A Probabilistic Model Based on n-Grams for Bilingual Word Sense Disambiguation*

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

Faculty of Computer Science
Benemérita Universidad Autónoma de Puebla, Mexico
{darnes,dpinto,mtovar,bbeltran}@cs.buap.mx

Abstract. Word Sense Disambiguation (WSD) is considered one of the most important problems in Natural Language Processing. Even if the problem of WSD is difficult, when we consider its bilingual version, 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. In this paper we propose a model based on n-grams (3-grams and 5-grams) that significantly outperforms the last results that we presented at the cross-lingual word sense disambiguation task at the SemEval-2 forum. We use a naïve Bayes classifier for determining the probability of a target sense (in a target language) given a sentence which contains the ambiguous word (in a source language). For this purpose, we use a bilingual statistical dictionary, which is calculated with Giza++ by using the EUROPARL parallel corpus, in order to determine 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). As we mentioned, the results were compared with those of an international competition, obtaining a good performance.

Keywords: Bilingual word sense disambiguation, Naïve Bayes classifier, Parallel corpus.

1 Introduction

Word Sense Disambiguation (WSD) is a task that has been studied for a long time. The aim of WSD is to select the correct sense of a given ambiguous word in some context. The fact that automatic WSD still being an open problem has motivated a great interest on the computational linguistics community, therefore, many approaches has been introduced in the last years [1]. Different studies have demonstrated that many real applications may get benefit from WSD (see for

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

G. Sidorov et al. (Eds.): MICAI 2010, Part I, LNAI 6437, pp. 82–91, 2010.

instance, [2,3]). The selection of the appropiate sense for a given ambiguous word, is commonly carried out by considering the words sorrounding the ambiguous word. A very complete survey of several approaches may be found in [1]. As may be seen, a lot of work has been done on finding the best supervised learning approach for WSD (see for instance [4,5,6,7]), but despite the wide range of learning algorithms, it has been noted that some classifiers such as Naïve Bayes are very competitive and their performance basically relies on the representation schemata and their feature selection process.

Monolingual word sense disambiguation is known for being a difficult task, however, when we consider its cross-lingual version, 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, thus, we attempted the *bilingual* version of WSD. 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[1], by using the EUROPARL parallel corpus[2].

This bilingual statistical dictionary feeds a Naïve Bayes classifier in order to determine the probability of sense given a source sentence which contains the ambiguous word. The manner in which we filter the content words of each sentence lead to present three different approaches based on n-grams whose performance is shown in this paper. Given a sentence S, we first consider its representation by using one $|S|$-gram. The second approach split the sentence into different 3-grams with the constraint of having each the ambiguous word. The third approach considers all the 5-grams extracted from the original sentence that again contain the ambiguous word. For each approach proposed, we obtain a candidate set of translations for the source ambiguous word by applying the probabilistic model on the basis of the n-grams selected.

Noticed that there are other works in literature that have used parallel corpora (bilingual or multilingual) for dealing with the problem of WSD (see for instance [8,9]). However, at difference of these approaches in which it is expected to find the best sense in the same language (despite using other languages for training the learning model), in this case we are interested on finding the best translated word, i.e., with the correct sense in a different language.

The rest of this paper is structured as follows. Section 2 presents the problem of bilingual word sense disambiguation. In Section 3 we define the probabilistic model used as classifier for the bilingual WSD task. The experimental results obtained with the two datasets used are shown in Section 4. Finally, the conclusions and further work are given in Section 5.

---

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

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

2 Bilingual Word Sense Disambiguation

Word sense disambiguation is an important task in multilingual scenarios due to fact that the meanings represented by an ambiguous word in one source language may be represented by multiple words in another language. Consider the word "bank" which may have up to 42 different meanings[3]. If we select one of these meanings, let's say: *put into a bank account* (to bank). The corresponding meaning in other languages would be *to make a deposit*. In Spanish, for instance, you would never say *She banks her paycheck every month* (Ella *bankea* su cheque cada mes), but *She deposites her paycheck every month* (Ella *deposita* su cheque cada mes). Therefore, the ability for disambiguating a polysemous word from one language to another one is essential to the task of machine translation and those Natural Language Processing (NLP) tasks related with it, such as cross-lingual lexical substitution [10].

In the task of bilingual word sense disambiguation we are required to obtain those translations of a given ambiguous word that match with the original word sense. In the following example, we have a sentence as input with one polysemous word to be disambiguated. The expected results are also given as follows.

Input sentence: ... equivalent to giving fish to people living on the *bank* of the river ... [English]

Output sense label:
Sense Label = {oever/dijk} [Dutch]
Sense Label = {rives/rivage/bord/bords} [French]
Sense Label = {Ufer} [German]
Sense Label = {riva} [Italian]
Sense Label = {orilla} [Spanish]

The bilingual WSD system is able to find the corresponding translation of "bank" in the target language with the same sense meaning. In order to approach this problem we propose the use of a probabilistic model based on n-grams. This propossal is discussed in the following section.

3 A Naïve Bayes Approach to Bilingual WSD

We have approached the bilingual word sense disambiguation task by means of a probabilistic system based on Naïve Bayes, which considers the probability of a word sense (in a target language), given a sentence (in a source language) containing the ambiguous word. 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

[3] http://ardictionary.com/Bank/742

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.

We will start this section by explaining the manner we represent the source documents (n-grams) in order to approach the bilingual word sense disambiguation problem. We further discuss the particularities of the general approach for each task evaluated.

3.1 The n-Grams Model

In order to represent the input sentence we have considered a model based on n-grams. In the experiments presented in this paper, we have considered three different approaches, which are described as follows.

Given a sentence S, we first consider its representation by using one $|S|$-gram. The second approach split the sentence into different 3-grams with the constraint of having each the ambiguous word. The third approach considers all the 5-grams extracted from the original sentence that again contain the ambiguous word.

Consider the following example for the ambiguous word *execution* and its pre-processed version which was just obtained by eliminating punctuation symbols and stop words (none other pre-processing step was performed):

Input sentence: Allegations of Iraqi army brutality, including summary *executions* and the robbing of civilians at gun-point for food, were also reported frequently during February.

Pre-processed input sentence: Allegations Iraqi army brutality including summary *executions* robbing civilians gun-point food reported frequently during February

n-gram model:

$|S|$-gram: Allegations Iraqi army brutality including summary *executions* robbing civilians gun-point food reported frequently during February

3-gram: {including, summary, *executions*}, {summary, *executions* robbing}, {*executions* robbing, civilians}

5-gram: {army, brutality, including, summary, *executions*}, {brutality, including, summary, *executions* robbing}, {including, summary, *executions* robbing, civilians}, {summary, *executions* robbing, civilians, gun-point}, {*executions* robbing, civilians, gun-point, food}

For each approach of the n-grams sentence representation proposed, we obtain a candidate set of translations for the source ambiguous word by applying one probabilistic model on the basis of the n-grams selected. See the following section for further details.

3.2 The Probabilistic Model

Given an English sentence S_E, we consider its representation based on n-grams as discussed in the previous section. Let $S = \{w_1, w_2, \cdots, w_k, \cdots, w_{k+1}, \cdots\}$ be the n-gram representation of S_E by bringing together all the n-grams, where w_k is the ambiguos word. 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 interested 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. 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 $i = 1, \cdots, N$.

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[4]. 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.

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

Algorithm 1. A Naïve Bayes approach to bilingual 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**
          $P_{l,i} = P_{l,i} * p(w_j|t_i^k)$;
        else
          $P_{l,i} = P_{l,i} * \epsilon$;
        end
      end
    end
  end
end
return $\arg\max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k)$

In Figure 1 we may see the complete process in which we approach the problem of bilingual WSD.

The experimental results of the different sentence representations based on n-grams for bilingual word sense disambiguation are given in the following section.

4 Experimental Results

In this section we present the obtained results for the bilingual word sense disambiguation task. We first describe the corpus used in the experiments and, thereafter, we present the evaluation of the three different sentence representations based on n-grams.

4.1 Datasets

For the experiments conducted 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 as the 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. Noticed that this corpus does not contains a repository of senses, since the task requires to find the most probable translation (with the correct sense) of a given ambiguous word.

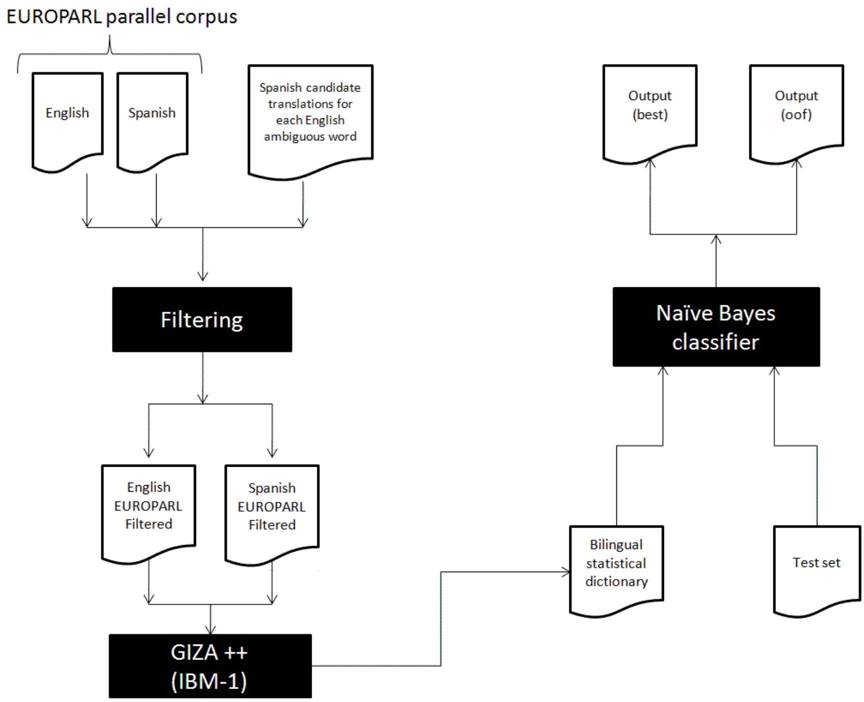

Fig. 1. An overview of the presented approach for bilingual word sense disambiguation

Table 1. Test set for the bilingual WSD task

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

4.2 Evaluation of the n-Gram Based Sentence Representation

In Table 2 we may see the results obtained with the different versions of n-gram sentence representation when we evaluated the model with the corpus presented in Table 1. The runs are labeled as follows:

3-gram: A representation of the sentence based on trigrams.
5-gram: A representation of the sentence based on 5-grams.
$|S|$**-gram:** A sentence representation based on a unique n-gram of length $|S|$.

With the purpose of observing the performance of the proposed approaches, we show in the same Table the results obtained by other systems at the SemEval-2 competition. A simple comparison lead to verify that two of the proposed sentence representations improve the rest of the approaches.

Table 2. Evaluation of the bilingual word sense disambiguation task - Five best translations (oof)

System name	*Precision (%)*	*Recall (%)*
3-gram	70.36	70.36
5-gram	54.81	54.81
UvT-WSD1	42.17	42.17
UvT-WSD2	43.12	43.12
$\|S\|$-*gram*	40.76	40.76
UHD-1	38.78	31.81
UHD-2	37.74	31.3
ColEur2	35.84	35.46

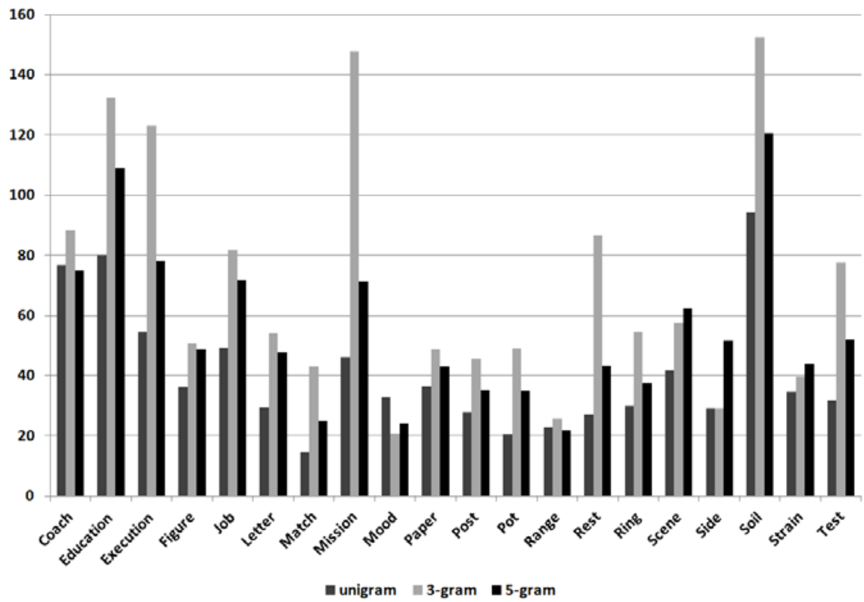

Fig. 2. An overview of the presented approach for bilingual word sense disambiguation

By observing the behaviour of precision over the different ambiguous words (see Figure 2), we may have a picture of the significant level of improving that may be reached when representing the sentence with 3-grams. We consider that again, the hypothesis of Harris[5] [11] is confirmed. The closer the words to the polysemous one, the better they can be used for disambiguating the ambiguous word. In Figure 2 we may also see that there are some words that are easier to be disambiguated (e.g. soil) than others (e.g. mood). For research purposes, we also consider important to focus the investigation on those words that are hard to be disambiguated.

[5] Words with similar syntactic usage have similar meaning.

5 Conclusions and Further Work

Bilingual word sense disambiguation is the task of obtaining those translations of a given ambiguous word that match with the original word sense. Different approaches have been presented in evaluations forums for dealing with this particular problem.

In this paper we propose a model based on n-grams (3-grams and 5-grams) that significantly outperforms the last results presented at the cross-lingual word sense disambiguation task at the SemEval-2 forum.

We use a Naïve Bayes classifier for determining the probability of a target sense (in a target language) given a sentence which contains the ambiguous word (in a source language). For this purpose, we use a bilingual statistical dictionary, which is calculated with Giza++ by using the EUROPARL parallel corpus, in order to determine 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). In order to represent the input sentence we have considered a model based on n-grams. For each approach of the n-grams sentence representation proposed, we obtain a candidate set of translations for the source ambiguous word by applying one probabilistic model on the basis of the n-grams selected.

As we mentioned, the results were compared with those of an international competition, obtaining a very good performance.

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. Florian, R., Yarowsky, D.: Modeling consensus: Classifier combination for word sense disambiguation. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 25–32 (2002)
5. Lee, Y.K., Ng, H.T.: An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 41–48 (2002)
6. Mihalcea, R.F., Moldovan, D.I.: Pattern learning and active feature selection for word sense disambiguation. In: Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), pp. 127–130 (2001)
7. Yarowsky, D., Cucerzan, S., Florian, R., Schafer, C., Wicentowski, R.: The johns hopkins senseval2 system descriptions. In: Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), pp. 163–166 (2001)

8. Ng, H.T., Wang, B., Chan, Y.S.: Exploiting parallel texts for word sense disambiguation: An empirical study. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 455–462 (2003)
9. Alonso, G.B.: Spanish word sense disambiguation with parallel texts (In spanish: Desambiguacion de los sentidos de las palabras en español usando textos paralelos). PhD thesis, Instituto Politécnico Nacional, Centro de Investigación en Computación (2010)
10. Sinha, R., McCarthy, D., Mihalcea, R.: Semeval-2010 task 2: Cross-lingual lexical substitution. In: Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, Association for Computational Linguistics, pp. 76–81 (2009)
11. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)