(OLink)$, and $J_i(ILink)$) retrieved by each of $TQ_{Term(e_i)}$, $TQ_{Context(e_i)}$, $TQ_{OLink(e_i)}$, and $TQ_{ILink(e_i)}$, where $n \leq N \leq 4 \times n$. For indexing and retrieving documents, we used "The Lemur Toolkit"[6] with its default Okapi BM-25 parameters: $k_1 = 1.2$, $b = 0.75$, and $k_3 = 7$.

$$Sim(TQ_{x(e_i)}, D_{x(j_k)}) = \sum w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \quad (1)$$

where $w^{(1)} = log\frac{N-DF(t)+0.5}{DF(t)+0.5}$ is the weight of term t, K is $k_1 \times ((1-b) + b \times \frac{|D|}{AVG(|D|)})$, and qtf and tf are the frequencies of term t occurring in query $TQ_{x(e_i)}$ and document $D_{x(j_k)}$, respectively. $|D|$ is the length of document D in words, $AVG(|D|)$ is the average document length in words, and $DF(t)$ is the number of documents containing term t.

3.2 Features for Cross-Lingual Similarity

3.2.1 Features based on $Sim(TQ_{x(e_i)}, D_{x(j_k)})$

We normalize $Sim(TQ_{x(e_i)}, D_{x(j_{ik})})$ from the viewpoints of $TQ_{x(e_i)}$ and $D_{x(j_{ik})}$ as described in Eqs. (2) and (3), where E represents a set of EWikiNet nodes.

$$SQ_x(e_i, j_{ik}) = \frac{Sim(TQ_{x(e_i)}, D_{x(j_{ik})})}{max_{j_{il} \in J_i}(Sim(TQ_{x(e_i)}, D_{x(j_{il})}))} \quad (2)$$

$$SD_x(e_i, j_{ik}) = \frac{Sim(TQ_{x(e_i)}, D_{x(j_{ik})})}{max_{e_j \in E}(Sim(TQ_{x(e_j)}, D_{x(j_{ik})}))} \quad (3)$$

Let $Z_{out}(j_{ik})$ be a set of JWikiNet nodes j_{ir}, where $j_{ir} \in J_i$ and $EN(j_{ir}) \in OLink(j_{ik})$, and let $Z_{in}(j_{ik})$ be a set of JWikiNet nodes j_{iq}, where $j_{iq} \in J_i$ and $EN(j_{iq}) \in ILink(j_{ik})$. We take into account content similarities between $j_{ir} \in Z_{out}(j_{ik})$ and e_i and between $j_{iq} \in Z_{in}(j_{ik})$ and e_i as features of j_{ik} as described in Eq. (4), where $y(\mathcal{A}) = \{Term(\mathcal{A}), Context(\mathcal{A})\}$ and $j_{il} \in J_i$. In other words, we consider neighborhoods of j_{ik} connected by IntraWiki links.

We call SQ_x and SD_x, where x=Term or Context, **content similarity (CS) features**. SQ_x, SD_x, OSQ_y, OSD_y, ISQ_y, and ISD_y, where x=OLink or ILink, are called **link similarity (LS) features**. We use 4 CS features and 12 LS features in Eqs. (2)–(4) for learning a discriminative model (Section 3.3).

[5] We could use about 0.17 million Wikipedia articles for WikiBLs and WikiBAs.

[6] Available at `http://www.lemurproject.org`

$$OSQ_y(e_i, j_{ik}) = \frac{\sum_{z \in Z_{out}(j_{ik})} SQ_y(e_i, z)}{max_{j_{il}}(\sum_{z \in Z_{out}(j_{il})} SQ_y(e_i, z))} \quad (4)$$

$$OSD_y(e_i, j_{ik}) = \frac{\sum_{z \in Z_{out}(j_{ik})} SD_y(e_i, z)}{max_{j_{il}}(\sum_{z \in Z_{out}(j_{il})} SD_y(e_i, z))}$$

$$ISQ_y(e_i, j_{ik}) = \frac{\sum_{z \in Z_{in}(j_{ik})} SQ_y(e_i, z)}{max_{j_{il}}(\sum_{z \in Z_{in}(j_{il})} SQ_y(e_i, z))}$$

$$ISD_y(e_i, j_{ik}) = \frac{\sum_{z \in Z_{in}(j_{ik})} SD_y(e_i, z)}{max_{j_{il}}(\sum_{z \in Z_{in}(j_{il})} SD_y(e_i, z))}$$

3.2.2 Features based on Translation Models

We construct translation models by using WikiBLs and WikiPSs and name them in accordance with the resources used to construct them — TM_D (dictionary) and TM_S (parallel sentences), respectively. The translation models was designed to show cross-lingual similarities between EN(e_i) and EN(j_{ik}). To do so, we calculate translation probabilities between EN(e_i) and EN(j_{ik}) based on their term constituents. GIZA++ [5] was used to acquire correspondence between term constituents in WikiBLs and WikiPSs. Because GIZA++ produces different alignment results and translation models depending on the source and target languages, we construct two translation models for each resource ($E \rightarrow J$ and $J \rightarrow E$). We used IBM Model 3 to calculate $P_{TM_D}(y|x)$, $P_{TM_D}(x|y)$, $P_{TM_S}(y|x)$, and $P_{TM_S}(x|y)$, where $x = EN(e_i)$ and $y = EN(j_{ik})$ [6, 5]. In this paper, we call $P_{TM_D}(y|x)$, $P_{TM_D}(x|y)$, $P_{TM_S}(y|x)$, and $P_{TM_S}(x|y)$ **entity translation (ET) features**.

3.3 Cross-Lingual Similarity

A support vector machine (SVM) regression algorithm is used to learn a discriminative model of cross-lingual similarity [7]. Given a vector of feature functions f between e_i and j_{ik}, $Score(e_i, j_{ik})$ is represented as an inner product between a weight vector and the feature vector in a kernel space:

$$Score(e_i, j_{ik}) = w \cdot \phi(f(e_i, j_{ik})) \quad (5)$$

where, ϕ is a mapping from the input feature space onto the kernel space, and w is the weight vector in the kernel space. Once w is learned by the SVM regression training, we can use Eq. (5) to estimate the cross-lingual similarity between e_i and every $j_{ik} \in J_i$. Finally, we can find the JWikiNet node most similar to e_i.

4 Evaluation

We built training and development sets by randomly selecting 1,688 entries of WikiBLs, respectively. A test set was built by selecting 1,791 terms from the JST dictionary[7], where the English and Japanese terms both correspond to WikiNet nodes but have no InterWiki link between them.

[7]The JST (Japan Science and Technology Agency) dictionary is a Japanese-English technical dictionary, which contains about 643,000 entries in several scientific domains including physics, biology, chemistry, medical and so on.

Note that, in our experiments, we exclude the entries in the training, development, and test sets from WikiBLs, which were used as a translation dictionary in candidate extraction and training data for translation model TM_D. The training set was used for SVM regression training to get a discriminative model of cross-lingual similarity. The development set was used to optimize several parameters. The test set was used to see how well our proposed method learns new translation lexicons from Wikipedia. The evaluation metric was precision — the number of translation lexicons found by our system divided by the number of translation lexicons in the gold standard.

4.1 Results

System	TOP1	TOP3	TOP5	TOP10
CS	35.8	52.2	59.8	69.1
LS	26.6	45.6	54.3	64.4
ET	59.6	67.7	68.8	69.4
CS+LS	42.4	61.8	68.5	77.6
CS+ET	73.6	87.2	90.6	93.6
LS+ET	70.7	84.5	89.2	92.3
ALL	**75.9**	**90.2**	**93.2**	**95.8**
UB	97.1			

Table 2: Results with different feature combinations (%)

We evaluated the performance of our method with different combinations of features — **content similarity (CS)**, **link similarity (LS)**, and **entity translation (ET)**. Table 2 summarizes the results; UB is a system that always outputs a correct Japanese counterpart of given English term if it is extracted as one of candidates in the candidate extraction step, and All is our proposed method, which uses all features. As more features were used, performance consistently improved. However, each feature made a different contribution to performance. The contribution of each feature to the performance of All can be estimated by investigating the performance gap between CS+LS and ALL, between CS+ET and ALL, and so on. We found that ET made the largest contribution. Because Wikipedia provides plenty of bilingual resources including parallel sentences (WikiPSs) and a translation dictionary (WikiBLs), we could construct translation models for entity names, which is effective in learning a discriminative model of cross-lingual similarity. We can regard UB as the upper bound on the performance of our method or the performance of candidate extraction. Because we used the CS and LS features for candidate extraction, the performance of UB (97.1%) shows that CS and LS for candidate extraction played their roles effectively. Moreover, both were good features for learning translation lexicons.

5 Related work

Translation lexicons have been extracted on the basis of co-occurrence patterns of words across languages in parallel or comparable corpora [8, 9, 10]. Because our con-

tent similarity (CS) feature uses information similar to that used in previous work — the co-occurrence patterns of words in Term($\mathcal{A}$) and Context($\mathcal{A}$) across languages —, we can indirectly compare previous approaches with ours by comparing CS with All in Table 2. In previous approaches, each sentence (or document) in the parallel or comparable corpora is usually regarded as an independent and unstructured unit. In contrast, Wikipedia has a rich structure with related articles linked to each other. Our method uses the rich structure of Wikipedia in both the source and target languages to find new translation lexicons.

Many researchers have used Wikipedia as a knowledge resource in natural language processing such as question answering, word sense disambiguation, named entity recognition, and ontology building [11, 12, 13, 14]. However, most of their work has focused on monolingual aspects (especially English). While the method proposed by Adafre and de Rijke [11] for finding parallel sentences uses WikiBAs and WikiBLs, it lacks the ability to use various types of information encoded in the rich structure of Wikipedia that are useful for multilingual knowledge acquisition.

6 Conclusion

We have described a method of mining new translation lexicons from Wikipedia. First, we construct a network of nodes representing Wikipedia articles and links representing the dependencies between the articles. Second, we acquire existing translation lexicons explicitly expressed by InterWiki links in the network. Third, we learn a discriminative model for cross-lingual similarity by investigating the co-occurrence patterns of existing bilingual translation lexicons in the network. Finally, we use the learned model to find new translation lexicons in the network. Experiments showed that our method finds new translation lexicons with high precision.

Because Wikipedia is a multilingual resource and we learn a translation lexicon model without relying other language resources except Wikipedia, our method can be easily extensible to other language pairs.

References

[1] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[2] Sadao Kurohashi and Daisuke Kawahara. Nihongo keitaikaiseki sisutemu JUMAN [Japanese morphological analysis system JUMAN] version 5.1, 2005.

[3] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *ACL '03*, pages 72–79, 2003.

[4] S.E. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life: Okapi at trec. *Information Processing and Management*, 36(3):95–108, 2000.

[5] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[6] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[7] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, 2002.

[8] Pascale Fung. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *AMTA '98*, pages 1–17, 1998.

[9] Li Shao and Hwee Tou Ng. Mining new word translations from comparable corpora. In *COLING '04*, pages 618–624, 2004.

[10] D. Wu and X. Xia. Large-scale automatic extraction of an English-Chinese translation lexicon. *Machine Translation*, 9(3-4):285–313, 1995.

[11] Sisay Fissaha Adafre and M. de Rijke. Finding similar sentences across multiple languages in Wikipedia. In *Proc. of EACL '06 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources*, 2006.

[12] Jun'ichi Kazama and Kentaro Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In *EMNLP-CoNLL '07*, pages 698–707, 2007.

[13] Rada Mihalcea. Using Wikipedia for automatic word sense disambiguation. In *NAACL-HLT '07; Proceedings of the Main Conference*, pages 196–203. Association for Computational Linguistics, 2007.

[14] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *WWW 2007*, New York, NY, USA, 2007. ACM Press.