Bilingual Co-Training for Monolingual Hyponymy-Relation Acquisition

Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa
Language Infrastructure Group, MASTAR Project,
National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai Seika-cho, Soraku-gun, Kyoto 619-0289 Japan
{rovellia,uchimoto,torisawa}@nict.go.jp

Abstract

This paper proposes a novel framework called *bilingual co-training* for a large-scale, accurate acquisition method for *monolingual* semantic knowledge. In this framework, we combine the independent processes of monolingual semantic-knowledge acquisition for two languages using bilingual resources to boost performance. We apply this framework to large-scale hyponymy-relation acquisition from Wikipedia. Experimental results show that our approach improved the F-measure by 3.6–10.3%. We also show that bilingual co-training enables us to build classifiers for two languages in tandem with the same combined amount of data as required for training a single classifier in isolation while achieving superior performance.

1 Motivation

Acquiring and accumulating semantic knowledge are crucial steps for developing high-level NLP applications such as question answering, although it remains difficult to acquire a large amount of highly accurate semantic knowledge. This paper proposes a novel framework for a large-scale, accurate acquisition method for monolingual semantic knowledge, especially for semantic relations between nominals such as hyponymy and meronymy. We call the framework *bilingual co-training*.

The acquisition of semantic relations between nominals can be seen as a classification task of semantic relations – to determine whether two nominals hold a particular semantic relation (Girju et al., 2007). Supervised learning methods, which have often been applied to this classification task, have shown promising results. In those methods, however, a large amount of training data is usually required to obtain high performance, and the high costs of preparing training data have always been a bottleneck.

Our research on bilingual co-training sprang from a very simple idea: perhaps training data in a language can be enlarged without much cost if we translate training data in another language and add the translation to the training data in the original language. We also noticed that it may be possible to further enlarge the training data by translating the reliable part of the classification results in another language. Since the learning settings (feature sets, feature values, training data, corpora, and so on) are usually different in two languages, the reliable part in one language may be overlapped by an unreliable part in another language. Adding the translated part of the classification results to the training data will improve the classification results in the unreliable part. This process can also be repeated by swapping the languages, as illustrated in Figure 1. Actually, this is nothing other than a bilingual version of co-training (Blum and Mitchell, 1998).

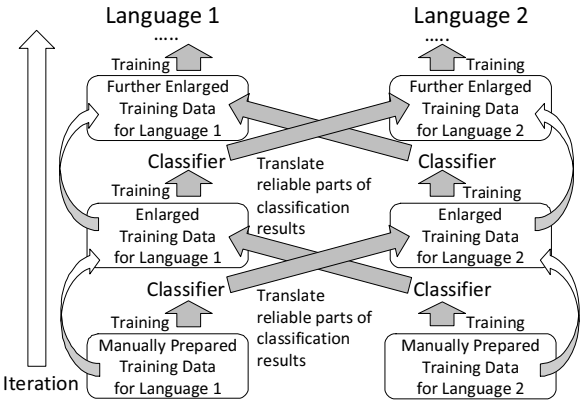


Figure 1: Concept of *bilingual co-training*

Let us show an example in our current task: hyponymy-relation acquisition from Wikipedia. Our original approach for this task was super-

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 432–440,
Suntec, Singapore, 2-7 August 2009.

vised learning based on the approach proposed by Sumida et al. (2008), which was only applied for Japanese and achieved around 80% in F-measure. In their approach, a common substring in a hypernym and a hyponym is assumed to be one strong clue for recognizing that the two words constitute a hyponymy relation. For example, recognizing a proper hyponymy relation between two Japanese words, 酵素 (***kouso*** meaning *enzyme*) and 加水分解酵素 (*kasuibunkai****kouso*** meaning *hydrolase*), is relatively easy because they share a common suffix: **kouso**. On the other hand, judging whether their English translations (*enzyme* and *hydrolase*) have a hyponymy relation is probably more difficult since they do not share any substrings. A classifier for Japanese will regard the hyponymy relation as valid with high confidence, while a classifier for English may not be so positive. In this case, we can compensate for the weak part of the English classifier by adding the English translation of the Japanese hyponymy relation, which was recognized with high confidence, to the English training data.

In addition, if we repeat this process by swapping English and Japanese, further improvement may be possible. Furthermore, the reliable parts that are automatically produced by a classifier can be larger than manually tailored training data. If this is the case, the effect of adding the translation to the training data can be quite large, and the same level of effect may not be achievable by a reasonable amount of labor for preparing the training data. This is the whole idea.

Through a series of experiments, this paper shows that the above idea is valid at least for one task: large-scale monolingual hyponymy-relation acquisition from English and Japanese Wikipedia. Experimental results showed that our method based on bilingual co-training improved the performance of monolingual hyponymy-relation acquisition about 3.6–10.3% in the F-measure. Bilingual co-training also enables us to build classifiers for two languages in tandem with the same combined amount of data as would be required for training a single classifier in isolation while achieving superior performance.

People probably expect that a key factor in the success of this bilingual co-training is how to translate the training data. We actually did translation by a simple look-up procedure in the existing translation dictionaries without any machine translation systems or disambiguation processes. Despite this simple approach, we obtained consistent improvement in our task using various translation dictionaries.

This paper is organized as follows. Section 2 presents bilingual co-training, and Section 3 precisely describes our system. Section 4 describes our experiments and presents results. Section 5 discusses related work. Conclusions are drawn and future work is mentioned in Section 6.

2 Bilingual Co-Training

Let S and T be two different languages, and let CL be a set of class labels to be obtained as a result of learning/classification. To simplify the discussion, we assume that a class label is binary; i.e., the classification results are "yes" or "no." Thus, $CL = \{\textit{yes}, \textit{no}\}$. Also, we denote the set of all nonnegative real numbers by R^+.

Assume $X = X_S \cup X_T$ is a set of instances in languages S and T to be classified. In the context of a hyponymy-relation acquisition task, the instances are pairs of nominals. Then we assume that classifier c assigns class label cl in CL and confidence value r for assigning the label, i.e., $c(x) = (x, cl, r)$, where $x \in X$, $cl \in CL$, and $r \in R^+$. Note that we used support vector machines (SVMs) in our experiments and (the absolute value of) the distance between a sample and the hyperplane determined by the SVMs was used as confidence value r. The training data are denoted by $L \subset X \times CL$, and we denote the learning by function $LEARN$; if classifier c is trained by training data L, then $c = LEARN(L)$. Particularly, we denote the training sets for S and T that are manually prepared by L_S and L_T, respectively. Also, bilingual instance dictionary D_{BI} is defined as the translation pairs of instances in X_S and X_T. Thus, $D_{BI} = \{(s,t)\} \subset X_S \times X_T$. In the case of hyponymy-relation acquisition in English and Japanese, $(s,t) \in D_{BI}$ could be (s=(*enzyme*, *hydrolase*), t=(酵素 (meaning *enzyme*), 加水分解酵素 (meaning *hydrolase*))).

Our bilingual co-training is given in Figure 2. In the initial stage, c_S^0 and c_T^0 are learned with manually labeled instances L_S and L_T (lines 2–5). Then c_S^i and c_T^i are applied to classify instances in X_S and X_T (lines 6–7). Denote CR_S^i as a set of the classification results of c_S^i on instances X_S that is not in L_S^i and is registered in D_{BI}. Lines 10–18 describe a way of selecting from CR_S^i newly la-

1: $i = 0$
2: $L_S^0 = L_S$; $L_T^0 = L_T$
3: **repeat**
4: $c_S^i := LEARN(L_S^i)$
5: $c_T^i := LEARN(L_T^i)$
6: $CR_S^i := \{c_S^i(x_S) | x_S \in X_S,$
$\forall cl\ (x_S, cl) \notin L_S^i, \exists x_T\ (x_S, x_T) \in D_{BI}\}$
7: $CR_T^i := \{c_T^i(x_T) | x_T \in X_T,$
$\forall cl\ (x_T, cl) \notin L_T^i, \exists x_S\ (x_S, x_T) \in D_{BI}\}$
8: $L_S^{(i+1)} := L_S^i$
9: $L_T^{(i+1)} := L_T^i$
10: **for** each $(x_S, cl_S, r_S) \in TopN(CR_S^i)$ **do**
11: **for** each x_T such that $(x_S, x_T) \in D_{BI}$ and $(x_T, cl_T, r_T) \in CR_T^i$ **do**
12: **if** $r_S > \theta$ **then**
13: **if** $r_T < \theta$ or $cl_S = cl_T$ **then**
14: $L_T^{(i+1)} := L_T^{(i+1)} \cup \{(x_T, cl_S)\}$
15: **end if**
16: **end if**
17: **end for**
18: **end for**
19: **for** each $(x_T, cl_T, r_T) \in TopN(CR_T^i)$ **do**
20: **for** each x_S such that $(x_S, x_T) \in D_{BI}$ and $(x_S, cl_S, r_S) \in CR_S^i$ **do**
21: **if** $r_T > \theta$ **then**
22: **if** $r_S < \theta$ or $cl_S = cl_T$ **then**
23: $L_S^{(i+1)} := L_S^{(i+1)} \cup \{(x_S, cl_T)\}$
24: **end if**
25: **end if**
26: **end for**
27: **end for**
28: $i = i + 1$
29: **until** a fixed number of iterations is reached

Figure 2: Pseudo-code of *bilingual co-training*

beled instances to be added to a new training set in T. $TopN(CR_S^i)$ is a set of $c_S^i(x)$, whose r_S is top-N highest in CR_S^i. (In our experiments, $N = 900$.) During the selection, c_S^i acts as a teacher and c_T^i as a student. The teacher instructs his student in the class label of x_T, which is actually a translation of x_S by bilingual instance dictionary D_{BI}, through cl_S only if he can do it with a certain level of confidence, say $r_S > \theta$, and if one of two other condition meets ($r_T < \theta$ or $cl_S = cl_T$). $cl_S = cl_T$ is a condition to avoid problems, especially when the student also has a certain level of confidence in his opinion on a class label but disagrees with the teacher: $r_T > \theta$ and $cl_S \neq cl_T$. In that case, the teacher does nothing and ignores the instance. Condition $r_T < \theta$ enables the teacher to instruct his student in the class label of x_T in spite of their disagreement in a class label. If every condition is satisfied, (x_T, cl_S) is added to existing labeled instances $L_T^{(i+1)}$. The roles are reversed in lines 19–27 so that c_T^i becomes a teacher and c_S^i a student.

Similar to co-training (Blum and Mitchell, 1998), one classifier seeks another's opinion to select new labeled instances. One main difference between co-training and bilingual co-training is the space of instances: co-training is based on different features of the same instances, and bilingual co-training is based on different spaces of instances divided by languages. Since some of the instances in different spaces are connected by a bilingual instance dictionary, they seem to be in the same space. Another big difference lies in the role of the two classifiers. The two classifiers in co-training work on the same task, but those in bilingual co-training do *the same type of* task rather than the same task.

3 Acquisition of Hyponymy Relations from Wikipedia

Our system, which acquires hyponymy relations from Wikipedia based on bilingual co-training, is described in Figure 3. The following three main parts are described in this section: candidate extraction, hyponymy-relation classification, and bilingual instance dictionary construction.

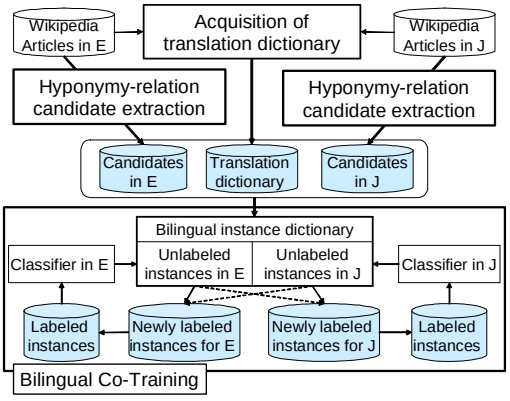
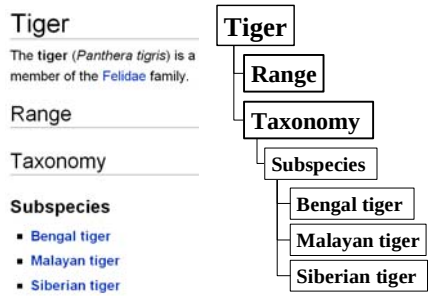


Figure 3: System architecture

3.1 Candidate Extraction

We follow Sumida et al. (2008) to extract hyponymy-relation candidates from English and Japanese Wikipedia. A layout structure is chosen



(a) Layout structure of article TIGER (b) Tree structure of Figure 4(a)

Figure 4: Wikipedia article and its layout structure

as a source of hyponymy relations because it can provide a huge amount of them (Sumida et al., 2008; Sumida and Torisawa, 2008)[1], and recognition of the layout structure is easy regardless of languages. Every English and Japanese Wikipedia article was transformed into a tree structure like Figure 4, where layout items *title*, *(sub)section headings*, and *list items* in an article were used as nodes in a tree structure. Sumida et al. (2008) found that some pairs consisting of a node and one of its descendants constituted a proper hyponymy relation (e.g., (TIGER, SIBERIAN TIGER)), and this could be a knowledge source of hyponymy relation acquisition. A hyponymy-relation candidate is then extracted from the tree structure by regarding a node as a hypernym candidate and all its subordinate nodes as hyponym candidates of the hypernym candidate (e.g., (TIGER, TAXONOMY) and (TIGER, SIBERIAN TIGER) from Figure 4). 39 M English hyponymy-relation candidates and 10 M Japanese ones were extracted from Wikipedia. These candidates are classified into proper hyponymy relations and others by using the classifiers described below.

3.2 Hyponymy-Relation Classification

We use SVMs (Vapnik, 1995) as classifiers for the classification of the hyponymy relations on the hyponymy-relation candidates. Let **hyper** be a hypernym candidate, **hypo** be a **hyper**'s hyponym candidate, and (**hyper, hypo**) be a hyponymy-relation candidate. The lexical, structure-based, and infobox-based features of (**hyper, hypo**) in Table 1 are used for building English and Japanese classifiers. Note that SF_3–SF_5 and IF were not used in Sumida et al. (2008) but LF_1–LF_5 and SF_1–SF_2 are the same as their feature set.

Let us provide an overview of the feature sets used in Sumida et al. (2008). See Sumida et al. (2008) for more details. Lexical features LF_1–LF_5 are used to recognize the lexical evidence encoded in **hyper** and **hypo** for hyponymy relations. For example, (**hyper,hypo**) is often a proper hyponymy relation if **hyper** and **hypo** share the same head morpheme or word. In LF_1 and LF_2, such information is provided along with the words/morphemes and the parts of speech of **hyper** and **hypo**, which can be multi-word/morpheme nouns. TagChunk (Daumé III et al., 2005) for English and MeCab (MeCab, 2008) for Japanese were used to provide the lexical features. Several simple lexical patterns[2] were also applied to hyponymy-relation candidates. For example, "List of artists" is converted into "artists" by lexical pattern "list of X." Hyponymy-relation candidates whose hypernym candidate matches such a lexical pattern are likely to be valid (e.g., (List of artists, Leonardo da Vinci)). We use LF_4 for dealing with these cases. If a typical or frequently used section heading in a Wikipedia article, such as "History" or "References," is used as a hyponym candidate in a hyponymy-relation candidate, the hyponymy-relation candidate is usually not a hyponymy relation. LF_5 is used to recognize these hyponymy-relation candidates.

Structure-based features are related to the tree structure of Wikipedia articles from which hyponymy-relation candidate (**hyper,hypo**) is extracted. SF_1 provides the distance between **hyper** and **hypo** in the tree structure. SF_2 represents the type of layout items from which **hyper** and **hypo** are originated. These are the feature sets used in Sumida et al. (2008).

We also added some new items to the above feature sets. SF_3 represents the types of tree nodes including root, leaf, and others. For example, (**hyper,hypo**) is seldom a hyponymy relation if **hyper** is from a root node (or title) and **hypo** is from a **hyper**'s child node (or section headings). SF_4 and SF_5 represent the structural contexts of **hyper** and **hypo** in a tree structure. They can provide evidence related to similar hyponymy-relation candidates in the structural contexts.

An infobox-based feature, IF, is based on a

[1] Sumida et al. (2008) reported that they obtained 171 K, 420 K, and 1.48 M hyponymy relations from a definition sentence, a category system, and a layout structure in Japanese Wikipedia, respectively.

[2] We used the same Japanese lexical patterns in Sumida et al. (2008) to build English lexical patterns with them.

Type	Description	Example
LF_1	Morphemes/words	hyper: tiger*, hypo: Siberian, hypo: tiger*
LF_2	POS of morphemes/words	hyper: NN*, hypo: NP, hypo: NN*
LF_3	**hyper** and **hypo**, themselves	hyper: Tiger, hypo: Siberian tiger
LF_4	Used lexical patterns	hyper: "List of X", hypo: "Notable X"
LF_5	Typical section headings	hyper: History, hypo: Reference
SF_1	Distance between **hyper** and **hypo**	3
SF_2	Type of layout items	hyper: title, hypo: bulleted list
SF_3	Type of tree nodes	hyper: root node, hypo: leaf node
SF_4	LF_1 and LF_3 of **hypo**'s parent node	LF_3:Subspecies
SF_5	LF_1 and LF_3 of **hyper**'s child node	LF_3: Taxonomy
IF	Semantic properties of **hyper** and **hypo**	hyper: (taxobox,species), hypo: (taxobox,name)

Table 1: Feature type and its value. $*$ in LF_1 and LF_2 represent the head morpheme/word and its POS. Except those in LF_4 and LF_5, examples are derived from (TIGER, SIBERIAN TIGER) in Figure 4.

Wikipedia infobox, a special kind of template, that describes a tabular summary of an article subject expressed by attribute-value pairs. An attribute type coupled with the infobox name to which it belongs provides the semantic properties of its value that enable us to easily understand what the attribute value means (Auer and Lehmann, 2007; Wu and Weld, 2007). For example, infobox template *City Japan* in Wikipedia article *Kyoto* contains several attribute-value pairs such as "Mayor=Daisaku Kadokawa" as *attribute=its value*. What *Daisaku Kadokawa*, the attribute value of *mayor* in the example, represents is hard to understand alone if we lack knowledge, but its attribute type, *mayor*, gives a clue–*Daisaku Kadokawa* is a *mayor* related to *Kyoto*. These semantic properties enable us to discover semantic evidence for hyponymy relations. We extract triples (*infobox name, attribute type, attribute value*) from the Wikipedia infoboxes and encode such information related to **hyper** and **hypo** in our feature set IF.[3]

3.3 Bilingual Instance Dictionary Construction

Multilingual versions of Wikipedia articles are connected by cross-language links and usually have titles that are bilinguals of each other (Erdmann et al., 2008). English and Japanese articles connected by a cross-language link are extracted from Wikipedia, and their titles are regarded as translation pairs[4]. The translation pairs between English and Japanese terms are used for building bilingual instance dictionary D_{BI} for hyponymy-relation acquisition, where D_{BI} is composed of translation pairs between English and Japanese hyponymy-relation candidates[5].

4 Experiments

We used the MAY 2008 version of English Wikipedia and the JUNE 2008 version of Japanese Wikipedia for our experiments. 24,000 hyponymy-relation candidates, randomly selected in both languages, were manually checked to build training, development, and test sets[6]. Around 8,000 hyponymy relations were found in the manually checked data for both languages[7]. 20,000 of the manually checked data were used as a training set for training the initial classifier. The rest were equally divided into development and test sets. The development set was used to select the optimal parameters in bilingual co-training and the test set was used to evaluate our system.

We used TinySVM (TinySVM, 2002) with a polynomial kernel of degree 2 as a classifier. The maximum iteration number in the bilingual co-training was set as 100. Two parameters, θ and $TopN$, were selected through experiments on the development set. $\theta = 1$ and $TopN$=900 showed

[3]We obtained 1.6 M object-attribute-value triples in Japanese and 5.9 M in English.

[4]197 K translation pairs were extracted.

[5]We also used redirection links in English and Japanese Wikipedia for recognizing the variations of terms when we built a bilingual instance dictionary with Wikipedia cross-language links.

[6]It took about two or three months to check them in each language.

[7]Regarding a hyponymy relation as a positive sample and the others as a negative sample for training SVMs, "positive sample:negative sample" was about 8,000:16,000=1:2

the best performance and were used as the optimal parameter in the following experiments.

We conducted three experiments to show effects of bilingual co-training, training data size, and bilingual instance dictionaries. In the first two experiments, we experimented with a bilingual instance dictionary derived from Wikipedia cross-language links. Comparison among systems based on three different bilingual instance dictionaries is shown in the third experiment.

Precision (P), recall (R), and F_1-measure (F_1), as in Eq (1), were used as the evaluation measures, where *Rel* represents a set of manually checked hyponymy relations and *HRbyS* represents a set of hyponymy-relation candidates classified as hyponymy relations by the system:

$$\begin{aligned} P &= |Rel \cap HRbyS|/|HRbyS| \quad (1) \\ R &= |Rel \cap HRbyS|/|Rel| \\ F_1 &= 2 \times (P \times R)/(P + R) \end{aligned}$$

4.1 Effect of Bilingual Co-Training

	ENGLISH			JAPANESE		
	P	R	F_1	P	R	F_1
SYT	78.5	63.8	70.4	75.0	77.4	76.1
INIT	77.9	67.4	72.2	74.5	78.5	76.6
TRAN	76.8	70.3	73.4	76.7	79.3	78.0
BICO	78.0	83.7	80.7	78.3	85.2	81.6

Table 2: Performance of different systems (%)

Table 2 shows the comparison results of the four systems. SYT represents the Sumida et al. (2008) system that we implemented and tested with the same data as ours. INIT is a system based on initial classifier c^0 in bilingual co-training. We translated training data in one language by using our bilingual instance dictionary and added the translation to the existing training data in the other language like bilingual co-training did. The size of the English and Japanese training data reached 20,729 and 20,486. We trained initial classifier c^0 with the new training data. TRAN is a system based on the classifier. BICO is a system based on bilingual co-training.

For Japanese, SYT showed worse performance than that reported in Sumida et al. (2008), probably due to the difference in training data size (ours is 20,000 and Sumida et al. (2008) was 29,900). The size of the test data was also different – ours is 2,000 and Sumida et al. (2008) was 1,000.

Comparison between INIT and SYT shows the effect of SF_3–SF_5 and IF, newly introduced feature types, in hyponymy-relation classification. INIT consistently outperformed SYT, although the difference was merely around 0.5–1.8% in F_1.

BICO showed significant performance improvement (around 3.6–10.3% in F_1) over SYT, INIT, and TRAN regardless of the language. Comparison between TRAN and BICO showed that bilingual co-training is useful for enlarging the training data and that the performance gain by bilingual co-training cannot be achieved by simply translating the existing training data.

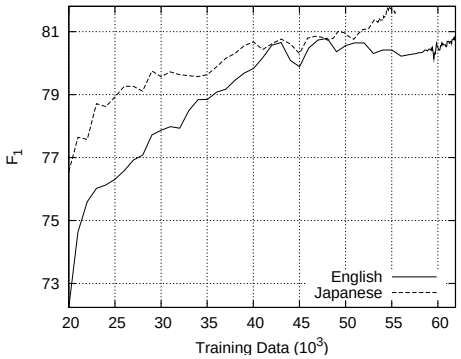


Figure 5: F_1 curves based on the increase of training data size during bilingual co-training

Figure 5 shows F_1 curves based on the size of the training data including those manually tailored and automatically obtained through bilingual co-training. The curve starts from 20,000 and ends around 55,000 in Japanese and 62,000 in English. As the training data size increases, the F_1 curves tend to go upward in both languages. This indicates that the two classifiers cooperate well to boost their performance through bilingual co-training.

We recognized 5.4 M English and 2.41 M Japanese hyponymy relations from the classification results of BICO on all hyponymy-relation candidates in both languages.

4.2 Effect of Training Data Size

We performed two tests to investigate the effect of the training data size on bilingual co-training. The first test posed the following question: "If we build $2n$ training samples by hand and the building cost is the same in both languages, which is better from the monolingual aspects: $2n$ monolingual training samples or n bilingual training samples?" Table 3 and Figure 6 show the results.

In INIT-E and INIT-J, a classifier in each language, which was trained with $2n$ monolingual training samples, did not learn through bilingual co-training. In BICO-E and BICO-J, bilingual co-training was applied to the initial classifiers trained with n training samples in both languages. As shown in Table 3, BICO, with half the size of the training samples used in INIT, always performed better than INIT in both languages. This indicates that bilingual co-training enables us to build classifiers for two languages in tandem with the same combined amount of data as required for training a single classifier in isolation while achieving superior performance.

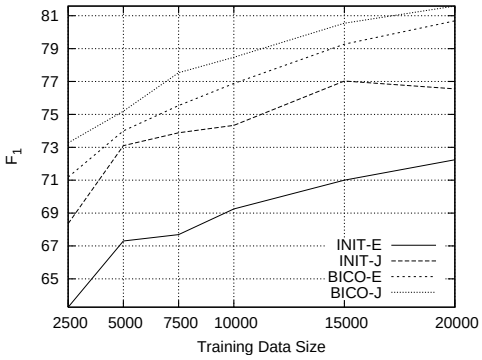


Figure 6: F_1 based on training data size: with/without bilingual co-training

n	*2n*		*n*	
	INIT-E	INIT-J	BICO-E	BICO-J
2500	67.3	72.3	70.5	73.0
5000	69.2	74.3	74.6	76.9
10000	72.2	76.6	76.9	78.6

Table 3: F_1 based on training data size: with/without bilingual co-training (%)

The second test asked: "Can we always improve performance through bilingual co-training with one strong and one weak classifier?" If the answer is yes, then we can apply our framework to acquisition of hyponymy-relations in other languages, i.e., German and French, without much effort for preparing a large amount of training data, because our strong classifier in English or Japanese can boost the performance of a weak classifier in other languages.

To answer the question, we tested the performance of classifiers by using all training data (20,000) for a strong classifier and by changing the training data size of the other from 1,000 to 15,000 ($\{$1,000, 5,000, 10,000, 15,000$\}$) for a weak classifier.

	INIT-E	BICO-E	INIT-J	BICO-J
1,000	72.2	79.6	64.0	72.7
5,000	72.2	79.6	73.1	75.3
10,000	72.2	79.8	74.3	79.0
15,000	72.2	80.4	77.0	80.1

Table 4: F_1 based on training data size: when English classifier is strong one

	INIT-E	BICO-E	INIT-J	BICO-J
1,000	60.3	69.7	76.6	79.3
5,000	67.3	74.6	76.6	79.6
10,000	69.2	77.7	76.6	80.1
15,000	71.0	79.3	76.6	80.6

Table 5: F_1 based on training data size: when Japanese classifier is strong one

Tables 4 and 5 show the results, where "INIT" represents a system based on the initial classifier in each language and "BICO" represents a system based on bilingual co-training. The results were encouraging because the classifiers showed better performance than their initial ones in every setting. In other words, a strong classifier always taught a weak classifier well, and the strong one also got help from the weak one, regardless of the size of the training data with which the weaker one learned. The test showed that bilingual co-training can work well if we have one strong classifier.

4.3 Effect of Bilingual Instance Dictionaries

We tested our method with different bilingual instance dictionaries to investigate their effect. We built bilingual instance dictionaries based on different translation dictionaries whose translation entries came from different domains (i.e., general domain, technical domain, and Wikipedia) and had a different degree of translation ambiguity. In Table 6, D1 and D2 correspond to systems based on a bilingual instance dictionary derived from two handcrafted translation dictionaries, EDICT (Breen, 2008) (a general-domain dictionary) and "The Japan Science and Technology Agency Dictionary," (a translation dictionary for technical terms) respectively. D3, which is the same as BICO in Table 2, is based on a bilingual

instance dictionary derived from Wikipedia. ENTRY represents the number of translation dictionary entries used for building a bilingual instance dictionary. E2J (or J2E) represents the average translation ambiguities of English (or Japanese) terms in the entries. To show the effect of these translation ambiguities, we used each dictionary under two different conditions, α=5 and ALL. α=5 represents the condition where only translation entries with less than five translation ambiguities are used; ALL represents no restriction on translation ambiguities.

DIC TYPE		F_1		DIC STATISTICS		
		E	J	ENTRY	E2J	J2E
D1	α=5	76.5	78.4	588K	1.80	1.77
D1	ALL	75.0	77.2	990K	7.17	2.52
D2	α=5	76.9	78.5	667K	1.89	1.55
D2	ALL	77.0	77.9	750K	3.05	1.71
D3	α=5	80.7	81.6	197K	1.03	1.02
D3	ALL	80.7	81.6	197K	1.03	1.02

Table 6: Effect of different bilingual instance dictionaries

The results showed that D3 was the best and that the performances of the others were similar to each other. The differences in the F_1 scores between α=5 and ALL were relatively small within the same system triggered by translation ambiguities. The performance gap between D3 and the other systems might explain the fact that both hyponymy-relation candidates and the translation dictionary used in D3 were extracted from the same dataset (i.e., Wikipedia), and thus the bilingual instance dictionary built with the translation dictionary in D3 had better coverage of the Wikipedia entries consisting of hyponymy-relation candidates than the other bilingual instance dictionaries. Although D1 and D2 showed lower performance than D3, the experimental results showed that bilingual co-training was always effective no matter which dictionary was used (Note that F_1 of INIT in Table 2 was 72.2 in English and 76.6 in Japanese.)

5 Related Work

Li and Li (2002) proposed bilingual bootstrapping for word translation disambiguation. Similar to bilingual co-training, classifiers for two languages cooperated in learning with bilingual resources in bilingual bootstrapping. However, the two classifiers in bilingual bootstrapping were for a bilingual task but did different tasks from the monolingual viewpoint. A classifier in each language is for word sense disambiguation, where a class label (or word sense) is different based on the languages. On the contrary, classifiers in bilingual co-training cooperate in doing the same type of tasks.

Bilingual resources have been used for monolingual tasks including verb classification and noun phrase semantic interpolation (Merlo et al., 2002; Girju, 2006). However, unlike ours, their focus was limited to bilingual features for one monolingual classifier based on supervised learning.

Recently, there has been increased interest in semantic relation acquisition from corpora. Some regarded Wikipedia as the corpora and applied hand-crafted or machine-learned rules to acquire semantic relations (Herbelot and Copestake, 2006; Kazama and Torisawa, 2007; Ruiz-casado et al., 2005; Nastase and Strube, 2008; Sumida et al., 2008; Suchanek et al., 2007). Several researchers who participated in SemEval-07 (Girju et al., 2007) proposed methods for the classification of semantic relations between simple nominals in English sentences. However, the previous work seldom considered the bilingual aspect of semantic relations in the acquisition of monolingual semantic relations.

6 Conclusion

We proposed a bilingual co-training approach and applied it to hyponymy-relation acquisition from Wikipedia. Experiments showed that bilingual co-training is effective for improving the performance of classifiers in both languages. We further showed that bilingual co-training enables us to build classifiers for two languages in tandem, outperforming classifiers trained individually for each language while requiring no more training data in total than a single classifier trained in isolation.

We showed that bilingual co-training is also helpful for boosting the performance of a weak classifier in one language with the help of a strong classifier in the other language without lowering the performance of either classifier. This indicates that the framework can reduce the cost of preparing training data in new languages with the help of our English and Japanese strong classifiers. Our future work focuses on this issue.

References

Sören Auer and Jens Lehmann. 2007. What have Innsbruck and Leipzig in common? Extracting semantics from wiki content. In *Proc. of the 4th European Semantic Web Conference (ESWC 2007)*, pages 503–517. Springer.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100.

Jim Breen. 2008. EDICT Japanese/English dictionary file, The Electronic Dictionary Research and Development Group, Monash University.

Hal Daumé III, John Langford, and Daniel Marcu. 2005. Search-based structured prediction as classification. In *Proc. of NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, Whistler, Canada.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. 2008. A bilingual dictionary extracted from the Wikipedia link structure. In *Proc. of DASFAA*, pages 686–689.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proc. of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18.

Roxana Girju. 2006. Out-of-context noun phrase semantic interpretation with cross-linguistic evidence. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 268–276.

Aurelie Herbelot and Ann Copestake. 2006. Acquiring ontological relationships from Wikipedia using RMRS. In *Proc. of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*.

Jun'ichi Kazama and Kentaro Torisawa. 2007. Exploiting Wikipedia as external knowledge for named entity recognition. In *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 698–707.

Cong Li and Hang Li. 2002. Word translation disambiguation using bilingual bootstrapping. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 343–351.

MeCab. 2008. MeCab: Yet another part-of-speech and morphological analyzer. `http://mecab.sourceforge.net/`.

Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. 2002. A multilingual paradigm for automatic verb classification. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 207–214.

Vivi Nastase and Michael Strube. 2008. Decoding Wikipedia categories for knowledge acquisition. In *Proc. of AAAI 08*, pages 1219–1224.

Maria Ruiz-casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic extraction of semantic relationships for Wordnet by means of pattern learning from Wikipedia. In *Proc. of NLDB*, pages 67–79. Springer Verlag.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proc. of the 16th international conference on World Wide Web*, pages 697–706.

Asuka Sumida and Kentaro Torisawa. 2008. Hacking Wikipedia for hyponymy relation acquisition. In *Proc. of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, pages 883–888, January.

Asuka Sumida, Naoki Yoshinaga, and Kentaro Torisawa. 2008. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

TinySVM. 2002. `http://chasen.org/~taku/software/TinySVM`.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Fei Wu and Daniel S. Weld. 2007. Autonomously semantifying Wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50.