FCC: Modeling Probabilities with GIZA++ for Task #2 and #3 of SemEval-2

Darnes Vilariño, Carlos Balderas, David Pinto, Miguel Rodríguez, Saul León
Faculty of Computer Science, BUAP
Puebla, Mexico
{darnes,mrodriguez,dpinto}@cs.buap.mx

Abstract

In this paper we present a naïve approach to tackle the problem of cross-lingual WSD and cross-lingual lexical substitution which correspond to the Task #2 and #3 of the SemEval-2 competition. We used 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 obtained values show that the unweighted version performs better thant the weighted one.

1 Introduction

Word Sense Disambiguation (WSD) is considered one of the most important problems in Natural Language Processing (Agirre and Edmonds, 2006). 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. 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 (Agirre and Edmonds, 2006). Different studies on this issue have demonstrated that those applications benefit from WSD, such as in the case of machine translation (Chan et al., 2007; Carpuat and Wu., 2007). 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.

Since we are describing the modules of our system, we did not provide information of the datasets used. For details about the corpora, see the task description paper for both tasks (#2 and #3) in this volume (Mihalcea et al., 2010; Lefever and Hoste, 2010). Description about the other teams are also described in the same papers.

2 A Naïve 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 Cross-Lingual Word Sense Disambiguation (C-WSD) problem.

2.1 Cross-Lingual Word Sense Disambiguation

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). 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 sen-

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 112–116,
Uppsala, Sweden, 15-16 July 2010.

tences which included some senses of the ambiguous word which were obtained by translating this ambiguous word on the Google search engine.

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

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 ambiguous 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.

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

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

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

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

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).

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

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 the implementation.

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

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)
```

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

$$BestSense_w(w_k) =$$

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

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[1]. 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

[1] http://translate.google.com.mx/

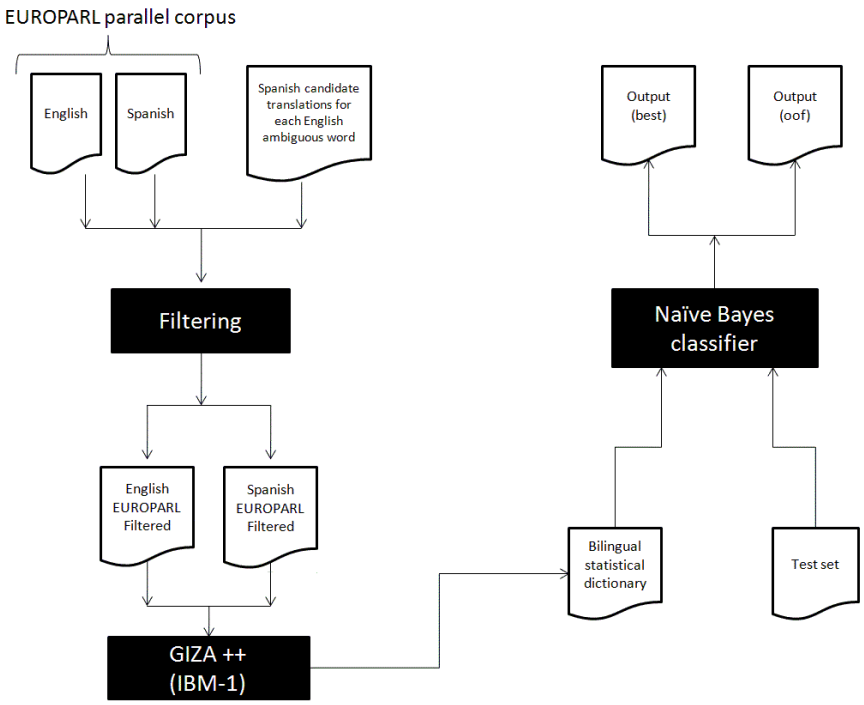

Figure 1: An overview of the presented approach for cross-lingual word sense disambiguation

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

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 3.

2.2 Cross-Lingual Lexical Substitution

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. Thereafter, 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), and we return those synonyms as possible lexical substitutes. Notice again that this task is complemented by the WSD solver.

In Figure 2 we may see the complete process of approaching the problem of cross-lingual lexical substitution.

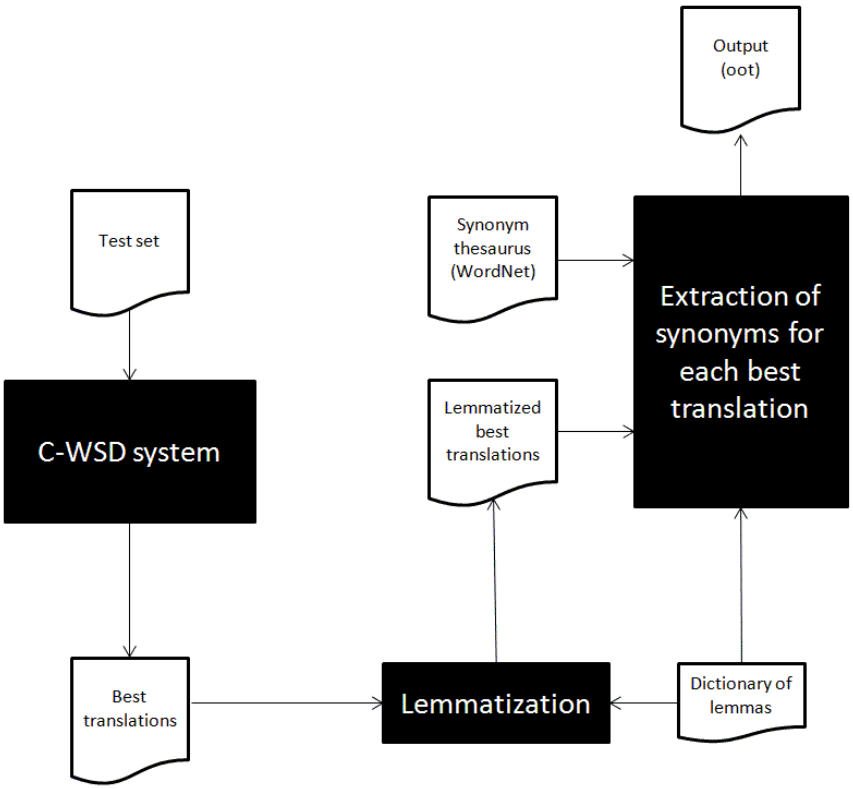

Figure 2: An overview of the presented approach for cross-lingual lexical substitution

3 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.

3.1 Cross-Lingual Word Sense Disambiguation

In Table 2 we may see the results we have obtained with the different versions of the presented approach. In the same Table we can find a comparison of our runs with others presented at the SemEval-2 competition. 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 1.

We obtained a better performance with those runs that were evaluated with the five best translations (oof) 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 of the proposed classifier improved the weighted one. This behavior was unexpected, because in the development dataset, the results were opposite. We consider that the problem comes from taking into account the entire sentence instead of a neighborhood (windows) around the ambiguous word. We will further investigate about this issue. 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.

3.2 Cross-Lingual Lexical Substitution

In Table 3 we may see the obtained results for the cross-lingual lexical substitution task. The obtained results are low in comparison with the best one. 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.

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

Table 1: Description of runs

System name	*Precision (%)*	*Recall (%)*
UvT-v	23.42	23.42
UvT-g	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
T3-COLEUR	19.78	19.59

a) Best translation

System name	*Precision (%)*	*Recall (%)*
UvT-v	42.17	42.17
UvT-g	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
T3-COLEUR	35.84	35.46

b) Five best translations (oof)

Table 2: Evaluation of the cross-lingual word sense disambiguation task

System name	*Precision (%)*	*Recall (%)*
SWAT-E	174.59	174.59
SWAT-S	97.98	97.98
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
DICT	44.04	44.04
DICTCORP	42.65	42.65

Table 3: Evaluation of the cross-lingual lexical substitution task (the ten best results - *oot*)

4 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 fed 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.

Acknowledgments

This work has been partially supported by CONACYT (Project #106625) and PROMEP (Grant #103.5/09/4213).

References

[Agirre and Edmonds2006] E. Agirre and P. Edmonds. 2006. *Word Sense Disambiguation, Text, Speech and Language Technology*. Springer.

[Carpuat and Wu.2007] M. Carpuat and D. Wu. 2007. 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)*, pages 61–72.

[Chan et al.2007] Y.S. Chan, H.T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40.

[Lefever and Hoste2010] E. Lefever and V. Hoste. 2010. Semeval-2010 task3:cross-lingual word sense disambiguation. In *Proceedings of the Fifth International Workshop on Semantic Evaluations (SemEval-2010)*. Association for Computational Linguistics.

[Mihalcea et al.2010] R. Mihalcea, R. Sinha, and D. McCarthy. 2010. Semeval-2010 task2:cross-lingual lexical substitution. In *Proceedings of the Fifth International Workshop on Semantic Evaluations (SemEval-2010)*. Association for Computational Linguistics.