Choosing the Best Dictionary for Cross-Lingual Word Sense Disambiguation

Andres Duque*, Juan Martinez-Romo, Lourdes Araujo

NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos.
Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain

Abstract

The choice of the dictionary that provides the possible translations a system has to choose when performing Cross-Lingual Word Sense Disambiguation (CLWSD) is one of the most important steps in such a task. In this work, we present a comparison between different dictionaries, in two different frameworks. First of all, a technique for analysing the potential results of an ideal system using those dictionaries is developed. The second framework considers the particular unsupervised CLWSD system CO-Graph, and analyses the results obtained when using different bilingual dictionaries providing the potential translations. Two different CLWSD tasks from the 2010 and 2013 SemEval competitions are used for evaluation, and statistics from the words in the test datasets of those competitions are studied. The conclusions of the analysis of dictionaries on a particular system lead us to a proposal that substantially improves the results obtained in that framework. In this proposal a hybrid system is developed, by combining the results provided by a probabilistic dictionary, and those obtained with a Most Frequent Sense (MFS) approach. The hybrid approach also outperforms the results obtained by other unsupervised systems in the considered competitions.

Keywords: Word Sense Disambiguation, Bilingual Dictionary, Natural Language Processing

*Corresponding author
Email addresses: aduque@lsi.uned.es (Andres Duque*), juaner@lsi.uned.es (Juan Martinez-Romo), lurdes@lsi.uned.es (Lourdes Araujo)

1. Introduction

Cross-Lingual Word Sense Disambiguation (CLWSD) can be defined as the task of automatically determining the contextually appropriate translation for a given word, from a source language to a target one. This is a particular case of the Word Sense Disambiguation (WSD) problem, which has been widely studied in the NLP community (Ide and Veronis, 1998). WSD is an essential and necessary step for many processes, such as automatic summarization, information retrieval, topic detection, and in general, any NLP process in which the semantic level of the words is important. WSD has been frequently treated as a supervised learning problem (Màrquez et al., 2006; Mihalcea, 2006), based on techniques that depend on semantically tagged corpora or lexical databases like Wordnet (Fellbaum, 1998). On the other hand, unsupervised techniques, also known as Word Sense Induction (WSI) techniques, do not require those kinds of resources. Their objective is to induce the different senses of a specific word in a given text by selecting groups of words related to a particular sense of the word. The motivation of the CLWSD task comes from the scarcity of sense inventories and sense-tagged corpora, and the need to evaluate the performance of WSD systems in real problems (Lefever and Hoste, 2010b).

A Cross-Lingual Word Sense Disambiguation task proposes a set of instances in which a target word can be found. This target word needs to be disambiguated, from an original language (typically English) to a final one. Figure 1 illustrates this task with an example. The bilingual dictionary that provides translations, both for words surrounding the target word (context) and for the target word itself, is a key part of the disambiguation process. This dictionary offers the potential translations of the target word, and any system which performs the disambiguation has to choose, among the translations, those which are considered most suitable for the particular sentence. This selection is then matched against an expected output or gold standard to determine a score for that specific test instance. In this example, the context taken into account for performing the disambiguation is only composed by nouns, although any other word (e.g. verbs, adjectives) can also be considered.

Many issues arise along the disambiguation process, the choice of an adequate bilingual dictionary being one of the most important for ensuring the good performance of a system. We compare the use of bilingual dictionaries of different nature: manually created by experts, semi-automatic, i.e. extracted with automatic tool but with human supervision or intervention, collaboratively edited by different authors, and statistical dictionaries. This last type of dictionaries,

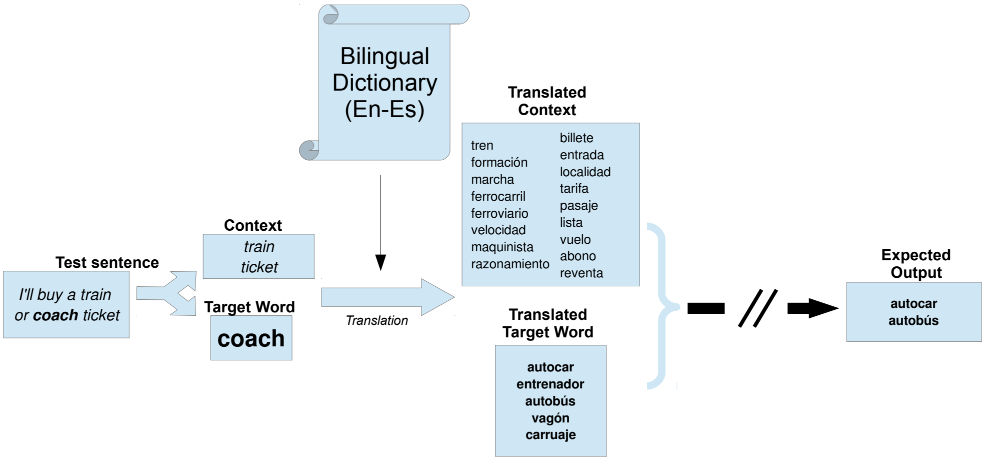

Figure 1: Example of a general disambiguation process of a sentence containing the target word coach, with Spanish as target language.

automatically created without human supervision, provide a much larger number of translations, at the price of introducing noise. However, apart from their size and the coverage they can present (denoted by the number of different translations for each word), this kind of dictionaries provide information about the translation probabilities, since their construction is based on statistical characteristics. The other dictionaries do not usually present this kind of information. Considering that CLWSD tasks are based on translations of words used in general sentences, we can expect that information about the most frequent translations would be useful.

In this work, we analyse different dictionaries that provide the candidate translations, and compare the results obtained using them, both in ideal conditions, and inside a particular unsupervised CLWSD system (Duque Fernandez et al., 2013). These results show the potential variations of the effectiveness of the CLWSD system according to the choice of the bilingual dictionary.

1.1. Background Work

For the purposes of this work, we have selected some evaluation tasks related to Cross-Lingual Word Sense Disambiguation, as a framework in which the effect of the selected dictionary can be tested. Specifically, we have selected task 3 of

2010 SemEval competition (Lefever and Hoste, 2010b) and task 10 of 2013 SemEval competition (Lefever and Hoste, 2013), both of them based on the Europarl parallel corpus (Koehn, 2005). Many different systems were proposed for these two tasks, and the use of bilingual dictionaries is a common practice inside the proposed algorithms, both for supervised and unsupervised systems. The OWNS system (Mahapatra et al., 2010) is a supervised system which participated in the 2010 SemEval competition. It uses nearest neighbors classifiers based on pairwise similarity measures. Most of its lexical information is extracted from WordNet (Fellbaum, 1998), although it uses a noisy statistical dictionary learnt from the Europarl corpus for proposing possible translations. Other supervised methods also participated in the 2010 competition: UvT-WSD (van Gompel, 2010), applying the K-NN algorithm, and FCC (Vilariño et al., 2010), using a Naive Bayes classifier. In those cases, the tool used for extracting bilingual dictionaries was GIZA++ (Och and Ney, 2003), which has proven to be the preferred tool for aligning the corpus at word level and extracting translations. Regarding unsupervised systems participating in the 2010 competition, in (Silberer and Ponzetto, 2010), a co-occurrence graph based on the aligned contexts of the target word is built for performing the disambiguation. This graph aggregates words from different languages and the disambiguation is made through the extraction of the minimum spanning tree. In this work, multilingual dictionaries such as EuroWordNet (Vossen, 1998), and PanDictionary (Mausam et al., 2009) are proposed for extracting translations, frequencies and characteristics. The other unsupervised system of the 2010 competition, T3-COLEUR (Guo and Diab, 2010) is based on probability tables extracted from the Europarl corpus, and also uses a GIZA-based bilingual dictionary. In this competition, the best results for the Spanish language were obtained by the supervised system UvT-WSD, while the best unsupervised system was T3-COLEUR.

In regard to the 2013 competition, the only system that did not make use of the GIZA++ tool was the supervised system HLDTI (Rudnick et al., 2013). It used maximum entropy classifiers, trained on local context features, to perform the disambiguation, and the aligning tool selected for extracting translations was the Berkeley Aligner (DeNero and Klein, 2007). The other systems of this competition used GIZA-based dictionaries, independently of the final languages of the translations. In this group, we can find supervised systems such as WSD2 (van Gompel and van den Bosch, 2013), the new version of the UvT-WSD also based on a K-NN classifier. Unsupervised systems also used this resource: LIMSI (Apidianaki, 2013) addressed the problem by using vectors of features extracted from the corpus. XLING (Tan and Bond, 2013) generated topic models from the source corpus using Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The main hypothesis is that the different senses of a target word will be classified into different topics by the LDA algorithm. The NRC-SMT system (Carpuat, 2013) uses a statistical machine translation approach, extracting knowledge only from the Europarl corpus in its first run, and adding information from news data in a second run of the system. In the 2013 competition also a supervised system, HLDTI, obtained the best results. The best unsupervised system was LIMSI.

Finally, we can find other systems that did not participate in any of the competitions, although they present results for some of the proposed datasets: the ParaSense system (Lefever et al., 2011) is a supervised, memory-based algorithm that builds different classifiers using both local context features and binary bag-of-words features. Unsupervised systems as the multilingual system described in (Navigli and Ponzetto, 2012) also addressed the problem without participating in the competitions. This system exploits the multilingual knowledge base BabelNet (Navigli and Ponzetto, 2010), for performing WSD and CLWSD, obtaining very competitive results. Both works make use of the GIZA++ tool, the first one as a main aligner for extracting a bilingual dictionary, and the second one for proposing the most frequent sense translations when no sense assignment is attempted.

1.2. Main Objectives

In this work we analyse the effect of bilingual dictionaries, both inside an ideal system, and a particular CLWSD system, named CO-Graph. This system is based on an unsupervised algorithm for extracting co-occurrence graphs from text documents (Martinez-Romo et al., 2011). In this case, we focus on the English-Spanish cross-lingual disambiguation, and on the out-of-five evaluation proposed in both SemEval tasks already mentioned. This evaluation scheme requires the systems to provide up to five guesses for each target word in each context, without penalising them due to the number of guesses.

The first objective of this work is the design of some experiments to compare different dictionaries in a general framework of a disambiguation task. For this purpose, we have developed a frame in which theoretical limits can be found for the performance of each of the analysed dictionaries for well defined CLWSD tasks. Once that we find these limits (upper bounds), we intend to explore the actual performance of a particular CLWSD system in the task, and analyse its results depending on the dictionary. Finally, on the basis of our observations, we develop a technique which combines the information provided by different sources. This technique has allowed us to outperform other unsupervised systems taking part in the SemEval competitions. Interestingly, this technique is not only valid for a particular system, but for any unsupervised system using weights for selecting the most appropriate translations.

The rest of the paper is organized as follows: section 2 describes the main characteristics of the CO-Graph system, which is used through the rest of the work to compare the different dictionaries. Section 3 explains in detail the different considered dictionaries. The characteristics of the evaluation framework used for testing the dictionaries are shown in section 4. Section 5 analyses the results that could be achieved by an ideal system, depending on the bilingual dictionary used. An error analysis concerning those results is conducted in section 6. In Section 7, the dictionaries are tested within the CO-Graph system previously mentioned, and the obtained results are analysed. Section 8 describes the development of a combined approach that can improve the previous results. Finally, conclusions and future work are detailed in section 9.

2. CLWSD System Description: CO-Graph

In this section we describe the main characteristics of the particular unsupervised CLWSD system used for the disambiguation, namely CO-Graph. In this system we need to select the five most suitable translations given each context, according to the SemEval 2010 and 2013 evaluation framework. The base of knowledge for all the steps of the disambiguation system is the Europarl parallel corpus (Koehn, 2005), which was compiled and sentence-aligned from the proceedings of the European Parliament between 1996 and 2011. Although the corpus is presented in many languages, for this particular work we focus on the English-Spanish translation.

The whole disambiguation system is composed of several steps. A test instance can be divided into the target word to be disambiguated and the context (rest of words in the test sentence). Using a bilingual dictionary, we translate the target word and the context words. From the corpus written in Spanish, we extract a co-occurrence graph, and then a community graph is generated from this graph. This community graph links clouds of words, each one of them containing related words, in terms of co-occurrence. After this step, the translations of words (in this case, nouns) of the context and the potential translations of the target word are found inside the community graph, and the distances between communities containing translations of the target word and communities containing translations of words of the context are calculated. Finally, the scores of the possible translations are ranked in order to select the five most suitable translations for the target word in this particular context. Figure 2 illustrates the complete CO-Graph system, with all its phases: the extraction of words from the test instance, the translation of those words, the construction of the co-occurrence graph and the community graph, and finally the disambiguation step, involving the community graph and the translated words. In later subsections each step of the process will be described in detail.

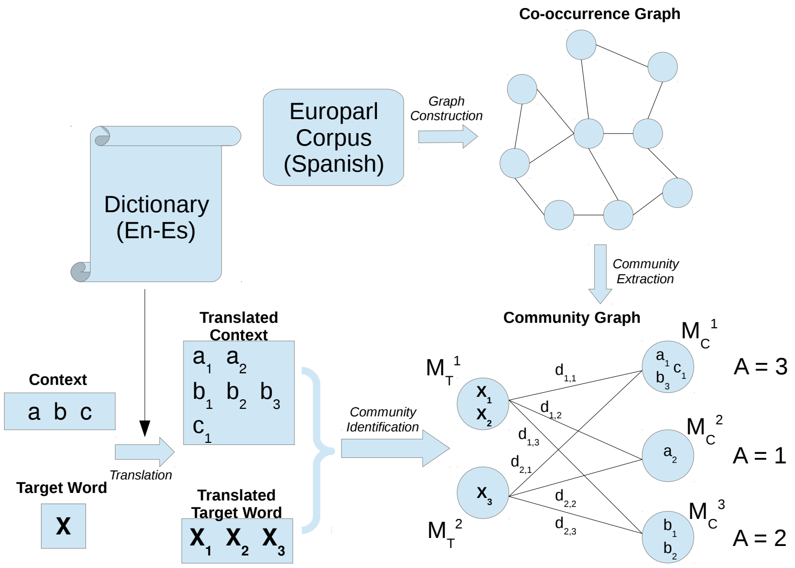

Figure 2: Diagram and example of the CLWSD system. The community graph is extracted from the co-occurrence graph, and used to compute the distances between words from the context and the target word. Communities named with "M_T" contain translations of the target word, and communities named with "M_C" contain translations of the words of the context. The letter "A" represents the number of translations from words of the context that can be found in each of the "M_C" communities.

2.1. *Co-occurrence Graph Construction*

The system used for disambiguation is based on a co-occurrence graph. The main hypothesis for building the graph considers a document to be a coherent piece of information, and thus words in a document tend to (statistically) adopt a related sense. However, this is not exactly true, since some words may appear in a document without really being related to its main sense. So we only consider co-occurrences that are statistically significant (do not happen by chance).

First steps for building the graph require part-of-speech tagging of the documents, which is done with the TreeTagger tool (Schmid, 1994), and selection of the words that will become part of the graph. In this work, we only use nouns as nodes of the graph, which are linked depending of the importance of their co-occurrences.

From the tagged documents written in the final language (in this case Spanish), we are now able to build the co-occurrence graph. In order to check if the co-occurrence of two words in the same document is statistically significant, a null model is defined, that represents what is considered pure chance. In this null model, for each pair of words co-occurring k times, a random and independent distribution is generated among a set of documents, and the probability of those two words co-occurring k times by pure chance is calculated. Then, given two words randomly and independently distributed among N documents, and appearing in n_1 and n_2 documents respectively, the probability of those words co-occurring in exactly k documents is given by:

$$p(k) = \frac{\binom{N}{k}\binom{N-k}{n_1-k}\binom{N-n_1}{n_2-k}}{\binom{N}{n_1}\binom{N}{n_2}} \tag{1}$$

if $\max\{0, n_1 + n_2 - N\} \leq k \leq \min\{n_1, n_2\}$, and zero otherwise. The denominator, formed by $\binom{N}{n_1}\binom{N}{n_2}$, represents the number of ways in which the subsets n_1 and n_2 of documents can be randomly and independently selected from the complete collection, which contains N documents. On the other hand, considering k as the number of coincidences between the first and second selection of sets, in the numerator we represent how many of the choices of n_1 and n_2 contain exactly k coincidences. We have four different types of documents: k documents containing co-occurrences of the words, $n_1 - k$ documents containing only the first word, $n_2 - k$ containing only the second word, and $N-k-(n_1-k)-(n_2-k) = N-n_1-n_2+k$ documents not containing any of the words. The number of choices then is represented by the multinomial coefficient $\binom{N}{k, n_1-k, n_2-k}$, that can be also expressed as shown by the numerator of Formula 1.

If this probability is high, that is, if the null model can easily generate the co-occurrences between this specific pair of words, then it is not considered to be statistically significant, and hence no link is created between those words in the graph. More specifically, a p-value p is calculated for the co-occurrence of two words inside the null model. If $p \ll 1$ (lies below a given threshold next to 0), then the appearance of the two words in a document is significant (their meaning is probably related). Moreover, we can quantify this significance by taking the median (corresponding to $p = 1/2$) as a reference, and hence, the weight of a link established between two words inside the graph is $\ell = -\log(2p)$, that is, a measurement of the deviation from the median. The p-value can be seen as a measurement of the restrictiveness of the graph, since we can establish a minimum value for p, above which no link is created between two words. Therefore, as the threshold for p decreases, the graph becomes more restrictive. This threshold value is a parameter to be set in the experiments.

2.2. Community Extraction

After the co-occurrence graph is built, we need to define a way to perform the final disambiguation of a word. The construction of the co-occurrence graph gives us a structured representation of the information inside the corpus. We now need to select from the graph those nodes closely related that can be interpreted as a specific sense. The technique used for this step is named "community detection": A community is a sub-graph whose nodes present some kind of structural or dynamic affinity. In this technique, we assume that words belonging to the same community have a related sense, different from those represented by other communities. There exist many different community extraction algorithms. In this work, we use the *Walktrap* algorithm (Pons and Latapy, 2005). This proposal is based on the fact that a random walker that jumps between nodes inside the graph, gets more easily trapped in those parts of the graph (sub-graphs) that are densely connected. These sub-graphs would then become the desired communities. The distance between two nodes is small, and hence they belong to the same community, if the accessibility of any third node is somewhat similar from any of the two nodes. The algorithm then generalizes to a more coarse-grained structure, for performing a community merging phase. In this phase distance between communities, instead of nodes, is computed for selecting those adjacent communities that can be merged due to their proximity.

With the communities obtained by the algorithm, we build a new graph, called community graph (CG). In this graph, each community is represented by a node, and an edge is added linking communities C_1 and C_2 if and only if any word $x \in C_1$ is linked in the co-occurrence graph to any word $y \in C_2$. It is important to notice that the input of the Walktrap algorithm must be a connected graph, that is, a graph without isolated vertices. For ensuring this, the giant component of the co-occurrence graph is used as input of the Walktrap algorithm. This fact also guarantees the connectivity of the community graph, and hence the distance between any two communities can be calculated.

2.3. Disambiguation

For every test instance that contains a target word to be disambiguated and a set of words surrounding it (context), we perform a part-of-speech tagging using the TreeTagger tool (Schmid, 1994). Then, we select the nouns and obtain their translations, according to the bilingual dictionary that we have selected. After that, we need to identify, inside the community graph CG, those communities that contain at least one of the translations, either from words of the context or from the target word. As a result, we obtain two sets of communities: the set M_T includes communities that contain at least one translation from the target word, and the set M_C is composed by communities containing at least one translation from any word of the context. Through the community graph we can calculate the distances between any community $M_C^i \in M_C$ and any community $M_T^j \in M_T$. Since a translation of a target word can belong to the same community that a translation of a context word ($M_C^i = M_T^j$), the distance in that case would be 1, which is the minimum distance we consider. In any other case, we add the number of links in the shortest path between M_C^i and M_T^j. Hence, if the path between M_C^i and M_T^j contains one link, the distance between them, for our purposes, would be 2, if the path contains 2 links, the distance would be 3, and so on.

Our hypothesis for this algorithm is that the translation of the target word that is nearer (in average) to the translations of the context words, is more likely to be the most suitable one for that target word in that context. Hence, we establish a formula for ranking the potential translations of the target word, based on two factors: the score of a translation is inversely proportional to the distance between the community to which it belongs and any community containing context translations, but directly proportional to the number of context translations inside the community. Thus, the weight or score of a translation of the target word, w_t, is given by:

$$w_t = \max_{M_C^i \in M_C} \frac{A_C^i}{(d_{M_C^i M_T^t} + 1)} \tag{2}$$

where A_C^i is the number of context translations inside M_C^i, and $d_{M_C^i M_T^t}$ is the distance (number of steps) between M_C^i and M_T^t, that is, the community in which translation t is located. By ranking the scores of all the possible translations for the target word given by the dictionary, the system can propose the most suitable ones as a solution.

3. Bilingual Dictionaries

In this section, we present the main characteristics of the different considered bilingual dictionaries. They are four English-Spanish dictionaries: a manually created dictionary, built by experts, which will be denoted as "external dictionary" along the rest of the paper, a collaboratively edited dictionary, a semi-automatic dictionary and a statistical, automatically created parallel corpus-based dictionary. All of them are described below.

- **External dictionary**: This dictionary, described in (López-Ostenero, 2002) is completely external to the main task. It is a generic bilingual dictionary, which has no relation to the source of knowledge in the task (the Europarl corpus). The results offered by this dictionary, both for the ideal system and for the CO-Graph system, are considered to be a baseline for this work, and hence a goal of the other dictionaries is to overcome those results.

- **MCR dictionary**: The Multilingual Central Repository (Atserias et al., 2004) is a lexical knowledge base (LKB) that constitutes a multilingual large scale linguistic resource for many semantic processes, due to the amount of multilingual knowledge that it contains (Agirre and Soroa, 2008). This LKB contains lexical information about five different languages: English, Spanish, Basque, Catalan and Galician, and is based on the WordNet and EuroWordNet projects. Synsets from different languages are linked through the Inter-Lingual Indices (ILIs). From the ILIs present in MCR 3.0 (Gonzalez-Agirre et al., 2012), we have extracted direct translations from English to Spanish to create our bilingual dictionary.

- **BabelNet dictionary**: BabelNet (Navigli and Ponzetto, 2010) is a very large semantic multilingual network that links Wikipedia information to WordNet synsets in an automatic way. The whole resource could be considered as a semi-automatic dictionary, since multilingual information comprises both manual translations from Wikipedia, and translations obtained by applying machine translation to the SemCor corpus (Miller et al., 1993). For any word in the English language, we can obtain all the possible senses of the word, and their corresponding translations in the final language (in our case, Spanish).

- **GIZA++ dictionary**: The statistical aligner GIZA++ is able to extract one-to-many translations from a target word and their corresponding probabilities of occurrence. For this aim, it uses a parallel corpus as knowledge base,

> in our case the Europarl corpus. In the first step, the GIZA++ tool performs a word alignment over the initial corpus, without any preprocessing. Once that the alignment is done, we obtain a probability table. This table links every word in the original language (in this case, English) to each of its possible translations in the final language (in this case, Spanish), and assigns a probability of occurrence. Due to the automatic and statistical nature of the algorithm implemented by GIZA++, the number of translations that it proposes for each English word is very high. This fact may introduce noise in the translation process so a technique to reduce this inducted noise and thereby improve the accuracy is needed. For this purpose, we performed the alignment in the other direction, i.e., obtaining a one-to-many word alignment from Spanish to English, and then calculated the intersection of both probability tables. In this way, we obtain an English-Spanish dictionary, ensuring that every English-Spanish translation has an equivalent Spanish-English translation. We have excluded stop words for building the dictionary.

Table 1 shows some statistics about the dictionaries used in this work. Specifically, we can observe the number of entries, maximum number of translations presented by a word, and the average number of translations for all the words in the dictionary.

	Entries	**Max # translations**	**Average # translations**
External	50,911	87	2.32
MCR	35,440	56	2.09
BabelNet	384,832	89	2.62
GIZA++	34,815	1,344	7.51

Table 1: Statistics from the bilingual dictionaries. Column "Entries" represents the total number of entries of the dictionary. Column "Max # translations" shows the maximum number of translations for a word. Column "Average # translations" shows the average number of translations in the complete dictionary.

Regarding the number of entries in the dictionary, we can observe that the BabelNet dictionary presents many more words than any other dictionary. This can be due to the completeness of the dictionary, which can be considered more as a encyclopaedic dictionary, since not only synsets from WordNet, but also entities from Wikipedia, are collected to build the dictionary. However, the total number of entries is not important for this work, given that all the words in the test sentences are covered by all the dictionaries. The average number of translations is a more important fact when we want to analyse the impact of each dictionary. In this case, we can observe that most of the dictionaries offer an average number of translations between 2 and 3. Nevertheless, the GIZA++ dictionary offers many more translations per word than the other dictionaries. This can lead to a wider coverage of the problem. On the other hand, and regarding a real system, this fact may imply a drawback, considering that a high number of possible translations for a target word could prevent the system from finding the most suitable ones. That is, the coverage would be high, but the precision may decrease.

4. Datasets and Evaluation

The evaluation setting adopted in our experiments is based on the one proposed in task 3 of SemEval 2010 and task 10 of SemEval 2013 competitions. Evaluation is carried out, in both tasks, over a test dataset with 20 different words and 50 sentences for each of them. The gold standard used for evaluating the participant systems is built from the Europarl corpus, proposed as knowledge base. For this purpose, a word-level alignment was performed and manually evaluated for all the sentences of the corpus containing target words. After that, a manual clustering by meaning was carried out, for every target word. The output of this process was a sense inventory (Lefever and Hoste, 2010a). Annotators of the gold standard used the clustered sense inventory for selecting the most appropriate translations of each target word. The translations are weighted depending on how many annotators selected each of them. Example 3 shows the gold standard for the Spanish language provided by the annotators for a given sentence in which we can find the target word "***coach***".

(3) SENTENCE 2: *A branch line train took us to Aubagne where a* ***coach*** *picked us up for the journey up to the camp.*

coach.n.es 2 :: autocar 3;autobus 3;diligencia 1;

In the evaluation scheme, called "out-of-five" evaluation, the system has to select five of the potential translations for each test instance.

We use the F-Measure value for illustrating the results achieved, and for comparing them with other systems participating in the SemEval competitions.

Regarding the datasets, Table 2 offers more information about the statistics of the dictionaries, focused on the 20 words composing the datasets. More specifically, it shows the number of translations offered by each dictionary for each possible target word in the test datasets.

Word	**External**	**MCR**	**BabelNet**	**GIZA++**
coach	15	13	27	**8**
education	6	4	10	52
execution	4	6	14	30
figure	29	25	25	146
job	17	14	28	133
letter	**3**	4	6	46
match	15	26	18	101
mission	6	7	8	35
mood	4	**3**	**4**	32
paper	10	8	12	64
post	30	21	11	72
pot	**43**	**41**	**80**	21
range	25	17	30	100
rest	22	11	13	87
ring	31	13	21	34
scene	15	9	19	46
side	19	15	26	**191**
soil	10	5	10	10
strain	31	13	32	48
test	15	7	7	89
Mean	**17.50**	**13.10**	**20.05**	**67.25**

Table 2: Number of translations of the words in the datasets, for each dictionary: External (second column), MCR (third column), BabelNet (fourth column) and GIZA++ (fifth column). Bold represents maximum and minimum values for each dictionary.

The table clearly shows the differences in number of translations for each target word depending on the bilingual dictionary. We can observe that the external dictionary, the dictionary based on MCR and the dictionary based on BabelNet present similar behaviour. In the three cases, the word which presents the highest number of translations is “pot”, while the word “mood” presents the lowest number of translations for the MCR and BabelNet dictionary, and the second lowest for the external dictionary. On the other hand, the behaviour of the GIZA++ dictionary is completely different, as the word presenting the highest number of translations is “side” and the word presenting the lowest number is “coach”. These differences can be due to the automatic nature of the dictionary generated with GIZA++. The other dictionaries present human intervention in their construction, which can lead to a different number of translations. Apart from this fact, it is important to notice the high number of possible translations produced in the GIZA++ dictionary, which may lead to decrease the performance. To avoid this decrease, we also considered a restricted GIZA-based dictionary, with a maximum of ten possible translations per word. These ten translations are those that present the highest probabilities of occurrence. Some experiments regarding the value of maximum translations per word have been done, showing that a pruning value of ten translations per word provides the best results. This dictionary will be denoted as “GIZA10” along the rest of the paper.

5. Analysing the Influence of the Dictionaries on an Ideal System

A good indicator for understanding how the dictionary can modify the performance of a system in a CLWSD task is the highest score that could be achieved by a perfect system for a given dictionary. In this particular case, we define the upper bound for a given dictionary as the best result that a system that uses this dictionary can achieve, according to the gold standard. Since we are working with datasets from two past competitions, we have access to the gold standards used for the evaluation. Then, for building the best guessing that a system could give, we take for every context of every target word those translations from the dictionary that are also in the solution provided by the gold standard. If there are words in the gold standard for this context that are not present in the dictionary, random words are selected to complete the requested five word guessing. In the proposed dictionaries we do not take into account those translations that contain more than one word.

Tables 3 and 4 show the highest precision, for each word in average, that can be achieved by any system using the five considered dictionaries. The last column represents an upper bound obtained by applying the same process to the gold standard itself, but excluding from the proposed solution those translations containing more than one word, since the co-occurrence graph used in CO-Graph only considers one-word translations (nodes of the graph represent one single word). Specifically, table 3 shows the results for the 2010 test set, and table 4 the results for the 2013 test set.

The dictionary obtained with GIZA++ and without restrictions (Column **GIZA**) is the resource that would allow an ideal system to obtain the best results. How-

Upper Bounds 2010						
Word	**ExtDic**	**MCR**	**BabelNet**	**GIZA10**	**GIZA**	**Gold**
coach	63.17	58.31	**76.89**	**76.89**	**76.89**	96.60
education	77.82	77.82	80.88	84.13	**94.00**	98.19
execution	53.26	53.26	62.94	67.77	**80.00**	89.35
figure	46.97	44.27	49.02	62.63	**84.90**	95.03
job	54.38	31.55	53.01	61.58	**74.10**	83.02
letter	37.51	37.51	40.94	42.68	**57.66**	93.19
match	46.74	55.79	55.79	26.41	**71.80**	99.71
mission	55.06	55.06	55.06	56.19	**76.12**	99.18
mood	14.20	23.27	26.42	62.32	**68.97**	77.64
paper	39.45	25.41	28.08	43.33	**64.92**	97.69
post	47.27	37.28	**49.46**	16.94	39.30	83.57
pot	**55.15**	32.37	45.57	38.60	48.71	89.70
range	17.66	15.15	21.29	17.96	**45.44**	84.77
rest	30.90	33.27	34.85	26.08	**36.48**	89.73
ring	42.04	29.00	30.49	50.65	**66.86**	98.83
scene	42.46	42.46	46.88	61.44	**80.86**	90.08
side	40.55	33.26	36.30	43.28	**70.43**	84.98
soil	63.06	63.06	73.69	**98.07**	**98.07**	99.27
strain	26.55	26.55	39.02	67.07	**83.17**	93.41
test	68.92	59.11	66.38	80.20	**87.00**	95.22
Mean	**46.16**	**41.69**	**48.65**	**54.22**	**70.28**	**91.97**

Table 3: Upper bounds (F-Measure in %) for SemEval 2010 test dataset, obtained with different translation dictionaries: external dictionary (column **ExtDic**), dictionary based on the Multilingual Central Repository (column **MCR**), BabelNet-based dictionary (column **BabelNet**), complete GIZA++ dictionary (column **GIZA**) and pruned GIZA++ dictionary (column **GIZA10**). Last column represents results obtained by the gold standard without considering multi-word translations. Bold represents best results for each word without taking the gold standard into account.

ever, due to the noise that the high number of translations of the dictionary induces, in the rest of the work we will use GIZA10. In the tables we can also observe that the dataset for 2013 ideally allows the systems to achieve better results, as the upper bounds are higher in all cases. The last column, representing the modified gold standard (without translations containing more than one word), gets close to a perfect performance. However, its accuracy is not 100% due to the men-

Upper Bounds 2013						
Word	**ExtDic**	**MCR**	**BabelNet**	**GIZA10**	**GIZA**	**Gold**
coach	76.50	73.53	**83.83**	**83.83**	**83.83**	100.00
education	77.17	76.83	75.34	83.83	**88.98**	92.67
execution	50.29	50.29	65.53	61.48	**75.81**	86.68
figure	57.49	52.77	56.00	69.55	**88.83**	99.53
job	66.93	40.54	56.99	63.51	**76.54**	84.34
letter	59.06	59.06	60.21	62.00	**76.49**	97.23
match	48.63	50.17	50.17	23.20	**76.67**	95.03
mission	71.78	71.78	71.78	78.99	**92.06**	100.00
mood	25.03	29.20	34.20	67.78	**74.28**	80.00
paper	65.47	52.79	54.54	65.23	**77.33**	99.71
post	**76.90**	59.15	65.89	34.67	48.68	96.99
pot	**58.97**	29.67	55.47	26.37	29.20	82.80
range	28.64	21.75	26.19	21.30	**50.31**	87.98
rest	35.19	39.14	42.87	25.78	**40.30**	91.08
ring	69.37	53.36	54.65	59.86	**72.23**	100.00
scene	42.67	42.67	51.00	65.94	**86.06**	90.69
side	53.75	47.03	48.27	59.62	**80.65**	93.70
soil	76.81	76.81	86.49	**96.60**	**96.60**	100.00
strain	27.40	27.40	44.44	63.66	**86.30**	94.32
test	74.55	65.29	71.66	76.21	**81.19**	91.96
Mean	**57.13**	**50.96**	**57.78**	**59.47**	**74.12**	**93.24**

Table 4: Upper bounds (F-Measure in %) for SemEval 2013 test dataset, obtained with different translation dictionaries: external dictionary (column **ExtDic**), dictionary based on the Multilingual Central Repository (column **MCR**), BabelNet-based dictionary (column **BabelNet**), complete GIZA++ dictionary (column **GIZA**) and pruned GIZA++ dictionary (column **GIZA10**). Last column represents results obtained by the gold standard without considering multi-word translations. Bold represents best results for each word without taking the gold standard into account.

tioned exclusion of multi-word translations. Hence, it provides some clues about the reduction of accuracy due to this exclusion. There are some words for which the external dictionary obtains a higher upper bound than the GIZA++ dictionary ("*post*" and "*pot*"). This may be due to the specific characteristics of those words (number of translations, differences between translations, ...). Overall, most of the words present significant potential improvements in their upper bounds when we use the GIZA++ dictionary. A deeper analysis regarding the words which present better performance with the other dictionaries is done in section 6. Comparing Tables 3 and 4 with Tables 1 and 2 we observe a direct correlation between the translation average in a dictionary and the performance (average F-Measure) of an ideal system using that dictionary. As the upper bounds are basically representing the coverage of each dictionary (the maximum performance that could be achieved), this correlation is expected: as the number of possible translations increases, the probability of covering more words from the gold standard is higher, and hence the ideal performance of the system also increases.

Figure 3 shows an example of the process of construction of the upper bounds for any dictionary. Given a sentence and its gold standard, we extract from the dictionary those words (highlighted in bold letters in the example) that appear in the gold standard. The rest of the words, until five, are randomly selected from those proposed by the dictionary. In the example, the external, MCR and BabelNet dictionaries contain two words appearing in the gold standard ("escena" and "panorama"). On the other hand, the GIZA10 dictionary contains three coincident words ("ambito", "escena" and "panorama"). Hence, an ideal system based on GIZA would obtain a better result for this particular instance.

6. Error Analysis

In this section we intend to analyse in detail the results offered by Tables 3 and 4. In particular, we want to focus on the results obtained by the ideal system using the GIZA10 dictionary. We can observe in the tables that there are some words for which other dictionaries ideally outperform the GIZA10 approach. We analyse the translation probabilities provided by this dictionary in order to look for possible explanations of this issue. Table 5 contains the number of translations of each word in the complete GIZA++ dictionary. After pruning the dictionary and obtaining the GIZA10 dictionary, we calculate the mean and standard deviation of the translation probabilities for each target word.

We focus on those words for which other dictionaries (external, MCR-based or BabelNet-based) obtain better results for ideal systems, in both test datasets (SemEval 2010 and SemEval 2013). Those words are "match", "post", "pot", "range" and "rest". We can observe that four of those five words (excluding "post") present low mean (around 0.1) and low standard deviation (below 0.18). These facts (specially the low standard deviation) indicate that most of the translations have similar probability of occurrence, that is, the distribution adopts similar values. Hence, it is more likely that some important translations that also have a similar proba-

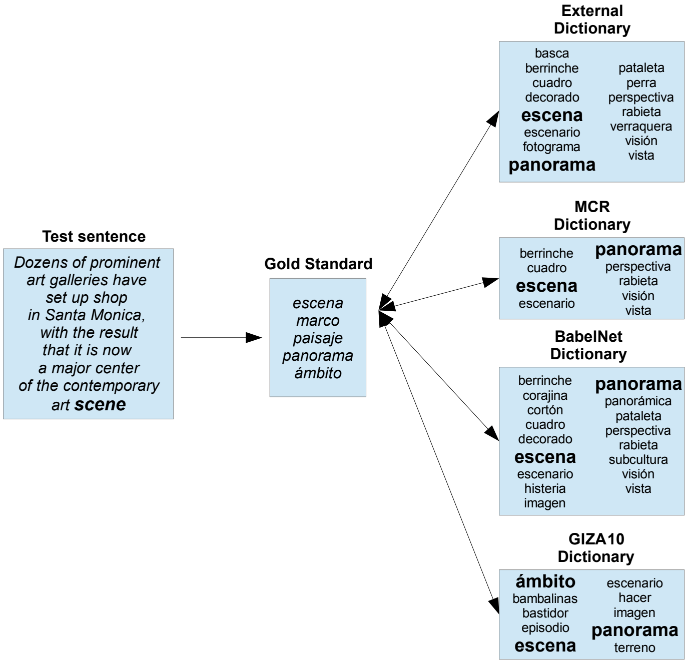

Figure 3: Example of the construction of the upper bounds for the considered dictionaries.

bility of occurrence, although slightly smaller, were discarded when the GIZA++ dictionary was pruned. Other words that present similar characteristics, such as “ring”, also present worse performance in ideal systems using GIZA10, but only in one of the test datasets (in this case, SemEval 2013).

7. Dictionary Comparison on a Particular Unsupervised CLWSD System: CO-Graph

Once we have compared the behaviour of different dictionaries inside an ideal system, we want to consider those dictionaries inside the specific unsupervised CLWSD system described in section 2. As it is stated before, the unsupervised graph construction algorithm on which the system relies depends on an initial threshold value for the p-value p. This threshold has to be determined in order to indicate the highest value of p for which the number of co-occurrences of two words is considered to be statistically significant and therefore a link is created

Word	**Trans. (GIZA)**	**Mean (GIZA10)**	**SD (GIZA10)**
coach	8	0.136	0.255
education	52	0.093	0.206
execution	30	0.104	0.301
figure	146	0.100	0.225
job	133	0.142	0.203
letter	46	0.106	0.240
match	**101**	**0.098**	**0.174**
mission	35	0.186	0.379
mood	32	0.118	0.087
paper	64	0.158	0.219
post	**72**	**0.129**	**0.219**
pot	**21**	**0.114**	**0.145**
range	**100**	**0.103**	**0.088**
rest	**87**	**0.071**	**0.162**
ring	34	0.116	0.134
scene	46	0.111	0.129
side	191	0.105	0.171
soil	10	0.126	0.295
strain	48	0.109	0.057
test	89	0.094	0.184

Table 5: Statistics for translations of words in the datasets. Second column contains the number of translations, third column the mean of the translation probabilities of the ten most probable translations, and fourth column the standard deviation of the same ten translations. Bold represents words for which the GIZA10 approach does not overcome the other dictionaries in neither SemEval test dataset (2010 nor 2013).

between them.

In previous experiments, we used the trial dataset provided in the SemEval 2010 competition for analysing the influence of the threshold in an exhaustive way. We varied the threshold from $p = 10^{-5}$ to $p = 10^{-17}$ and obtained F-Measure results for all the p-values. Higher values of p lead to huge graphs that usually become unmanageable in terms of time and memory consumption. As the threshold decreases, the graph becomes more restrictive, and hence presents less nodes and less edges linking nodes. This restrictiveness of the graph can lead to better results, as we gain some specificity, but when the graph becomes too restrictive the performance of the algorithm may decrease. In those previous experiments, best results were achieved with values of the threshold between $p = 10^{-5}$ and $p = 10^{-11}$. Within this smaller range of thresholds, the F-Measure values are quite similar. According to this, we have selected a threshold value of $p = 10^{-11}$ for all the experiments in this section. This value is inside the range that offers the best results, and allows us to deal with a smaller graph. By selecting a fixed threshold, we want to test the robustness of our system under the same conditions that the systems participating in the SemEval competitions. This selection of a specific value for all the cases eliminates the risk of overfitting, since known gold standard data are not used for adjusting parameters.

Since we are performing a comparison between systems, it is useful to consider a baseline for studying whether the proposed systems are able to outperform it. We take as a baseline the results obtained by a system that would return the five most frequent translations for the target word, according to the GIZA++ dictionary. This approach will be denoted as Most Frequent Sense or MFS along the rest of the paper.

Table 6 shows the performance achieved by CO-Graph, using the different considered dictionaries for both the 2010 and 2013 test datasets. It also contains the results obtained with the MFS approach, for the same datasets.

Competition	**ExtDic**	**MCR**	**BabelNet**	**GIZA10**	**MFS**
SemEval 2010	37.04	33.94	34.60	42.03	44.02
SemEval 2013	43.87	41.35	38.95	47.06	49.75

Table 6: Results (F-Measure in %) obtained over 2010 and 2013 SemEval test datasets, for the out-of-five evaluation. Columns 2 to 5 contain the results achieved by the CO-Graph system when using the different bilingual dictionaries (external, MCR-based, BabelNet-based and GIZA++ pruned to ten translations per word). Last column represents the results obtained by the MFS (Most Frequent Sense) approach.

The results clearly show, on one hand, that the test dataset for the 2013 competition allows the system to obtain a higher performance. This is basically due to the use of the same words as in the 2010 competition, but modifying the contexts for evaluation. As we can observe, all approaches improve their performance from 2010 to 2013. On the other hand, we can observe that, as we expected, the use of the GIZA10 dictionary, allows the system to improve the results, when compared to those obtained with the other three dictionaries. We observe that the F-Measure achieved by the system using a particular dictionary is directly proportional to the average number of translations for each word in the dictionary, in a similar way to what happened with the ideal systems. As we stated above, we performed different tests regarding the pruning value of the GIZA++ dictionary, observing that when more than 10 words were used as maximum number of translations for each word, the performance of the system decreased. Hence, the key point of pruning the GIZA++ dictionary is to find a large enough maximum number of translations (coverage of the problem) that does not introduce too much noise into the system. Table 6 shows that the value of 10 translations per word offers good results. Since we select those translations with highest probability of occurrence, the overall performance of the system is better than that achieved when using the MCR-based dictionary for instance, a dictionary that uses a similar (average) number of translations for the target words in the datasets (see Table 2). Still, the Most Frequent Sense technique outperforms any of the proposed approaches. This fact indicates that when more than five translations are considered, the system does not effectively choose the most suitable ones. However, a deeper analysis of the F-Measure per individual word indicates that there are words for which GIZA++ outperforms the results of the MFS approach. Hence, a good step at this point would be the development of a hybrid system that combines the translations proposed by the MFS approach, and those proposed by the system.

8. Hybrid Approach

As we can confirm in Table 6, the Most Frequent Sense (MFS) is a baseline that has been proved difficult to overcome in many CLWSD tasks, including those under analysis in this work. Moreover, these tasks use a MFS approach based on a specific corpus used to represent knowledge, and hence its performance is even better than a MFS approach based on a more generalist corpus. The MFS can be extracted in an automatic way with the GIZA++ tool and has a different nature than the weights assigned by the CO-Graph system to each translation. This fact can be used in a hybrid approach for enriching the information given by the disambiguation algorithm. Hence, the combination of the weights given by the system and the probabilities given by GIZA++ may offer better results than those obtained by the original approach of our system and may also overcome the MFS approach. The intuition behind this hybrid approach is based on what we stated in section 6: when the values of the probabilities of translations from a target word are quite different (their standard deviation is high), CO-Graph is able to obtain a good performance, both in cases in which selecting the most frequent senses offer good results, and in cases in which the best translations do not present the highest probabilities. However, when this standard deviation of the probabilities is low, that is, when the distribution tends to be flat, CO-Graph can get lost, and hence the MFS information obtained from GIZA can be very useful. According to this intuition, our approach combines, for every potential translation, the weight according to CO-Graph, and the probability of translation, according to GIZA++. A final score is then assigned to each of the ten potential translations provided by the dictionary. Specifically, we consider $T = (t_1, t_2, ..., t_n)$ to be the set of most likely potential translations provided by GIZA++ for a given target word, where $n <= 10$. Each translation t_i has an associated probability p_i. After applying the disambiguation process, CO-Graph assigns a weight w_i to each of the potential translations. The final score of each translation s_i, which will be used for selecting the most appropriate translations for evaluation, is given by $s_i = p_i w_i$. Figure 4 shows an example of the behaviour of the hybrid approach.

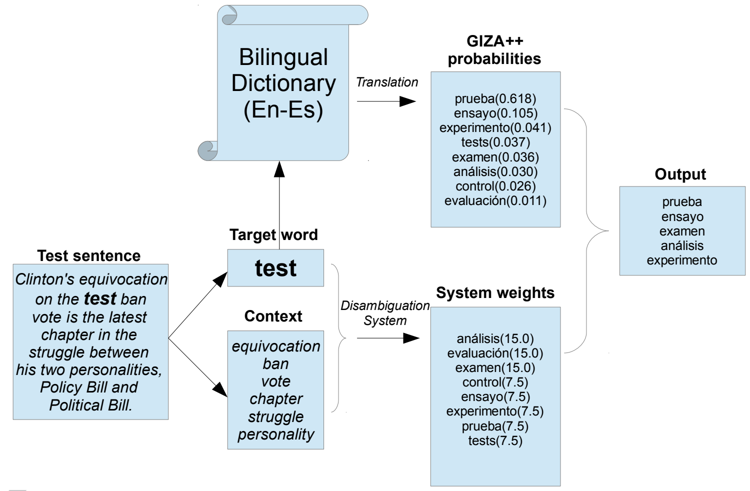

Figure 4: Example of the behaviour of the hybrid system which combines the output of CO-Graph with the translation probabilities given by GIZA++.

Table 7 completes Table 6 with the results, for the 2010 and 2013 datasets, obtained by the proposed hybrid approach.

The new column with respect to Table 6, **Hybrid**, contains the results for both datasets using the hybrid approach. We can observe that, for both datasets, the hybrid approach gets better results than the MFS approach. The performance of the system increases about 3.5 points for the 2010 and 2013 datasets. Moreover, the improvement over the system that uses the pruned GIZA dictionary is more

Data	ExtDic	MCR	BabelNet	GIZA10	MFS	Hybrid
2010	37.04	33.94	34.60	42.03	44.02	**47.41**
2013	43.87	41.35	38.95	47.06	49.75	**53.33**

Table 7: Final results (F-Measure in %) obtained over 2010 and 2013 test datasets.

than 5 points in the 2010 dataset and more than 6 points in the 2013 dataset. A two-tailed paired t-test for statistical significance testing has been performed over the results in the table. According to this test, the results obtained by the hybrid approach are significantly better than those obtained by the system using only any of the bilingual dictionaries. Also, in the 2010 dataset, the differences between the hybrid approach and the MFS approach are statistically significant, whereas in the 2013 dataset, although the results are also better, the significance is not achieved.

As we have stated along the whole paper, we have used only nouns for building the co-occurrence graph of the CO-Graph system, and for extracting the context of the target word in each test sentence. A last experiment was conducted for testing whether the selection of other important category of words, in this case verbs, could improve the overall performance of the system. For this purpose, a new co-occurrence graph that also considered verbs was built, and the disambiguation process was repeated for all the test instances, extracting also verbs from the context. Table 8 shows the comparative between results obtained by the system using only the best dictionary (GIZA10) and by the hybrid system, both using only nouns and using nouns and verbs for building the graph and extracting the context.

As we can observe, the inclusion of verbs in the construction of the graph does not improve the results. Including new words in the graph may lead to bigger, more difficult to handle graphs, and hence to more difficulties in the disambiguation process. Also, it is important to indicate that most of the target words in the test instances can be translated as nouns. Therefore, the increase of coverage that could be achieved by including verbs in the translations may not compensate the probable loss of precision due to the need of dealing with bigger graphs.

Tables 9 and 10 show the comparison between results obtained using the hybrid approach, and those obtained by other unsupervised systems participating in the 2010 (Table 9) and 2013 (Table 10) CLWSD competitions. The results ob-

	Words	GIZA10	Hybrid
SemEval 2010	**Nouns**	**42.03**	**47.41**
	Nouns+Verbs	34.70	45.00
SemEval 2013	**Nouns**	**47.06**	**53.33**
	Nouns+Verbs	39.58	51.33

Table 8: Comparative between results obtained by the best performing configurations of the system (GIZA10 and Hybrid), using only nouns for building the graph and extracting the context, and using nouns and verbs for these processes. Results (F-Measure in %) for the 2010 and 2013 test datasets.

tained by the best participating system (even if supervised) are also shown, as well as the baselines proposed in the competitions.

System	Task 3 SemEval 2010
Best	43.12
Hybrid	**47.41**
T3-COLEUR	35.65
UHD-1	34.95
UHD-2	34.22
Baseline	48.41

Table 9: Comparison of the F-Measure (%) achieved by the unsupervised systems participating in task 3 of SemEval 2010, and by the hybrid approach of our system (row **Hybrid**). The best participating system (even if supervised) is shown in row **Best**, while the baseline proposed by the organizers is shown at the bottom of the table, in row **Baseline**.

As we can observe, in both cases the hybrid approach outperforms the results obtained by other unsupervised systems. More specifically, the unsupervised systems in the 2010 task were the T3-COLEUR system, based on probability tables, and the UHD system, also based on co-occurrence graphs, but with different techniques for extracting the knowledge from the graph to perform the disambiguation. In the 2010 competition, we can also see that the best participating system (supervised) is also outperformed by the hybrid system. However, the baseline proposed by the organizers is still the best "system" in the task. We consider that

System	Task 10 SemEval 2013
Best	61.69
Hybrid	**53.33**
LIMSI	49.01
XLING snt	44.83
XLING merged	43.76
XLING tnt	39.52
NRC-SMT adapt2	41.65
NRC-SMT basic	37.98
Baseline	53.07

Table 10: Comparison of the F-Measure (%) achieved by the unsupervised systems participating in task 10 of SemEval 2013, and by the hybrid approach of our system (row **Hybrid**). The best participating system (even if supervised) is shown in row **Best**, while the baseline proposed by the organizers is shown at the bottom of the table, in row **Baseline**.

this baseline provided by the organizers must be an unrealistic approach to the problem, since not even supervised techniques are able to outperform it. In 2013, the unsupervised participants were the vector-based LIMSI system, the XLING system, using topic modelling techniques, and the NRC system, based on a statistical machine translation approach. Regarding this dataset, we observe that the best (supervised) system is better than our hybrid approach. In this case, the proposed baseline is outperformed by our system, but not by any of the unsupervised systems that participated in the competition.

9. Conclusions and Future Work

We have analysed the effect of the translation dictionary in the performance of a Cross-Lingual Word Sense Disambiguation system. The results obtained within an ideal framework indicate that when the dictionary is generated in a statistical automatic way from a corpus large enough to represent the characteristics of a language, the potential results for a disambiguation task are better. The best ideal results are achieved when considering all the possible translations obtained. However, this induces too much noise. Accordingly, the number of potential translations for each word has been pruned to the ten most probable ones for building the GIZA10 dictionary. For some target words, the other dictionaries are able to outperform the results obtained by an ideal GIZA-based dictionary. This fact can be due to the nature of the set of translations that GIZA extracts for each word: when too many translations are extracted, and their probabilities are similar, the coverage that can be achieved by a dictionary containing ten translations per word can be compromised. Nevertheless, the GIZA10 dictionary has been shown to be the best dictionary in ideal conditions. This selection has been confirmed using a particular CLWSD system. CO-Graph has been tested over the four different dictionaries, and the results have been compared to those obtained by a Most Frequent Sense (MFS) baseline. In this case, the GIZA10 dictionary has also proven to be the best choice among the analysed dictionaries for solving the CLWSD tasks. However, the MFS approach still outperforms its results. Considering this fact, and the unsupervised nature of the MFS approach, a hybrid approach has been built, using outputs from both CO-Graph and the MFS approach. The results obtained by this hybrid approach outperform the MFS reference baseline, and the other unsupervised systems participating in the 2010 and 2013 CLWSD competitions from SemEval. The main conclusion is that statistical information related to the possible translations of the target words, is a key knowledge for systems performing CLWSD. Accordingly, this way of selecting the candidate translations can be considered as one of the best options for unsupervised CLWSD systems.

Future work includes the refinement of the hybrid system by modifying the formula that determines the final score of each potential translation. Also, a deeper exploration of the dictionary extracted with GIZA++ is needed, in order to include more possible translations that would ideally allow the system to reach higher accuracy, as Tables 3 and 4, and section 6 suggest. Following this intuition, a good approach could be not restricting the number of translations per word to a fixed value, but varying that value depending on the statistical characteristics of the translations. The use of multi-word translations could improve the upper bounds of the dictionary, and hence the final results obtained by a CLWSD system. Finally, more work needs to be done in order to expand this work to other languages.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the project EXTRECM (TIN2013-46616-C2-2-R), and the Universidad Nacional de Educación a Distancia (UNED) (2013-025-UNED-PROY).

Agirre, E., Soroa, A., 2008. Using the multilingual central repository for graph-based word sense disambiguation. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco, http://www.lrec-conf.org/proceedings/lrec2008/.

Apidianaki, M., 2013. LIMSI: Cross-lingual word sense disambiguation using translation sense clustering. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., Vossen, P., 2004. The meaning multilingual central repository. In: Proceedings of the Second International WordNet Conference. pp. 80–210.

Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. In: Journal of Machine Learning Research 3, 993–1022.

Carpuat, M., 2013. NRC: A machine translation approach to cross-lingual word sense disambiguation (semeval-2013 task 10). In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 188–192.

DeNero, J., Klein, D., 2007. Tailoring word alignments to syntactic machine translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics, Prague, Czech Republic, pp. 17–24.

Duque-Fernandez, A., Araujo, L., Martinez-Romo, J., 2013. Una nueva técnica de construcción de grafos semánticos para la desambiguación bilingüe del sentido de las palabras. In: Procesamiento del Lenguaje Natural 51 (0).

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. Bradford Books.

Gonzalez-Agirre, A., Laparra, E., Rigau, G., 2012. Multilingual central repository version 3.0. In: Chair, N. C. C., Choukri, K., Declerck, T., Doan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey.

Guo, W., Diab, M., 2010. COLEUR and COLSLM: A WSD approach to multilingual lexical substitution, tasks 2 and 3 semeval 2010. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 129–133.

Ide, N., Veronis, J., 1998. Word sense disambiguation: The state of the art. In: Computational Linguistics 24, 1–40.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. In: MT summit. Vol. 5.

Lefever, E., Hoste, V., 2010a. Construction of a benchmark data set for cross-lingual word sense disambiguation. In: Chair, N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta.

Lefever, E., Hoste, V., 2010b. Semeval-2010 task 3: Cross-lingual word sense disambiguation. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 15–20.

Lefever, E., Hoste, V., 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 158–166.

Lefever, E., Hoste, V., De Cock, M., 2011. Parasense or how to use parallel corpora for word sense disambiguation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2. HLT '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 317–322.

López-Ostenero, F., 2002. Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el usuario. Ph.D. thesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia.

Mahapatra, L., Mohan, M., Khapra, M. M., Bhattacharyya, P., 2010. OWNS: Cross-lingual word sense disambiguation using weighted overlap counts and wordnet based similarity measures. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 138–141.

Màrquez, L., Exsudero, G., Martínez, D., Rigau, G., 2006. Supervised corpus-based methods for WSD. In: Word Sense Disambiguation: Algorithms and Applications. Vol. 33 of Text, Speech and Language Technology. Springer, Dordrecht, The Netherlands, pp. 167–216.

Martinez-Romo, J., Araujo, L., Borge-Holthoefer, J., Arenas, A., Capitán, J. A., Cuesta, J. A., 2011. Disentangling categorical relationships through a graph of co-occurrences. In: Physical Review E 84, 046108.

Mausam, Soderland, S., Etzioni, O., Weld, D. S., Skinner, M., Bilmes, J., 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1. ACL '09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 262–270.

Mihalcea, R., 2006. Knowledge-based methods for WSD. In: Word Sense Disambiguation: Algorithms and Applications. Vol. 33 of Text, Speech and Language Technology. Springer, Dordrecht, The Netherlands, pp. 107–132.

Miller, G. A., Leacock, C., Tengi, E., Bunker, R. T., 1993. A semantic concordance. In: Proceedings ARPA Human Language Technology Workshop. pp. 303–308.

Navigli, R., Ponzetto, S. P., 2010. Babelnet: Building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 216–225.

Navigli, R., Ponzetto, S. P., 2012. Joining forces pays off: Multilingual joint word sense disambiguation. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. EMNLP-CoNLL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1399–1410.

Och, F. J., Ney, H., Mar. 2003. A systematic comparison of various statistical alignment models. In: Computational Linguistics 29 (1), 19–51.

Pons, P., Latapy, M., 2005. Computing communities in large networks using random walks. In: Lecture Notes in Computer Science 3733, 284.

Rudnick, A., Liu, C., Gasser, M., 2013. HLTDI: CL-WSD using markov random fields for semeval-2013 task 10. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 171–177.

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing. Vol. 12. Manchester, UK, pp. 44–49.

Silberer, C., Ponzetto, S. P., 2010. UHD: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 134–137.

Tan, L., Bond, F., 2013. XLING: Matching query sentences to a parallel corpus using topic models for WSD. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 167–170.

van Gompel, M., 2010. UVT-WSD1: A cross-lingual word sense disambiguation system. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 238–241.

van Gompel, M., van den Bosch, A., 2013. WSD2: Parameter optimisation for memory-based cross-lingual word-sense disambiguation. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 183–187.

Vilariño, D., Balderas, C., Pinto, D., Rodríguez, M., León, S., 2010. FCC: Modeling probabilities with GIZA++ for task #2 and #3 of semeval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation. SemEval '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 112–116.

Vossen, P. (Ed.), 1998. EuroWordNet: a multilingual database with lexical semantic networks. Kluwer Academic Publishers, Norwell, MA, USA.