Evaluating *n*-gram Models for a Bilingual Word Sense Disambiguation Task

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

Facultad de Ciencias de la Computación Benemérita Universidad Autónoma of Puebla,
Puebla, Mexico
{dpinto,darnes,mtovar,bbeltran}@cs.buap.mx

Abstract. The problem of Word Sense Disambiguation (WSD) is about selecting the correct sense of an ambiguous word in a given context. However, even if the problem of WSD is difficult, when we consider its bilingual version, this problem becomes much more complex. In this case, it is necessary not only to find the correct translation, but such translation must consider the contextual senses of the original sentence (in the source language), in order to find the correct sense (in the target language) of the source word. In this paper we present a probabilistic model for bilingual WSD based on *n*-grams (2-grams, 3-grams, 5-grams and *k*-grams, for a sentence *S* of a length *k*). The aim is to analyze the behavior of the system with different representations of a given sentence containing an ambiguous word. We use a Naïve Bayes classifier for determining the probability of the target sense (in the target language) given a sentence which contains an ambiguous word (in the source language). For this purpose, we use a bilingual statistical dictionary, which is calculated with Giza++ by using the EUROPARL parallel corpus. On the average, the representation model based on 5-grams with mutual information demonstrated the best performance.

Keywords. Bilingual word sense disambiguation, machine translation, parallel corpus, Naïve Bayes classifier.

Evaluación de modelos de n-gramas para la tarea de desambiguación bilingüe del sentido de las palabras

Resumen. El problema de desambiguación del sentido de las palabras (WSD) consiste en seleccionar el sentido adecuado de una palabra polisémica, considerando el contexto en el que ésta se encuentra. Esta tarea se complica aún más cuando se desea desambiguar entre distintos idiomas; en el caso de dos idiomas, a este problema se le conoce como WSD bilingüe. Es necesario entonces no solamente encontrar la traducción correcta, sino también esta traducción debe considerar los sentidos de las palabras en el contexto de la oración original (en un idioma fuente), para encontrar el correcto sentido de la palabra ambigua (en un idioma destino). En este trabajo de investigación se presenta un modelo probabilístico para la desambiguación bilingüe basado en *n*-gramas (2-gramas, 3-gramas, 5-gramas y *k*-gramas, para una oración *S* de longitud *k*). El objetivo es analizar el comportamiento del sistema de desambiguación con diferentes representaciones de la oración que contiene la palabra ambigua. Para este propósito se usa el clasificador de Naïve Bayes para determinar la probabilidad de un sentido candidato (en un idioma destino), dada una oración que contiene la palabra ambigua (en un idioma fuente). Se emplea un diccionario estadístico bilingüe, el cual es calculado con el software Giza++ usando el corpus paralelo EUROPARL. Se evaluaron las diferentes representaciones llegando a la conclusión de que aquella basada en 5-gramas con esquema de filtrado por información mutua de bigramas ofrece el mejor valor de precisión.

Palabras clave. Desambiguación bilingüe del sentido de las palabras, traducción automática, corpus paralelo, clasificador de Naïve Bayes.

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 the automatic WSD continues to be an open problem has invoked a great interest in the computational linguistics community, therefore, many approaches has been introduced in the last years [1]. Competitions such as Senseval and recently SemEval[1] have also motivated the generation of new systems for WSD, providing an interesting environment for testing those

[1] http://www.senseval.org/; http://nlp.cs.swarthmore.edu/semeval/

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 [4], [3].

Monolingual word sense disambiguation is known to be a difficult task, however, when we consider its bilingual version, the problem becomes much more complex. In this case, it is necessary not only to find the correct translation, but such translation must consider the contextual senses of the original sentence (in the source language), in order to find the correct sense (in the target language) of the source word.

For the experiments reported in this paper, we 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 most of the WSD systems do. Instead, we try to find those senses automatically by means of a bilingual statistical dictionary which is calculated on the basis of the IBM-1 translation model[2] by using a filtered version of the EUROPARL parallel corpus[3]. This filtered version is obtained by selecting sentences of EUROPARL containing some of the ambiguous words of the test corpus.

The bilingual statistical dictionary is fed into a Naïve Bayes classifier in order to determine the probability of a target sense, given a source sentence (which contains the ambiguous word). Note that we do not use a training corpus of disambiguated words. Instead, we construct a classification model based on the probability of translating each ambiguous word (and those words that surround it). We are aware that other classification models exist such as CRF [8] and SVM [5]. However, since we have chosen a probabilistic model based on independent features (calculated by means of the IBM-1 translation model), in order to find the correct target sense, we believe that the Naïve Bayes classifier perfectly fits with this kind of approach.

The main aim of this research is to evaluate to what extent each word, in a neighborhood of the ambiguous word, contributes to improving the process of the bilingual WSD. We may hypothesize the following: “the closer a term is to the ambiguous word, the more it helps to find the correct target sense”. A natural document representation is the use of *n*-grams. In our work, we decided to evaluate six different approaches based on *n*-grams whose performance is further shown in this paper. A brief explanation of the general approach is as follows. Given a sentence *S*, we consider its representation by using one |*S*|-gram. We then propose the first approach considering the distance of each sentence term to the ambiguous word (weighted version). The second approach also uses one |*S*|-gram disregarding the distance of each term to the ambiguous word (unweighted version). The third approach considers the use of bigrams, i.e., a sequence of two terms containing the ambiguous word (a window size of 1 around the ambiguous word). The fourth approach uses 3-grams. The fifth and sixth approaches both use 5-grams; the former filters 5-grams using pointwise mutual information between each pair of terms of a 5-gram; the latter uses the student’s *t*-distribution in order to determine those bigrams that are likely to be a collocation, i.e., that they do not co-occur by chance. 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.

The rest of this paper is structured as follows. Section 2 presents some efforts reported in literature that we consider to be related with the present research. Section 3 introduces the problem of the bilingual word sense disambiguation. In Section 4 we define the probabilistic model used as a classifier for the bilingual WSD task. The experimental results obtained on the two datasets are shown in Section 5. Finally, the conclusions are given in Section 6.

[2] We used Giza++ (http://fjoch.com/GIZA++.html)
[3] http://www.statmt.org/europarl/

2 Related Work

The selection of the appropriate sense for a given ambiguous word is commonly carried out by considering the words surrounding the ambiguous word. A comprehensive 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 (for instance, see [6], [9], [11], [14]), 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.

There are other works described in literature in which parallel corpora (bilingual or multilingual) was used for dealing with the problem of WSD (for instance, see [12], [4]). Such approaches are expected to find the best sense in the same language (despite using other languages for training the learning model), however, in our research we are interested in finding the best translated word, i.e., the word with the correct sense in a different language.

3 Bilingual Word Sense Disambiguation

Word sense disambiguation is an important task in multilingual scenarios due to the fact that the meanings represented by an ambiguous word in the source language may be represented by multiple words in the target language. Consider the word “bank” which may have up to 42 different meanings[4]. Suppose we select one of these meanings, namely, “to 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 deposits her paycheck every month* (*Ella deposita su cheque cada mes*). Therefore, the ability for disambiguating a polysemous word in many languages is essential to the task of machine translation and to such a related Natural Language Processing (NLP) task as bilingual lexical substitution [13].

In the task of bilingual word sense disambiguation, we are required to obtain such translations of a given ambiguous word which match with the original word sense. As an example, let us consider the following sentence containing one polysemous word to be disambiguated. This sentence serves as the input, and the expected results are the following:

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 deal with this problem we propose to use a probabilistic model based on *n*-grams. This proposal is discussed in the following section.

4 A NaïveBayes Approachto Bilingual WSD

We 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 the target language), given a sentence (in the 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 EUROPARL parallel corpus. We filtered this corpus by selecting only sentences containing candidate senses of the ambiguous word (which

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

were obtained by translating the ambiguous word in the Google search engine).

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

4.1 The *n*-gram Model

In order to represent an input sentence we have considered a model based on *n*-grams. Remember, that we attempt to evaluate the degree of support in the process of bilingual disambiguation, for each word in a neighborhood of the ambiguous word. Thus, a natural document representation is the use of *n*-grams, and, therefore, each sentence is split into grams of *n* terms. In order to fully understand this process, let us consider the following example of the ambiguous word *execution*, and its pre-processed version which was obtained by eliminating punctuation symbols and stop words (no 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

In the experiments reported in this paper, we considered six different approaches, but only four types of *n*-grams (bigrams, 3-grams, 5-grams and the complete sentence, i.e., |*S*|-grams) which are described (including one example) as follows.

The representation of documents by means of bigrams is constructed by selecting sequences of two terms that sequentially co-occur in the sentence but considering that at least one of the terms is the ambiguous word. This consideration leads us to conform bigrams of terms in a neighborhood of window size one of the ambiguous word. For the example sentence presented above, we should obtain the following representation (two bigrams):

2-grams: {summary, *executions*}, {*executions,* robbing}

If we represent the sentence by using 3-grams, we must consider sequences of three terms containing the ambiguous word. For the same example sentence, we should get the following set of 3-grams:

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

In the case of representing the sentences by 5-grams, we should select sequences of five terms containing the ambiguous word, i.e., a window size of two around this word. The set of 5-grams for the same example sentence should be:

5-grams: {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}

Finally, if we consider the sentence *S* of example, we must consider all the terms inside it. The sentence representation by means of |*S*|-grams is as follows:

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

We experimented with two different *n*-grams filtering methods for the particular case of representing the sentences by 5-grams. Firstly, we discarded those bigrams belonging to the 5-grams that do not offer enough evidence of co-occurrence. For this purpose, we use the pointwise mutual information formulae which is presented in Eq. (1)

$$PMI(t_1, t_2) = log\left[\frac{N * freq(t_1 t_2)}{freq(t_1) * freq(t_2)}\right] \tag{1}$$

where the bigram t_1t_2, is the sequence of the two terms t_1 and t_2 which occurs in the 5-gram, *freq*(t_1) is the frequency of the term t_1 in the complete corpus, and *N* is the number of terms in the corpus.

The second method used for filtering the terms in the 5-grams is the student's *t*-distribution applied in order to eliminate those terms that co-occur by chance.

Given two terms t_1 and t_2 contained in one 5-gram, we considered the following hypotheses:

H0: P(t_1 t_2) = P(t_1) * P(t_2)
H1: P(t_1 t_2) > P(t_1) * P(t_2).

We assume that each term t_1 and t_2 was generated independently, therefore, the null hypothesis (H0) declares that the bigram t_1t_2 co-occur by chance, i.e., this bigram is not considered a collocation whereas the alternative hypothesis (H1) states that the bigram is in fact a collocation. The *t*-distribution is calculated as follows:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \qquad (2)$$

where

$$\bar{x} = P(t_1 t_2) = \frac{freq(t_1 t_2)}{N} \qquad (3)$$

and

$$\mu = P(t_1) * P(t_2) = \frac{freq(t_1)}{N} * \frac{freq(t_2)}{N} \qquad (4)$$

with *N* equal to the number of bigrams of the whole corpus. If the *t* value is greater than *2.576*, then we reject the null hypothesis with a significance level *α= 0.005*.

The representation of sentences by means of *n*-grams aims at studying the impact of the terms that co-occur with the ambiguous word. We believe that the importance of such terms should be emphasized and they should be employed in the process of bilingual word sense disambiguation. For each *n*-gram sentence representation proposed, we obtain a candidate set of translations for the source ambiguous word by applying a probabilistic model on the basis of the *n*-grams selected. Further details are explained in the next section.

4.2 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, …, w_k, …, w_{k+1}, …, $w_{|S|}$} be the *n*-gram representation of S_E constructed by bringing together all the *n*-grams (w_k is the ambiguous word). Let us consider *N* candidate translations of w_k, $\{t_1^k, t_2^k, \dots, t_N^k\}$ obtained somehow (we will discuss this issue later in this section). We are interested in finding the most likely candidate translations for the polysemous word w_k. Therefore, we may use a Naïve Bayes classifier which takes into account the probability of t_1^k given w_k. A formal description of the classifier is given as follows.

$$p(t_i^k|S) = p(t_i^k|w_1, w_2, \dots, w_k, \dots) \qquad (5)$$

$$p(t_i^k|w_1, w_2, \dots, w_k, \dots) = \frac{p(t_i^k)p(w_1, w_2, \dots, w_k, \dots|t_i^k)}{p(w_1, w_2, \dots, w_k, \dots)} \qquad (6)$$

We are looking for the argument that maximizes $p(t_i^k|S)$, therefore, the process of calculating the denominator may be avoided. Moreover, if we assume that all the different translations are equally distributed, then Eq. (6) must be approximated by Eq. (7).

$$p(t_i^k|w_1, w_2, \dots, w_k, \dots) \approx p(w_1, w_2, \dots, w_k, \dots|t_i^k) \qquad (7)$$

The complete calculation of Eq. (7) requires applying the chain rule. However, if it is assumed that the words of the sentence are independent, then Eq. (7) may be rewritten as Eq. (8).

$$p(t_i^k|w_1, w_2, \dots, w_k, \dots) \approx \prod_{j=1}^{|S|} p(w_j|t_i^k) \qquad (8)$$

The best translation is obtained as shown in Eq. (9). Irrespective of the position of the ambiguous word, we consider only the product of the probabilities of translation. Algorithm 1 provides details of implementation.

$$BestSense_w(S) = \text{arg } max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k) \quad (9)$$

where $i = 1, \ldots, N$.

An alternative approach (*the weighted version*) is proposed as well and shown in Eq. (10). The aim of this approach is to verify whether it is possible to obtain a better performance in the bilingual word sense disambiguation task when the distance of each term to the ambiguous word in the probabilistic model is considered. Algorithm 2 provides details about implementation.

$$BestSense_w(S) = \text{arg } max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k) * \frac{1}{k-j+1} \quad (10)$$

with i = 1, …, N.

We have used the Google translator[5] in order to obtain the N candidate translations of the polysemous word w_k, $\{t_1^k, t_2^k, \ldots, t_N^k\}$. Google provides all the possible translations for w_k with the corresponding grammatical category. Therefore, we are able to use 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 would have confirmed that by using the Google online translator we obtain the best performance. 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 on the EUROPARL parallel corpus.

```
Input: A set $Q$ of sentences: $Q = \{S_1, S_2, \ldots\}$;
Dictionary = $p(w|t)$: A bilingual statistical dictionary;
Output: The best word/sense for each ambiguous word $w_j \in S_1$.
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
                    $P_{l,i} = P_{l,i} * \varepsilon$;
                end
            end
        end
    end
end
return $\text{arg } max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k)$
```

Algorithm 1. A Naïve Bayes approach to bilingual WSD

```
Input: A set $Q$ of sentences: $Q = \{S_1, S_2, \ldots\}$;
Dictionary = $p(w|t)$: A bilingual statistical dictionary;
Output: The best word/sense for each ambiguous word $w_j \in S_1$.
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
                    $P_{l,i} = P_{l,i} * \varepsilon$;
                end
            end
        end
    end
end
return $arg\ max_{t_i^k} \prod_{j=1}^{|S|} p(w_j|t_i^k) * \frac{1}{k-j+1}$
```

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

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

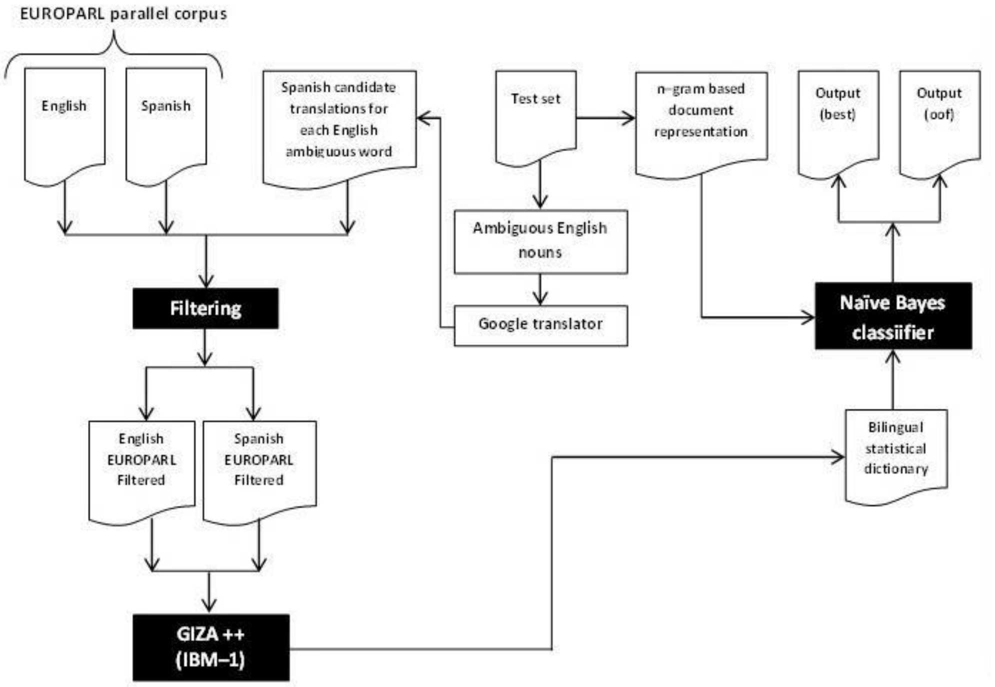

EUROPARL parallel corpus
English
Spanish
Spanish candidate translations for each English ambiguous word
Test set
n-gram based document representation
Output (best)
Output (oof)
Ambiguous English nouns
Google translator
Filtering
Naïve Bayes classiifier
English EUROPARL Filtered
Spanish EUROPARL Filtered
Bilingual statistical dictionary
GIZA ++ (IBM–1)

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

To summarize the above said, we have proposed a probabilistic model (see Figure 1) that uses a statistical bilingual dictionary, constructed with the IBM-1 translation model, on the basis of a filtered parallel corpus extracted from EUROPARL. This corpus includes Spanish translations of each different English ambiguous word. The probability of translation of each ambiguous word is then modeled by means of the Naïve Bayes classifier in order to further classify the original ambiguous words with the purpose of finding the correct translation in the Spanish language. As it may be seen, we do not use any training corpus of disambiguated words.

5 Experimental Results

The results of experiments with different sentence representations based on *n*-grams for bilingual word sense disambiguation are given in this section. First, we describe the corpus used in the experiments and further on, we present the evaluation of the six different approaches based on n-grams.

5.1 Datasets

In our experiments, 25 polysemous English nouns were used. 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 test corpus, 50 instances per noun were used. The list of the ambiguous nouns in the test corpus may be seen in Table 1. Notice that this corpus is not a sense repository, because the task requires finding the most probable translation (with the correct sense) of a given ambiguous word.

5.2 Evaluation of the *n*-gram Based Sentence Representation

In Figure 2, one can view the results obtained for the different approaches evaluated with the corpus presented in Table 1. The runs are labeled as follows:

2-grams: A representation of the sentence based on bigrams.

3-grams: A representation of the sentence based on trigrams.

tStudent 5-grams: A representation of the sentence based on 5-grams, removing all those bigrams with are not considered to be a collocation by means of the student's *t*-distribution.

PMI 5-grams: A representation of the sentence based on 5-grams, removing all those bigrams which are not considered to be a collocation by means of pointwise mutual information.

unweighted |*S*|-gram: A sentence representation based on a unique n-gram of length |*S*|.

weighted |*S*|-gram: A sentence representation based on a unique n-gram of length |*S*|, considering the distance of each sentence term to the ambiguous word.

The bigram model showed the worst performance. We think that the fact of using only one term (besides the ambiguous one) in the disambiguation model is responsible for this failure. Thus, the information needed in order to disambiguate the polysemous word is not sufficient. It may be seen that the model based on a 3-gram representation outperformed the bigram one, but the number the terms around the ambiguous word is still insufficient. With these results in mind, we proposed to use a representation with a greater number of terms (in this case, 5-grams were used). This representation model was analyzed with the purpose of detecting those bigrams, inside the 5-grams, that are actual collocations and not co-occur by chance. Therefore, we proposed two different filtering methods: pointwise mutual information and Student's *t*-distribution. The former filtering method obtained the best results. The reason is that PMI does not need so many occurrences of the bigram that the Student's *t*-distribution does in order to detect that a given bigram is in fact a collocation.

Finally, when we considered all the terms for the process of disambiguation (|*S*|−*gram*), we observed that some terms were positioned too far from the ambiguous word to provide valuable information. Actually, such terms introduce noise making the performance of the method to decrease. In the latter representation, we were interested in finding out whether the closeness of the terms in the sentence with respect to the ambiguous word had a positive impact on the process of disambiguation. Therefore, we proposed a weighted version of the representation model which gives less importance to those terms that are far and more importance to closer terms.

Unfortunately, the formulae did not give enough weight for emphasizing this characteristic. That is one of the reasons why the 5-gram representation reached a better performance. In other words, the 5-gram representation uses only the necessary terms and assigns a higher value of importance to all of them if they are closer to the ambiguous word.

With the purpose of observing the performance of the proposed approaches, Table 2 presents a comparison of our runs with others approaches presented at the SemEval-2[6] competition. The *UvT* team submitted two runs (UvT-WSD1 and UvT-WSD2) with an *oof* evaluation, which outputs the five best translations/senses. This team made use of a *k*-nearest neighbor classifier to build one word sense for each target ambiguous word, and selected translations from a bilingual dictionary obtained by executing the GIZA package on the EUROPARL parallel corpus [10].

The University of Heidelberg participated submitting two runs (UHD-1 and UHD-2). They approached the bilingual word sense disambiguation by finding the most appropriate translation in different languages on the basis of a multilingual co-occurrence graph automatically induced from the target words aligned with the texts found in the EUROPARL and JRC-Arquis parallel corpora [10].

Finally, there was another team which submitted one run (ColEur2) with a supervised approach using the translations obtained with GIZA from the EUROPARL parallel corpus in order to distinguish between senses in the English source sentences [10]. In general, it may be seen that all the teams used the GIZA software in order to build a bilingual statistical dictionary. Therefore, the main differences among all these approaches are in the way of representing the original ambiguous sentence (including the pre-processing stage), and the manner of filtering the results obtained by GIZA.

Table 2 is given only as a reference of the behavior of our approaches with respect to those presented in the literature. However, we must emphasize that these results are not comparable because the teams participating at the SemEval-2 competition were allowed to repeat the target translation/sense among the five possible outputs. This type of evaluation leads to higher performance (even greater than 100%) compared with the case when it is not allowed to repeat translations. Despite the unfair comparison, it can be seen that the approach named PMI 5-*gram* outperforms the best result obtained in the competition.

In Table 3, we compare our approaches which allow the repetition of translations. Again, it may be noticed that some of our approaches perform better than some other systems.

By observing the values of precision over the different ambiguous words (see Figure 3), we may have a picture of the significant level of improvement that may be reached when representing the sentence with 5-grams. In particular, we present the approach that filtered the terms using pointwise mutual information and obtained the best results over all the approaches analyzed. In Figure 3, it may also be seen that there are some words that are easier to disambiguate (e.g. *soil* and *education*) than others (e.g. *match*). For research purposes, we also consider it important to focus on those words that are hard to disambiguate.

Table 2. Evaluation of the bilingual WSD (removing repeated translations/senses); Five best translations (oof)

System name	Precision (%)	Recall (%)
PMI 5-gram	43.26	43.26
UvT-WSD2	43.12	43.12
UvT-WSD1	42.17	42.17
unweighted |S|-gram	40.82	40.82
weighted |S|-gram	40.76	40.76
UHD-1	38.78	31.81
UHD-2	37.74	31.3
3-gram	36.82	36.82
ColEur2	35.84	35.46
tStudent 5-gram	33.52	33.52
2-gram	21.25	21.25

6 Conclusions

Bilingual word sense disambiguation is the task of obtaining such translations of a given ambiguous word that match with the original word sense. In this paper, we presented an evaluation of different representations based on *n*-grams for sentences containing one ambiguous word. In particular, we used a Naïve Bayes classifier for determining the probability of a target sense (in the target language) given a

[6] http://semeval2.fbk.eu/

sentence which contains the ambiguous word (in the source language). The probabilities were modeled by means of a bilingual statistical dictionary calculated with Giza++ (using the EUROPARL parallel corpus). Six different approaches based on *n*-grams were evaluated. The 5-gram representation that employed mutual information demonstrated the best performance, slightly outperforming the results reported in the literature for the bilingual word sense disambiguation task at the SemEval-2 international competition.

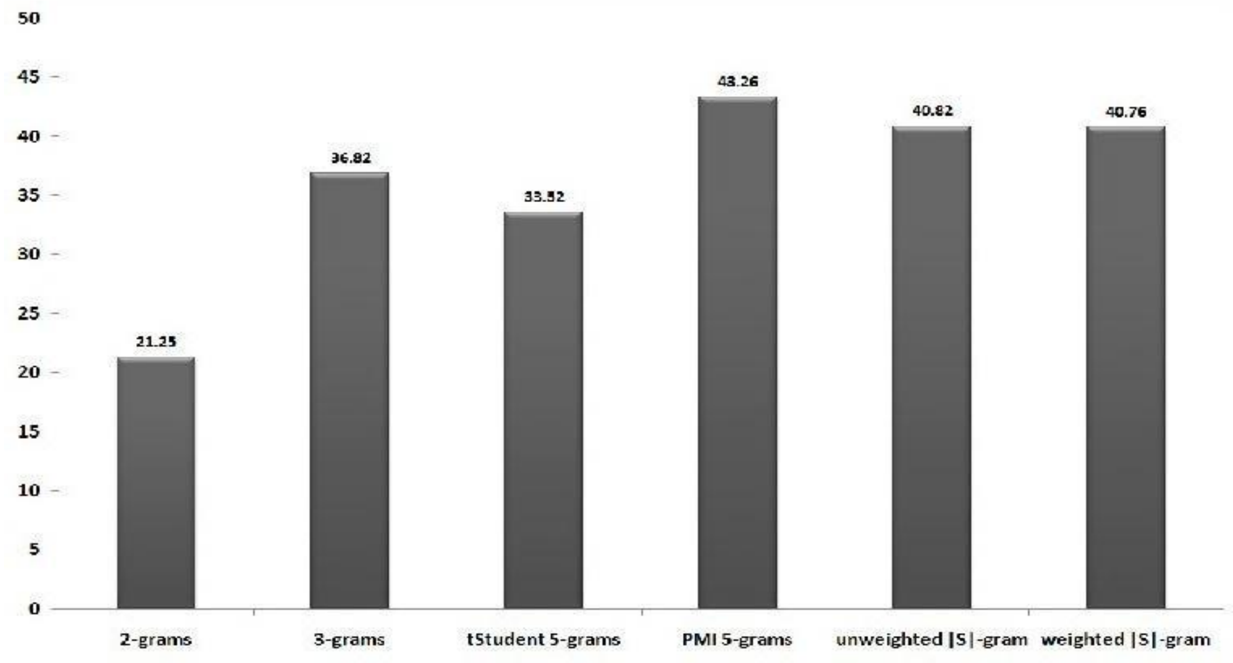


Fig. 2. A comparison among all the approaches proposed

Table 3. Evaluation of the bilingual WSD (considering repeated translations/senses); five best translations (oof)

System name	Precision (%)	Recall (%)
3-gram	70.36	70.36
PMI 5-gram	54.87	54.87
UvT-WSD2	43.12	43.12
UvT-WSD1	42.17	42.17
unweighted |S|-gram	40.76	40.76
UHD-1	38.78	31.81
weighted |S|-gram	38.46	38.46
UHD-2	37.74	31.3
ColEur2	35.84	35.46
tStudent 5-gram	33.52	33.52
2-gram	21.25	21.25

Adding a filtering step by means of pointwise mutual information allowed us to identify the terms which give the best support to the process of bilingual WSD.

We observed that in some cases the use of a reduced window size in the neighborhood of the ambiguous word may exclude some important terms that would help to improve the precision of finding the correct target sense. This leads us to conclude that statistical methods do have some limitations but they may be enriched by considering the use of linguistic and/or semantic techniques able to capture those terms.

Finally, we consider that the hypothesis of Harris [7] which states that the closer the words are to the polysemous word, the better they serve for disambiguating the polysemous word, although at the same time it is important to avoid Type I errors or "false positives" by using some techniques like pointwise mutual information.

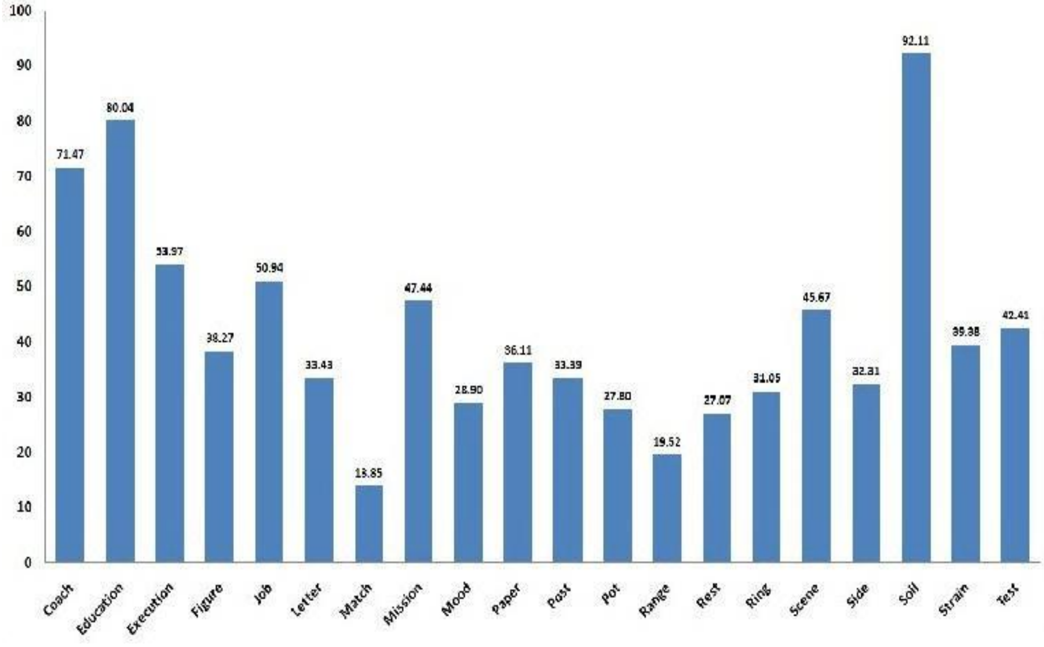


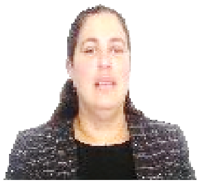
Fig. 3. An analysis of the evaluation of all the ambiguous words with the *PMI* 5-*grams* approach

References

1. **Aguirre, E. & Edmonds, P. (2006).** *Word Sense Disambiguation: algorithms and applications*. Dordrecht: Springer.
2. **Barceló, G., (2010).** *Desambiguación de los sentidos de las palabras en español usando textos paralelos*. Tesis de Doctorado, Instituto Politécnico Nacional, Centro de Investigación en Computación, México, D.F.
3. **Carpuat, M. & Wu, D. (2007).** Improving statistical machine translation using word sense disambiguation. *2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (*EMNLP-CoNLL 2007*). Prague, Czech Republic, 61-72.
4. **Chan, Y., Ng, H. & Chiang, D. (2007).** Word sense disambiguation improves statistical machine translation. *45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 33-40.
5. **Cortes, C. & Vapnik, V. (1995).** Support-vector networks. *Machine Learning,* 20 (3), 273–297.
6. **Florian, R. & Yarowsky, D**. **(2002).** Modeling consensus: Classifier combination for word sense disambiguation. *ACL-02 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, USA, 10, 25–32.
7. **Harris, Z. (1981).** Distributional structure. In Henry Hiz (Ed.), *Papers on syntax* (3–22). Boston: Kluwer Boston Inc.

8. **Lafferty, J.D., McCallum, A. & Pereira, F.C.N**. **(2001).** Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Eighteenth International Conference on Machine Learning, ICML '01*. Massachusetts, USA, 282–289.
9. **Lee, Y.K. & Ng, H.T. (2002).** An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. *ACL-02 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, USA, 10, 41–48.
10. **Lefever, E. & Hoste, V. (2010).** Semeval-2010 task 3: Cross-lingual word sense disambiguation. *NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions.* Colorado, USA, 82–87.
11. **Mihalcea, R.F. & Moldovan, D.I. (2001).** Pattern learning and active feature selection for word sense disambiguation. *Second International Workshop on Evaluating Word Sense Disambiguation Systems* (*SENSEVAL-2*). Toulouse, France, 127–130.
12. **Ng, H. T., Wang, B. & Chan, Y. S. (2003).** Exploiting parallel texts for word sense disambiguation: An empirical study. *41st Annual Meeting of the Association for Computational Linguistics (ACL'03)*. Sapporo, Japan, 455–462.
13. **Sinha, R., McCarthy, D. & Mihalcea, R. (2010).** Semeval-2010 task 2: Cross-lingual lexical substitution. *NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Colorado, USA, 76–81.
14. **Yarowsky, D., Cucerzan, S., Florian, R., Schafer, C. & Wicentowski, R**. **(2001).** The Johns Hopkins Senseval2 system descriptions. *Second International Workshop on Evaluating Word Sense Disambiguation Systems* (*SENSEVAL-2*). Toulouse, France, 163–166.

David Eduardo Pinto Avendaño obtained his PhD in computer science in the area of artificial intelligence and pattern recognition at the Polytechnic University of Valencia, Spain in 2008. At present he is a full time professor at the Faculty of Computer Science of the Benemérita Universidad Autónoma de Puebla (BUAP). His areas of interest include clustering, information retrieval, crosslingual NLP tasks and computational linguistics in general.



Darnes Vilariño Ayala obtained her PhD in mathematics in the area of optimization at the Havana's University of Cuba in 1997. At present she is a full time professor at the Faculty of Computer Science of the BUAP. Her areas of interest include artificial intelligence, business intelligence and computational linguistics.

Carlos Balderas is currently a master student at the Faculty of Computer Science of BUAP. His areas of interest include information retrieval and word sense disambiguation.

Mireya Tovar Vidal obtained her master degree in computer science at the Cinvestav - IPN in 2002. She is currently a PhD student at the CENIDET research institute. She is also a full time professor at the Faculty of Computer Science of BUAP. Her areas of interest include ontologies and computational linguistics.

Beatriz Beltrán Martínez obtained her master degree in computer science in 1997 at the Faculty of Computer Science of BUAP where she holds now a position of a full time professor. Her areas of interest include pattern recognition and computational linguistics.

Article received on 12/03/2010; accepted 05/02/2011.