A Naïve Bayes Approach to Cross-Lingual Word Sense Disambiguation and Lexical Substitution*

David Pinto, Darnes Vilariño, Carlos Balderas,
Mireya Tovar, and Beatriz Beltrán

Faculty of Computer Science
B. Autonomous University of Puebla, Mexico
{dpinto,darnes,mtovar,bbeltran}@cs.buap.mx

Abstract. Word Sense Disambiguation (WSD) is considered one of the most important problems in Natural Language Processing [1]. It is claimed that WSD is essential for those applications that require of language comprehension modules such as search engines, machine translation systems, automatic answer machines, second life agents, etc. Moreover, with the huge amounts of information in Internet and the fact that this information is continuosly growing in different languages, we are encourage to deal with cross-lingual scenarios where WSD systems are also needed. On the other hand, Lexical Substitution (LS) refers to the process of finding a substitute word for a source word in a given sentence. The LS task needs to be approached by firstly disambiguating the source word, therefore, these two tasks (WSD and LS) are somehow related. In this paper, we present a naïve approach to tackle the problem of cross-lingual WSD and cross-lingual lexical substitution. We use a bilingual statistical dictionary, which is calculated with Giza++ by using the EUROPARL parallel corpus, in order to calculate the probability of a source word to be translated to a target word (which is assumed to be the correct sense of the source word but in a different language). Two versions of the probabilistic model are tested: unweighted and weighted. The results were compared with those of an international competition, obtaining a good performance.

1 Introduction

Word Sense Disambiguation is a task that consists in selecting the correct sense of a given ambiguous word in a given context. There are several approaches that have been proposed for WSD [1], however, the problem of automatic WSD has not been resolved. Competitions such as Senseval[1] and recently SemEval[2] have also motivated the generation of new systems for WSD, providing an interesting

* This work has been partially supported by the CONACYT project #106625, as well as by the PROMEP/103.5/09/4213 grant.

[1] http://www.senseval.org/

[2] http://nlp.cs.swarthmore.edu/semeval/
http://semeval2.fbk.eu/

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2010, LNCS 6256, pp. 352–361, 2010.

environment for testing those systems. Despite the WSD task has been studied for a long time, the expected feeling is that WSD should be integrated into real applications such as mono and multi-lingual search engines, machine translation systems, automatic answer machines, etc [1]. Different studies on this issue have demonstrated that those applications benefit from WSD. For instance, the case of machine translation [2,3].

Even if the problem of WSD is difficult when dealing in only one language, when we consider its cross-lingual version (C-WSD), this problem becomes to be much more complex. In this case, it is needed not only to find the correct translation, but this translation must consider the contextual senses of the original sentence (in a source language), in order to find the correct sense (in the target language) of the source word.

For the experiments carried out in this paper, we have considered English as the source language and Spanish as the target language. We do not use an inventory of senses, as the most of the WSD systems do. Instead, we attempt to find those senses automatically by means of a bilingual statistical dictionary which is calculated on the basis of the IBM-1 translation model[3], by using the EUROPARL parallel corpus[4]. In this way, we obtain a set canditate translations for the source ambiguous word and applying a probabilistic model we may rank those translations in order to determine the most probable word/sense for the ambiguous word.

We have also considered the problem of Cross-lingual Lexical Substitution(C-LS) for the experiments presented in this paper. The aim was to test the results obtained in C-WSD to solve the problem of C-LS. In general, the C-LS problem may be defined as follows: given a paragraph and a source word, the goal is to provide several correct translations for that word in a given language, with the constraint that the translations fit the given context in the source language. We consider this task to be a step forward of the English lexical substitution task from SemEval-2007 [4], but this time the problem is considered in a cross-lingual scenario.

The rest of this paper is structured as follows. Section 2 presents the two datasets used in the experiments. In Section 3 we define the probabilistic model used as classifier for both, the cross-lingual WSD and LS. The experimental results are shown in Section 4 together with a discussion of findings. Finally, the conclusions and further work are given in Section 5.

2 Datasets

For the experiments conducted on cross-lingual word sense disambiguation we have used 25 polysemous English nouns. We selected five nouns (movement, plant, occupation, bank and passage), each with 20 example instances, for conforming a development corpus. The remaining polysemous nouns (twenty) were considered for a test corpus. In the case of the test corpus, we used 50 instances per noun. A list of the ambiguous nouns of the test corpus may be seen in Table 1.

[3] We used Giza++ (`http://fjoch.com/GIZA++.html`)

[4] `http://www.statmt.org/europarl/`

Table 1. Test set for the cross-lingual WSD task

Noun name			
coach	education	execution	figure
job	post	pot	range
rest	ring	mood	soil
strain	match	scene	test
mission	letter	paper	side

Table 2. Development set for the cross-lingual lexical substitution task

Polysemous word name									
take	v	stand	v	run	v	manage	v	lie	v
forget	v	fix	v	find	v	fear	v	bar	v
well	r	tight	r	severely	r	nearly	r	finally	r
wild	n	stand	n	gall	n	film	n	examination	n
dark	n	cross	n	can	n	bar	n	wild	a
rough	a	rich	a	reasonable	a	outdoor	a	neat	a
nasty	a	grim	a	cross	a	bright	a		

On the other hand, for the experiments carried out on the cross-lingual lexical substitution we employed two corpora: the development corpus and the test corpus. In Table 2 we may see the different polysemous words used in the development corpus. Whereas, Table 3 shows the different ambiguous words used. As may be seen in the case of the C-LS task, we have considered other grammatical categories different than nouns. We denoted verbs with v, nouns with n, adjectives with a and adverbs with r.

3 A Naïve Bayes Approach to WSD and LS

In this section it is presented an overview of the presented system, but also we further discuss the particularities of the general approach for each task evaluated. We will start this section by explaining the manner we deal with the C-WSD problem.

3.1 Cross-Lingual Word Sense Disambiguation

In Figure 1 we may see the complete process of approaching the problem of cross-lingual WSD.

We have approached the cross-lingual word sense disambiguation task by means of a probabilistic system which considers the probability of a word sense (in a target language), given a sentence (in a source language) containing the ambiguous word. In particular, we used the Naive Bayes classifier in two different ways. First, we calculated the probability of each word in the source language of being associated/translated to the corresponding word (in the target language).

Table 3. Test set for the cross-lingual lexical substitution task

Polysemous word name									
work	v	wind	v	touch	v	throw	v	tap	v
strike	v	skip	v	show	v	shed	v	see	v
return	v	render	v	put	v	pass	v	order	v
let	v	hold	v	go	v	fire	v	drop	v
draw	v	dismiss	v	clear	v	clean	v	check	v
charge	v	carry	v	call	v	burst	v	bring	v
acquire	v	yet	r	right	r	only	r	now	r
late	r	hard	r	forward	r	closely	r	away	r
around	r	about	r	way	n	test	n	strain	n
side	n	shot	n	shade	n	scene	n	ring	n
rest	n	range	n	pulse	n	pot	n	post	n
paper	n	mission	n	match	n	mass	n	lead	n
investigator	n	girl	n	function	n	figure	n	field	n
execution	n	coach	n	cap	n	bug	n	board	n
blow	n	account	n	tender	a	strong	a	straight	a
stiff	a	special	a	solid	a	soft	a	serious	a
rude	a	raw	a	profound	a	poor	a	open	a
new	a	live	a	light	a	liberal	a	informal	a
heavy	a	good	a	fundamental	a	fresh	a	flat	a
extended	a	dry	a	clear	a	civil	a	blue	a

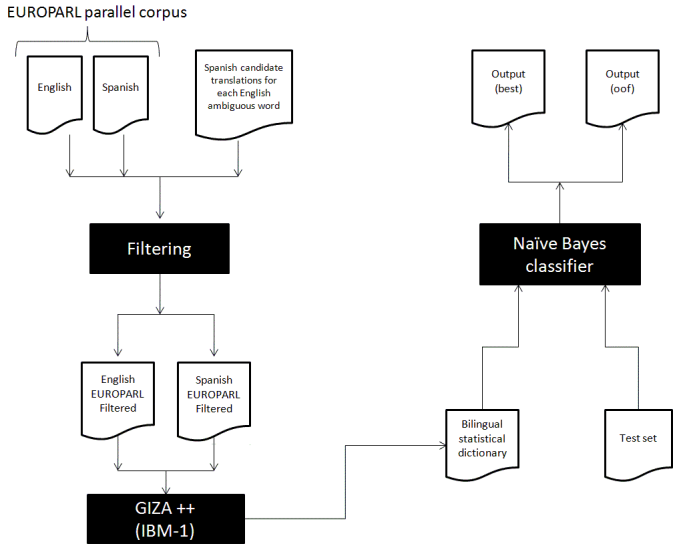

Fig. 1. An overview of the presented approach for cross-lingual word sense disambiguation

The probabilities were estimated by means of a bilingual statistical dictionary which is calculated using the Giza++ system over the EUROPARL parallel corpus. We filtered this corpus by selecting only those sentences which included some senses of the ambiguous word which were obtained by translating this ambiguous word on the Google search engine. The second approach considered a weighted probability for each word in the source sentence. The closer a word of the sentence to the ambiguous word, the higher the weight given to it.

In other words, given an English sentence $S = \{w_1, w_2, \cdots, w_k, \cdots, w_{k+1}, \cdots\}$ with the ambiguos word w_k in position k. Let us consider N candidate translations of w_k, $\{t_1^k, t_2^k, \cdots, t_N^k\}$ obtained somehow (we will further discuss about this issue in this section). We are insterested on finding the most probable candidate translations for the polysemous word w_k. Therefore, we may use a Naïve Bayes classifier which considers the probability of t_i^k given w_k. A formal description of the classifier is given as follows.

$$p(t_i^k|S) = p(t_i^k|w_1, w_2, \cdots, w_k, \cdots) \tag{1}$$

$$p(t_i^k|w_1, w_2, \cdots, w_k, \cdots) = \frac{p(t_i^k)p(w_1, w_2, \cdots, w_k, \cdots|t_i^k)}{p(w_1, w_2, \cdots, w_k, \cdots)} \tag{2}$$

We are interested on finding the argument that maximizes $p(t_i^k|S)$, therefore, we may avoid calculating the denominator. Moreover, if we assume that all the different translations are equally distributed, then Eq. (2) must be approximated by Eq. (3).

$$p(t_i^k|w_1, w_2, \cdots, w_k, \cdots) \approx p(w_1, w_2, \cdots, w_k, \cdots|t_i^k) \tag{3}$$

The complete calculation of Eq. (3) requires to apply the chain rule. However, if we assumed that the words of the sentence are independent, then we may rewrite Eq. (3) as Eq. (4).

$$p(t_i^k|w_1, w_2, \cdots, w_k, \cdots) \approx \prod_{j=1}^{|S|} p(w_j|t_i^k) \tag{4}$$

The best translation is obtained as shown in Eq. (5). Nevertheless the position of the ambiguous word, we are only considering a product of the probabilites of translation. Thus, we named this approach, the *unweighted version*. Algorithm 1 provides details about implementation.

$$BestSense_u(S) = \arg\max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k) \tag{5}$$

with $i = 1, \cdots, N$.

A second approach (*weighted version*) is also proposed as shown in Eq. (6). Algorithm 2 provides details about implementation.

$$BestSense_w(S) = \arg\max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k) * \frac{1}{k-j+1} \tag{6}$$

with $i = 1, \cdots, N$.

Algorithm 1. An unweighted naïve Bayes approach to cross-lingual WSD

Input: A set Q of sentences: $Q = \{S_1, S_2, \cdots\}$;
$Dictionary = p(w|t)$: A bilingual statistical dictionary;
Output: The best word/sense for each ambiguous word $w_j \in S_l$

```
for l = 1 to |Q| do
    for i = 1 to N do
        P_{l,i} = 1;
        for j = 1 to |S_l| do
            foreach w_j ∈ S_l do
                if w_j ∈ Dictionary then
                    P_{l,i} = P_{l,i} * p(w_j|t_i^k);
                else
                    P_{l,i} = P_{l,i} * ε;
                end
            end
        end
    end
end
return arg max_{t_i^k} ∏_{j=1}^{|S|} p(w_j|t_i^k)
```

With respect to the N candidate translations of the polysemous word w_k, $\{t_1^k, t_2^k, \cdots, t_N^k\}$, we have used of the Google translator[5]. Google provides all the possible translations for w_k with the corresponding grammatical category. Therefore, we are able to use those translations that match with the same grammatical category of the ambiguous word. Even if we attempted other approaches such as selecting the most probable translations from the statistical dictionary, we confirmed that by using the Google online translator we obtain the best results. We consider that this result is derived from the fact that Google has a better language model than we have, because our bilingual statistical dictionary was trained only with the EUROPARL parallel corpus.

The experimental results of both, the *unweighted* and the *weighted* versions of the presented approach for cross-lingual word sense disambiguation are given in Section 4.

3.2 Cross-Lingual Lexical Substitution

In Figure 2 we may see the complete process of approaching the problem of cross-lingual lexical substitution. Notice that this task is complemented by the WSD solver.

This module is based on the cross-lingual word sense disambiguation system. Once we knew the best word/sense (Spanish) for the ambiguous word (English), we lemmatized the Spanish word. We searched, at WordNet, the synonyms of this word (sense) that agree with the grammatical category (noun, verb, etc) of the query (source polysemous word).

[5] `http://translate.google.com.mx/`

Algorithm 2. A weighted naïve Bayes approach to cross-lingual WSD

Input: A set Q of sentences: $Q = \{S_1, S_2, \cdots\}$;
$Dictionary = p(w|t)$: A bilingual statistical dictionary;
Output: The best word/sense for each ambiguous word $w_j \in S_l$

for $l = 1$ *to* $|Q|$ **do**
  for $i = 1$ *to* N **do**
    $P_{l,i} = 1$;
    for $j = 1$ *to* $|S_l|$ **do**
      foreach $w_j \in S_l$ **do**
        if $w_j \in Dictionary$ **then**
          $P_{l,i} = P_{l,i} * p(w_j|t_i^k) * \frac{1}{k-j+1}$;
        else
          $P_{l,i} = P_{l,i} * \epsilon$;
        end
      end
    end
  end
end
return arg $\max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k) * \frac{1}{k-j+1}$

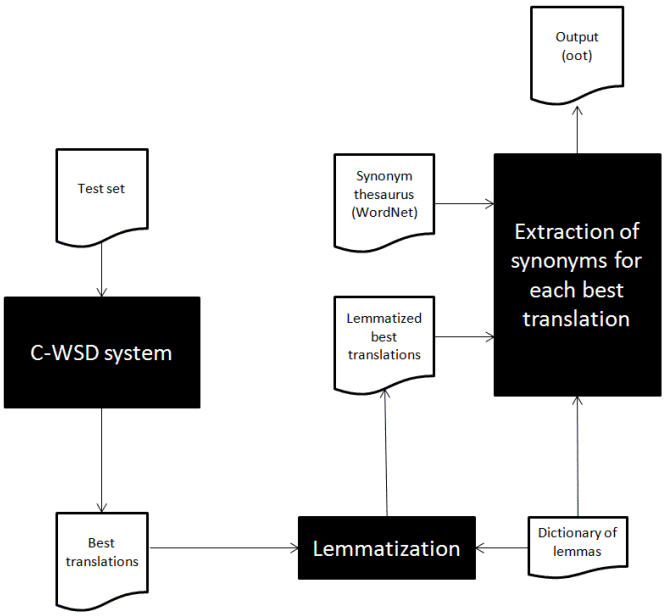

Fig. 2. An overview of the presented approach for cross-lingual lexical substitution

Table 4. Description of runs

Run name	*Description*
FCC-WSD1	Best translation (one target word) / unweighted version
FCC-WSD2	Ten best translations (ten target words - *oof*) / unweighted version
FCC-WSD3	Best translation (one target word) / weighted version
FCC-WSD4	Ten best translations (ten target words - *oof*) / weighted version

4 Experimental Results

In this section we present the obtained results for both, the cross-lingual word sense disambiguation task and the cross-lingual lexical substitution task.

4.1 Cross-Lingual Word Sense Disambiguation

In Table 5 we may see the results we have obtained with the different versions of the presented approach. In particular, we have tested four different runs which correspond to two evaluations for each different version of the probabilistic classifier. The description of each run is given in Table 4.

In the same Table we can find a comparison of our runs with others approaches presented at the SemEval-2 competition. The *UvT* team submitted four runs (UvT-WSD1 and UvT-WSD2 for the both *best* and the *oof* evaluation) which make use of a k-nearest neighbour classifier to build one word sense for each target ambiguous word, and select translations from a bilingual dictionary obtained by executing the GIZA package on the EUROPARL parallel corpus [5]. The University of Heidelberg participated submitting other four runs (UHD-1 and UHD-2 for both the *best* and the *oof* evaluation). They approached the cross-lingual word sense disambiguation by finding the most appropriate translation in different languages on the basis of a multilingual co-ocurrence graph, which is automatically induced from the target words aligned contexts found in the EUROPARL and JRC-Arquis parallel corpora [5]. Finally, there was another team which submitted two runs: ColEur1 (*best* evaluation) and ColEur2 (*oof* evaluation) with a supervised approach that uses the translations obtained with GIZA from the EUROPARL parallel corpus in order to distinguish between senses in the English source sentences [5]. In general, we may see that all the teams used the GIZA software in order to find a bilingual statistical dictionary. Therefore, the main differences among all these approaches are in the way that they represents the original ambiguous sentence (including the pre-processing stage), and the manner the teams filter the results obtained by GIZA.

We obtained a better performance with those runs that were evaluated with the ten best translations than with those that were evaluated with only the best ones. This fact lead us to consider in further work to improve the ranking of the translations found by our system. On other hand, the unweighted version

Table 5. Evaluation of the cross-lingual word sense disambiguation task

System name	*Precision (%)*	*Recall (%)*
UvT-WSD1	23.42	23.42
UvT-WSD2	19.92	19.92
FCC-WSD1	15.09	15.09
FCC-WSD3	14.43	14.43
UHD-1	20.48	16.33
UHD-2	20.2	16.09
ColEur1	19.78	19.59

a) Best translation

System name	*Precision (%)*	*Recall (%)*
UvT-WSD1	42.17	42.17
UvT-WSD2	43.12	43.12
FCC-WSD2	40.76	40.76
FCC-WSD4	38.46	38.46
UHD-1	38.78	31.81
UHD-2	37.74	31.3
ColEur2	35.84	35.46

b) Five best translations (oof)

of the proposed classifier improved the weighted one. This behavior was unexpected, because in the development dataset, the results were opposite. We got a better performance than other systems, and those runs that outperformed our system runs did it by around 3% of precision and recall in the case of the oof evaluation.

4.2 Cross-Lingual Lexical Substitution

In Table 6 we may see the obtained results for the cross-lingual lexical substitution task. The obtained results are low in comparison with the best one (the complete description of all the runs may be found in [6]). Since this task relies on the C-WSD task, then a lower performance on the C-WSD task will conduct to a even lower performance in C-LS. Firstly, we need to improve the C-WSD solver. In particular, we need to improve the ranking procedure in order to obtain a better translation of the source ambiguous word. Moreover, we consider that the use of language modeling would be of high benefit, since we could test whether or not a given translation together with the terms in its context would have high probability in the target language.

Table 6. Evaluation of the cross-lingual lexical substitution task (the ten best results - *oot*)

System name	*Precision (%)*	*Recall (%)*
UvT-v	58.91	58.91
UvT-g	55.29	55.29
UBA-W	52.75	52.75
WLVUSP	48.48	48.48
UBA-T	47.99	47.99
USPWLV	47.6	47.6
ColSlm	43.91	46.61
ColEur	41.72	44.77
TYO	34.54	35.46
IRST-1	31.48	33.14
FCC-LS	23.9	23.9
IRSTbs	8.33	29.74

5 Conclusions and Further Work

In this paper we have presented a system for cross-lingual word sense disambiguation and cross-lingual lexical substitution. The approach uses a Naïve Bayes classifier which is feed with the probabilities obtained from a bilingual statistical dictionary. Two different versions of the classifier, unweighted and weighted were tested. The results were compared with those of an international competition, obtaining a good performance. As further work, we need to improve the ranking module of the cross-lingual WSD classifier. Moreover, we consider that the use of a language model for Spanish would highly improve the results on the cross-lingual lexical substitution task.

References

1. Aguirre, E., Edmonds, P.: Word Sense Disambiguation, Text, Speech and Language Technology. Springer, Heidelberg (2006)
2. Chan, Y., Ng, H., Chiang, D.: Word sense disambiguation improves statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 33–40 (2007)
3. Carpuat, M., Wu, D.: Improving statistical machine translation using word sense disambiguation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL), pp. 61–72 (2007)
4. McCarthy, D., Navigli, R.: English lexical substitution task. In: SemEval 2007 Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 48–53 (2007)
5. Mihalcea, R., Sinha, R., McCarthy, D.: Semeval-2010 task2: cross-lingual lexical substitution. In: Proceedings of the Fifth International Workshop on Semantic Evaluations (SemEval-2010). Association for Computational Linguistics (2010)
6. Lefever, E., Hoste, V.: Semeval-2010 task3:cross-lingual word sense disambiguation. In: Proceedings of the Fifth International Workshop on Semantic Evaluations (SemEval-2010). Association for Computational Linguistics (2010)