Unsupervised Cross-Lingual Lexical Substitution

Marianna Apidianaki
Alpage, INRIA & Univ Paris Diderot
Sorbonne Paris Cité, UMRI-001
75013 Paris, France
Marianna.Apidianaki@inria.fr

Abstract

Cross-Lingual Lexical Substitution (CLLS) is the task that aims at providing for a target word in context, several alternative substitute words in another language. The proposed sets of translations may come from external resources or be extracted from textual data. In this paper, we apply for the first time an unsupervised cross-lingual WSD method to this task. The method exploits the results of a cross-lingual word sense induction method that identifies the senses of words by clustering their translations according to their semantic similarity. We evaluate the impact of using clustering information for CLLS by applying the WSD method to the SemEval-2010 CLLS data set. Our system performs better on the 'out-of-ten' measure than the systems that participated in the SemEval task, and is ranked medium on the other measures. We analyze the results of this evaluation and discuss avenues for a better overall integration of unsupervised sense clustering in this setting.

1 Introduction

Lexical Substitution (LS) aims at providing alternative substitute words (or phrases) for a target word in context, a process useful for monolingual tasks such as paraphrasing and textual entailment (McCarthy and Navigli, 2009). Its multilingual counterpart, Cross-Lingual Lexical Substitution (CLLS), aims at finding for a target word in context, alternative substitute words in another language. CLLS systems may assist human translators and language learners, while their output may constitute the input to cross-language Information Retrieval and Machine Translation (MT) systems (Sinha et al., 2009; Mihalcea et al., 2010).

The multilingual context in which CLLS is performed permits to override some issues common to monolingual semantic processing tasks, such as the selection of an adequate sense inventory and the definition of the granularity of the semantic descriptions. In a multilingual context, word senses can be easily identified using their translations in other languages (Resnik and Yarowsky, 2000). Although this conception of senses presents some theoretical and practical drawbacks, it provides a standard criterion for sense delimitation which explains its wide adoption in recent works on multilingual Word Sense Disambiguation (WSD) and WSD in MT (Carpuat and Wu, 2007; Ng and Chan, 2007).

In this paper, we explain how semantic clustering may provide answers to some of the issues posed by the traditional cross-lingual sense induction approach, and how it can be efficiently exploited for CLLS. Given that existing CLLS systems rely on predefined semantic resources, we show, for the first time, that CLLS can be performed in a fully unsupervised manner. The paper is organized as follows: in the next section, we present some arguments towards unsupervised clustering for cross-lingual sense induction. The clustering method used is presented in section 3. Section 4 describes the SemEval-2010 CLLS task, and section 5 presents the cross-lingual WSD method used for CLLS. In section 6, we proceed to a detailed analysis of the obtained results, before concluding with some avenues for future work.

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 13–23,
Edinburgh, Scotland, UK, July 27–31, 2011.

2 Cross-lingual sense induction

2.1 Related work

Word sense induction (WSI) methods offer an alternative to the use of predefined semantic resources for NLP. They automatically define the senses of words from textual data and may adapt the obtained descriptions to the WSD needs of specific applications. In a monolingual context, WSI is performed by exploiting more or less refined distributional information (Navigli, 2009), while in a multilingual context WSI is mostly based on translation information. In this setting, the senses of words in one language are identified by their translations in another language, usually found in a parallel corpus (Resnik and Yarowsky, 2000).

This empirical approach to sense induction offers a standard criterion for sense delimitation and, consequently, dissociates WSD from semantic theories and predefined semantic inventories. Moreover, by establishing semantic distinctions pertinent for translation between the implicated languages, it allows to tune sense induction to the needs of multilingual applications. It has thus been widely adopted in works on multilingual WSD and WSD in MT, where senses are derived from parallel data (Diab, 2003; Ide, 1999; Ide et al., 2002; Ng et al., 2003; Chan et al., 2007; Carpuat and Wu, 2007). By linking WSD and its evaluation to translation, this hypothesis also offers a solution to the problem of non-conformity of monolingual WSD methods in this setting.

Nevertheless, the assumption of biunivocal ('one-to-one') correspondences between senses and translations is rather simplistic. One word sense may be translated by different synonymous words in another language, whose relatedness should be considered during sense induction. Furthermore, this approach does not permit to account for cases of parallel ambiguities (Resnik, 2007), and cases where the senses of a word share some of their translations (Sinha et al., 2009). Additional problems arise at the practical level as the induced senses are uniform and, so, the constraints used during WSD for selecting between close and distant senses are similar. Furthermore, when WSD coincides with lexical selection in MT, the selection of a translation different from the reference is considered as wrong even if it is semantically correct. So, this conception of senses does not permit to penalize WSD errors relatively to their importance (Resnik and Yarowsky, 2000), unless semantic resources are used to identify semantic correspondences.

2.2 Cross-lingual sense clustering

Instead of using translations as straightforward sense indicators, it is possible to perform a more thorough semantic analysis during cross-lingual WSI by combining distributional and translation information. The sense clustering method proposed by Apidianaki (2008) identifies complex semantic relations between word senses and their translations. The method is based on the contextual hypotheses of meaning and of semantic similarity (Harris, 1954; Miller and Charles, 1991), which underlie monolingual WSI methods, and is combined to the assumption of a semantic correspondence between words and their translations in real texts (Chesterman, 1998). Following these hypotheses, information coming from the source contexts of a target word when translated with a precise translation in a parallel corpus, is used to reveal the senses carried by the translation. Furthermore, the similarity of the source contexts reveals the semantic relatedness of the translations.

This cross-lingual WSI method groups the semantically similar translations of ambiguous words into clusters that serve to describe their senses instead of the individual translations. For instance, the traditional cross-lingual WSI approach would propose three senses for the English noun *coach*, corresponding to each of its Spanish translations: *entrenador, autocar* and *autobús*.[1] However, this solution is not sound given that the translations *autocar* and *autobús* are semantically related and do not lexicalize distinct senses of the English word, as is the case with *entrenador*. Sense clustering permits to estimate the semantic similarity of the translations and to not consider synonymous translations as indicators of distinct senses. Consequently, the English word *coach* has two senses after sense clustering: one described by the cluster {*autocar, autobús*} (the "bus" sense) and one described by the cluster {*entrenador*} (the "trainer" sense). In the automat-

[1] This set of translations was extracted from the word aligned Europarl corpus (Koehn, 2005) after applying a set of filters that will be described in section 3.

ically built bilingual inventories, the senses of the words in one language are thus described by clusters of their translations in another language.

2.3 Applications

This type of sense clustering has proved to be useful in various application settings. When exploited in cross-lingual WSD, it permits to assign 'sense-tags' containing several semantically correct translations to new instances of words in context (Apidianaki, 2009). Moreover, the use of clustering information during evaluation allows for a differing penalization of WSD errors. In an MT evaluation setting, sense clusters have been integrated into an MT evaluation metric (METEOR) (Lavie and Agarwal, 2007) and brought about an increase of the metric's correlation with human judgments of translation quality in different languages (Apidianaki and He, 2010). The use of sense clusters in this setting permits to identify semantic correspondences between translations and hypotheses, and to circumvent the strict requirement for exact surface correspondences, one of the main critics addressed to MT evaluation metrics. The same notion of sense clusters has been adopted in the most recent SemEval Cross-Lingual WSD task (Lefever and Hoste, 2010). Instead of considering translations as indicators of distinct senses, as was the case in previous tasks, the senses of a small number of ambiguous words were described by manually created clusters of translations.

We consider that the sense cluster inventories created by the unsupervised WSI method proposed by Apidianaki (2008) would be useful in other applicative contexts as well and, especially, in CLLS. In unsupervised cross-lingual WSD, the clusters constitute the candidate senses from which one has to be selected for each new instance of the words in context. So, when an instance of a word is disambiguated, a cluster of semantically related translations is selected on the basis of the source context describing its sense. This is exactly the goal of CLLS, as described in the relevant task set up in SemEval-2010, where the systems had to provide for instances of words in context, several possible translations in another language (Sinha et al., 2009; Mihalcea et al., 2010). It seems thus that CLLS constitutes a suitable field for exploiting this sense clustering method and, in what follows, we will try to evaluate this assumption.

3 Unsupervised clustering for sense induction

3.1 Bilingual lexicons

The SemEval-2010 CLLS task concerned the pair of languages English (EN) - Spanish (SP). In order to apply our cross-lingual WSD method to the data of the SemEval-2010 CLLS task, an EN-SP sense cluster inventory had first to be built where the senses of English words would be described by clusters of their Spanish translations. The training corpus used for building the sense cluster inventory is the SP-EN part of Europarl (release v5), which contains 1,689,850 aligned sentence pairs (Koehn, 2005). Before clustering, some preprocessing steps are performed. First, the corpus is lemmatized and tagged by POS (Schmid, 1994). Then sentence pairs presenting a great difference in length (i.e cases where one sentence is three times longer than the other) are eliminated and the corpus is aligned at the level of word types using Giza++ (Och and Ney, 2003).

Two bilingual lexicons of content words are built from the alignment results, one for each translation direction (EN-SP/SP-EN). In the entries of these lexicons, source words are associated with the translations to which they are aligned. As these lexicons are automatically created, they contain some noise mainly due to spurious word alignments. In order to eliminate erroneous translation correspondences, we first apply a filter which discards translations with a probability below 0.001 (according to the scores assigned during word alignment). Then an intersection filter is applied which discards correspondences not found in lexicons of both directions. Finally, the two lexicons are filtered by POS, keeping for each *w* only its translations that pertain to the same POS category.[2] The translations of a word (*w*) used for clustering are the ones that translate *w* at least 20 times in the training corpus. This frequency threshold leaves out some translations of the source words but has a double merit: it eliminates erroneous translations

[2]For instance, for English nouns we retain their noun translations in Spanish; for verbs, we keep verbs, etc.

and reduces data sparseness issues which pose problems in distributional semantic analysis.

3.2 Clustering based on semantic similarity

The semantic clustering is performed in the target language by using source language feature vectors. Each translation of a word *w* is characterized by a vector built from the content words that cooccur with *w* whenever it is translated by this word in the aligned sentences of the training corpus.[3] The vector similarity is calculated using a variation of the Weighted Jaccard measure (Grefenstette, 1994) which weighs each source context feature according to its relevance for the estimation of the translations similarity.

The input of the similarity calculation consists of the frequency lists of *w*'s translations. The score assigned to a pair of translations indicates their degree of similarity. Each feature (*j*) gets a *total weight* (*tw*) relatively to a translation (*i*), which corresponds to the product of its *global* (*gw*) and its *local weight* (*lw*) with this translation. The *gw* is based on the dispersion of *j* in the contexts of *w*, and on its frequency of cooccurrence (*cooc_freq*) with *w* when translated by each *i* (cf. formula 1). So, it depends on the number of translations with which *j* is related (*nrels*) and on its probability of cooccurrence with each one of them (cf. formula 2). The *local weight* (*lw*) between *j* and *i* depends on their frequency of cooccurrence (cf. formula 3).

$$\mathrm{gw}(j) = 1 - \frac{\sum_i \mathrm{p}_{ij} \log(\mathrm{p}_{ij})}{\mathrm{nrels}} \quad (1)$$

$$\mathrm{p}_{ij} = \frac{\mathrm{cooc\_freq\ of}\ j\ \mathrm{with}\ i}{|j\mathrm{s}|\ \mathrm{for}\ i} \quad (2)$$

$$\mathrm{lw}(j,i) = \log(\mathrm{cooc\_freq\ of}\ j\ \mathrm{with}\ i) \quad (3)$$

The Weighted Jaccard (WJ) coefficient of two translations *m* and *n* is given by formula 4.

$$\mathrm{WJ}(m,n) = \frac{\sum_j \min(\mathrm{tw}(m,j)\mathrm{tw}(n,j))}{\sum_j \max(\mathrm{tw}(m,j)\mathrm{tw}(n,j))} \quad (4)$$

The pairwise similarity of the translations is thus estimated by comparing the corresponding weighted source feature vectors. A similarity score is assigned to each pair of translations and stored in a table that is being looked up by the clustering algorithm. The pertinence of the relation of each translation pair is estimated by comparing its score to a threshold defined locally for each *w* by the following iterative procedure.

1. The initial threshold (T) corresponds to the mean of the scores (above 0) of the translation pairs of *w*.
2. The set of translations is segmented into pairs whose score exceeds the threshold and pairs whose score is inferior to the threshold, creating two sets (G1, G2).
3. The average of each set is computed (*m1* = average value of G1, *m2* = average value of G2).
4. A new threshold is created that is the average of *m1* and *m2* (*T* = (m1 + m2)/2).
5. Go back to step 2, now using the new threshold computed in step 4, keep repeating until convergence has been reached.

The clustering algorithm groups the translations by exploiting the similarity calculation results. The condition for a translation to be included in a cluster is to have pertinent relations with all the elements already in the cluster. The clustering stops when all the translations of *w* are included in some cluster and all their relations have been checked. All the elements of the final clusters are linked to each other by pertinent relations. The translations not having any strong relations to other translations are included in separate one-element clusters.

3.3 The EN-SP sense cluster inventory

In the obtained semantic inventory, the senses of each English word are described by clusters of its semantically similar translations in Spanish.[4] Some entries from the EN-SP sense cluster inventory are presented in Table 1. We provide examples for words of different POS (nouns, verbs, adjectives and adverbs) and with varying degrees of polysemy. The

[3] We use a stoplist of English function words (conjunctions, prepositions and articles) that may be erroneously tagged as content words.

[4] The inventory contains entries for all English content words in the corpus. Here, we focus on the target words used in the CLLS task.

POS	EN word	# SP_Ts	# occ	Sense clusters
Nouns	coach	3	265	{entrenador}{autocar, autobús}
	test	11	3162	{prueba, ensayo, examen} {experimento, análisis, examen, ensayo} {evaluación} {comprobación} {experimentación, ensayo, análisis, experimento} {inspección} {experimento, control, análisis, examen} {experimentación, control, análisis, experimento} {criterio}
Verbs	drop	10	390	{disminuir, reducir, bajar, caer, descender} {retirar} {dejar, abandonar} {lanzar}
	check	5	1343	{examinar} {revisar} {controlar, verificar, comprobar}
Adjs	heavy	7	448	{elevado, fuerte, grave, grande}{elevado, enorme}{grave, duro, fuerte, grande} {grave, alto, elevado}
	open	6	6286	{público, libre, transparente} {público, franco, transparente} {abierto} {sincero, franco}
Advs	around	5	742	{alrededores}{casi, aproximadamente, cerca}{menos}
	now	9	33662	{aquí, actualmente, hoy, ahora bien} {actualmente, ahora, hoy} {entretanto, aquí, ahora bien} {de momento}, {adelante}, {por ahora, entretanto}

Table 1: *Entries from the EN-SP sense cluster inventory.*

third column of the table gives the number of Spanish words (SP_Ts) translating more than 20 occurrences of the English words in the corpus and retained for clustering. This threshold ensures that the words being clustered are good translations of the English words. The fourth column of the table shows the number of English word occurrences translated by the retained translations.

As is shown in these examples, the translations of the English words are not considered as straightforward indicators of their senses but are grouped into clusters describing senses. For instance, the word *drop*, which is translated by ten different words into Spanish (*disminuir, reducir, bajar, caer, descender, retirar, dejar, abandonar, lanzar*) is not considered as having ten distinct senses but four, described by each cluster of translations: {disminuir, reducir, bajar, caer, descender}: "decrease, reduce", {retirar}: "remove, withdraw", {dejar, abandonar}: "leave, abandon" and {lanzar}: "launch". The obtained clusters group semantically similar words which would be erroneously considered as indicators of distinct senses by the traditional cross-lingual sense induction method.

Another important point is that this algorithm performs a soft clustering, highly adequate in this setting. Given that the generated clusters describe senses, their overlaps describe the relations between the corresponding senses. For instance, the two senses of the word *test* described by the clusters {experimentación, control, análisis, experimento} and {experimento, control, análisis, examen} share three elements and are closer than those described by {experimentación, control, análisis, experimento} and {evaluación}, which have no element in common. The first two senses could also be considered as nuances of a coarser sense ("examination / analysis") that could be obtained by merging the overlapping clusters. Capturing inter-sense relations is important in lexical semantics and numerous works have been criticized for just enumerating word senses without describing their relations. Discovering these links automatically, as is done with this sense clustering method, permits to account for differences in the status of senses during WSD and its evaluation. It also offers the possibility to automatically modify the granularity of the obtained senses according to the WSD needs of the applications. Moreover, when the sense cluster inventory is used for cross-lingual WSD, it allows to capture subtle relations between word usages in cases where the senses of a word share some of their translations but not all of them, an issue highlighted in the SemEval CLLS task (Sinha et al., 2009) which will be presented in the next section.

4 The SemEval-2010 CLLS task

In the SemEval-2010 Cross-Lingual Lexical Substitution task, annotators and systems had to provide several alternative correct translations in Spanish for English target words in context. Given a paragraph containing an instance of an English target word, the annotators had to find as many good substitute translations as possible for that word in Spanish. Unlike a full-blown MT task, CLLS targets one word at a time rather than an entire sentence. So, annotators were asked to translate the target word and not entire sentences. Moreover, they were asked to supply, for each instance, as many translations as they felt were valid and not just one translation, which would be the case in MT.

The task of the participating systems was then to predict the translations provided by the annotators for each target word instance. By analyzing the context of the English target word instances, the systems had to provide for each instance, several correct Spanish translations which should fit the given source language context. The set of target words in the SemEval CLLS task is composed of Nouns, Verbs, Adjectives and Adverbs exhibiting a wide variety of substitutes. The annotators were allowed to use any resources they wanted to in order to supply substitutes for instances of the English target words. So, instances of the target words in context were tagged by sets of Spanish translations.[5] The inter-annotator agreement for this task was calculated as pairwise agreement between sets of substitutes from annotators and corresponds to 0.2777.

The sets of translations provided for different instances of a target word could overlap in different degrees, depending on the meaning of the instances. These overlaps reveal subtle relations between word usages in cases where they share some of their translations but not all of them (Sinha et al., 2009). This also shows the absence of clear divisions between usages and senses: usages overlap to different extents without having identical translations. Although no clustering of translations from a specific resource into senses was performed for this task, the interest of examining the possibility of clustering the translations provided by the annotators is highlighted (Mihalcea et al., 2010).

[5]The average numbers of substitutes provided by the annotators for words of different POS are: 4.47 for nouns, 5.2 for verbs, 4.99 for adjectives and 4.77 for adverbs.

5 Cross-lingual WSD

The source language features that revealed the similarity of the translations and served to their clustering (cf. section 3) can be exploited by an unsupervised WSD classifier (Apidianaki, 2009). In order to disambiguate a new instance of an English word *w*, cooccurrence information coming from its context is compared to these feature sets and the cluster that has the highest similarity with the new context is selected. We adopt this WSD method in order to exploit the sense clustering results and perform CLLS in an unsupervised manner. Instead of comparing the new contexts to the features that are common to all the translations in a cluster (i.e. the intersection of their source language features), as is done in the initial method, we compare them to the features shared by each pair of translations. This increases the coverage of the method, given that these source features sets are larger than the ones containing the intersection of the features of all the clustered translations. As the training corpus was lemmatized and POS-tagged prior to building the feature vectors (only content word cooccurrences were retained), the new contexts have to be lemmatized and POS-tagged as well.

If common features (CFs) are found between the new context and a translation pair, a score is assigned to this 'context-pair' association which corresponds to the mean of the weights of the CFs relatively to each translation of the pair. The weights used here are the total weights (*tw*s) that were assigned to the context features relatively to the translations during the semantic similarity calculation (cf. section 3.2). In formula 5, i is equal to 2 (i.e. the number of translations in the pair) and j is the number of CFs between the translation pair and the new context.

If the highest-ranked translation pair is found in just one sense cluster, this cluster is selected as describing the sense of the new instance. Otherwise, if the translation pair is found in different clusters, it is checked whether the CFs characterize the other translations in these clusters (or some of them). If this is the case, a score is assigned to each cluster

Test instance	WSD suggestion	Gold annotation
test.n 1698	prueba;ensayo;examen;	examen 4;prueba 4;test 1;
board.n 1781	consejo;bordo;junta;comité;cuenta; administración;	junta directiva 2;consejo 2;mesa directiva 1;junta 1;junta de ayuda 1;directiva 1;comite 1;comision 1;
drop.v 1288	bajar;disminuir;reducir;caer;descender	dejar caer 2;tirar 1;arrojar 1;lanzar 1;soltar 1;dejar 1;bajar 1;
check.v 851	comprobar;controlar;verificar;	verificar 3;checar 2;confirmar 1;anotar 1;rectificar 1;revisar 1;comprobar 1;
yet.r 1766	todavía;aún;sin embargo;	sin embargo 2;pero 2;no obstante 1;aun 1;todavia 1;
now.r 1019	hoy;aquí;actualmente;ahora bien;	hoy 2;ahora 2;este momento 2;a partir 1;el presente 1;de aqui 1;

Table 2: *Clusters suggested by the WSD method.*

depending on the weights of the features with the other translations, and the cluster with the highest score is selected as describing the sense of the new instance. The score is again calculated by formula 5 but this time *i* is equal to the number of translations in the cluster having CFs with the new context.

$$\text{score} = \frac{\sum_i \sum_j tw(i,j)}{i * j} \quad (5)$$

If no CFs are found using the translation pairs, the WSD algorithm considers each translation's feature set separately (which is naturally larger than the feature sets of the translation pairs). If CFs exist, the translation with the highest score is selected as well as the cluster containing it. If the translation is found in the intersection of different clusters, it is checked whether the CFs characterize some of the other translations found in the clusters. If this is the case, a score is assigned to the clusters depending on the weights of the features with the translations and the cluster with the highest score is selected. The cluster containing the translation pair with the highest similarity to the new context is retained as the sense of the new instance. If no CFs are found in this way neither, a most frequent sense heuristic is used which selects the most frequent cluster (i.e. the one assigned to most of the new instances of *w*).

For the 1000 test instances in the SemEval CLLS task, the WSD method proposes 625 clusters with more than one element and 118 one element clusters.[6] The most frequent translation is suggested in 210 cases while the most frequent cluster is chosen in 43 cases. A cluster is chosen randomly only in 3 cases. In Table 2, we present some suggestions made by the WSD method for target words of different POS (n: nouns, v: verbs, a: adjectives, r: adverbs) and the corresponding gold standard (GS) annotations. For instance, the following occurrence of the English noun *test*:

> Entries typically identify the age or school grade levels for which the **test** is appropriate, as well as any subtests.

is tagged by the Spanish cluster {*prueba, examen, ensayo*} during WSD, which is close to the GS annotation {*examen, prueba, test*} and correctly describes its sense.

The first translation provided in the results is the word of the cluster that translates most of the English target word instances in the corpus (and which is duplicated in order to be reinforced during the 'out-of-ten' evaluation, as we will explain in the next section). We observe that this most frequent word, although it is a correct translation (i.e. found in the GS annotations), does not coincide with the annotators' first choice. This explains the evaluation results that we present in the next section.

It is also important to note that the system suggests not only translations that have been proposed by the annotators, but also other semantically pertinent translations that were found in the training corpus and which do not exist in the GS annotations. This is the case, for instance, with the translation

[6] 262 clusters with two elements; 157 clusters with three; 73 with four; 64 with five; 69 clusters with more than five and less than ten elements; 23 clusters with ten elements and 22 clusters with more than ten elements.

”controlar” of the verb *check* and the translation ”ensayo” proposed for the noun *test*. This shows that the suggestions made by the WSD method greatly depend on the corpus used for training.

6 Evaluation

6.1 The setting

We evaluate our method on the SemEval-2010 CLLS task test set. The metrics used for evaluation are the *best* and *out-of-ten* (oot) precision (P) and recall (R) scores. In the SemEval task, the systems were allowed to supply as many translations as they felt fit the context. These suggestions were then given credit depending on the number of annotators that had picked each translation. The credit was divided by the number of annotator responses for the item. For the *best* score, the credit for the system answers for an item was also divided by the number of answers provided by the system, which allows more credit to be given to instances with less variation.

The *oot* scorer allows up to ten system responses and does not divide the credit attributed to each answer by the number of system responses. This scorer allows duplicates which means that systems can get inflated scores (i.e. > 100), as the credit for each item is not divided by the number of substitutes and the frequency of each annotator response is used. Allowing duplicates permits that the systems boost their scores with duplicates on translations with higher probability.[7]

Two baselines are used for evaluation: a dictionary-based one (DICT), which contains the Spanish translations of all target words provided by an SP-EN dictionary, and a dictionary and corpus-based one (DICTCORP), where the translations provided by the dictionary for a given target word are ranked according to their frequencies in the Spanish Wikipedia. In DICT, the *best* baseline is produced by taking the first translation provided by the dictionary while the *oot* baseline considers the first ten translations.

6.2 Results

In order to evaluate our WSD method, we proceed as follows. If the cluster selected by the WSD method contains ten translations (or more), all the translations are given in the *oot* results. Otherwise, the translations found in the cluster are proposed and the most frequent translation is duplicated till reaching ten elements. For *best*, we always retain the most frequent translation of the selected cluster.

Our intuition was that the WSD method, which assigns sense clusters (i.e. sets of semantically similar and, more or less, substitutable translations), would fit and perform well on the *oot* subtask of the SemEval CLLS task. This is confirmed by the results presented in Table 3.[8] Our method (denoted by 'WSD' in the table) outperforms the 14 systems that participated in the CLLS task as well as the recall (R) and precision (P) baselines. It is important to note that, contrary to our method which is totally unsupervised, all the systems that participated in the SemEval-2010 task used predefined resources. The second ranked system (SWAT-E), for instance, performs lexical substitution in English and then translates each substitute into Spanish using two predefined bilingual dictionaries, while SWAT-S does the inverse, performing lexical substitution in the translated text (Wicentowski et al., 2010).

Systems	R	P	Mode R	Mode P
WSD	**180.10**	**186.25**	**56.52**	**58.44**
SWAT-E	174.59	174.59	66.94	66.94
SWAT-S	97.98	97.98	79.01	79.01
UvT-v	58.91	58.91	62.96	62.96
UvT-g	55.29	55.29	73.94	73.94
DICT	44.04	44.04	73.53	73.53
DICTCORP	42.65	42.65	71.60	71.60

Table 3: **oot** results (%)

Another interesting point is that the sense cluster inventory used by the cross-lingual WSD method is derived from Europarl, which is the European Parliament Proceedings parallel corpus (Koehn, 2005). Despite this fact, the WSD method that exploits this inventory performs particularly well on this task which concerns the semantic analysis and translation of words of general language. We would thus expect the results to be even better if the sense induc-

[7] The metrics used for evaluation are defined in Mihalcea et al. (2010).

[8] We report the results obtained by the highest-ranked systems in the SemEval-2010 CLLS task. The full table of results can be found in Mihalcea et al. (2010).

tion and the WSD method were trained on a bigger, or more general, parallel corpus.

The mode recall and precision (Mode R and Mode P) metrics evaluate the performance of the systems in predicting the translation that was most frequently selected by the annotators, provided that such a translation extists. To identify the most frequent response, we order the system responses according to their frequency as translations of the target words in the training corpus. The relatively low scores obtained for the Mode R and Mode P metrics (compared to R and P) are explained by the fact that the most frequent translation in the training corpus does not always correspond to the translation that was most frequently selected by the annotators, although it may be a good translation for the target word.

The same reason explains the weaker performance of the method in the *best* evaluation subtask (cf. Table 4), where our system is ranked eighth compared to the 14 systems that participated in the task.[9] Here too, the *best* translation according to the annotators does not correspond to the most frequent translation in the corpus. This highlights the impact that the relevance of the training corpus to the domains of the processed texts has on unsupervised CLLS.

Systems	R	P	Mode R	Mode P
UBA-T	27.15	27.15	57.20	57.20
USPWLV	26.81	26.81	58.85	58.85
WLVUSP	25.27	25.27	52.81	52.81
WSD	**19.73**	**19.93**	**41.29**	**41.75**
UBA-W	19.68	19.68	39.09	39.09
SWAT-S	18.87	18.87	36.63	36.63
IRST-1	15.38	22.16	33.47	45.95
TYO	8.39	8.62	14.95	15.31
DICT	24.34	24.34	50.34	50.34
DICTCORP	15.09	15.09	29.22	29.22

Table 4: **best** results (%)

Another important factor that has to be taken into account is that the WSD method that we use is oriented towards multilingual applications (more precisely MT). In these applications, it is possible to filter the proposed sense clusters by reference to the target language context (for instance, by using a language model) in order to retain the most adequate translation. It is interesting to note that the systems that perform better in the *best* subtask get relatively low results in the *oot* subtask, and the inverse. This is the case, for instance, for UBA-T (Basile and Semeraro, 2010), while Aziz and Specia (2010) clearly specify that their main goal is to maximize the accuracy of their system (USPwlv) in choosing the *best* translation. A conclusion that can be drawn is that each subtask has different requirements, which may be satisfied by different types of methods.

In order to investigate other possible reasons behind the different behavior of the WSD method in the two evaluation subtasks, we performed the evaluation separately for each POS. The results are presented in Tables 5 and 6.

POS	R	P	Mode R	Mode P
Adjs	**287.94**	**296.41**	**72.44**	**74.43**
Nouns	127.01	141.65	37.78	42.29
Verbs	115.94	121.43	53.17	55.90
Advs	111.46	111.46	65.15	65.15

Table 5: **oot** results for different POS (%)

POS	R	P	Mode R	Mode P
Adjs	**30.77**	**31.00**	**63.56**	**64.13**
Nouns	14.61	16.29	25.78	28.86
Verbs	14.98	14.98	29.76	29.76
Advs	13.07	13.07	37.88	37.88

Table 6: **best** results for different POS (%)

In both the *oot* and *best* evaluation subtasks, the best scores are obtained for adjectives. Especially in the *best* subtask, where the method seemed to perform worse than the other systems, the recall and precision scores obtained for adjectives (with and without mode) are higher than those obtained by the highest-ranked system (cf. Table 4) and much higher than the baselines. A more detailed look at the obtained results proved that the most frequent translation of the English adjectives in our training corpus – proposed in the *best* evaluation subtask and emphasized in the *oot* subtask – is often the most frequent translation proposed by the annotators. This is not the case for the other POS, where the most frequent

[9] We report some indicative results from the *best* subtask. The full table of results can be found in Mihalcea et al. (2010).

translation in the corpus often does not correspond to the annotators' first choice. Furthermore, the translation proposed by the system is not the same as the most frequent translation of the word in the general dictionary and the Spanish Wikipedia which were used, respectively, for the DICT and DICT-CORP baselines. Consequently, this issue could probably be resolved if a more balanced corpus was used for training the WSI and WSD methods.

7 Conclusions and future work

We have shown that Cross-Lingual Lexical Substitution can be performed in a totally unsupervised manner, if a parallel corpus is available. We applied an unsupervised cross-lingual WSD method based on semantic clustering to the SemEval-2010 CLLS task. The method performs well compared to the systems that participated in the task, which exploit predefined lexico-semantic resources. It is ranked first on the *out-of-ten* measure and medium on measures that concern the choice of the *best* translation. We wish to pursue this work and explore other ways for selecting *best* translations than solely relying on frequency information. As unsupervised methods heavily rely on the training data, it would also be interesting to experiment with different corpora in order to evaluate the impact of the type and the size of the corpus on CLLS.

The sense clusters assigned to target word instances during CLLS contain semantically similar translations of these words, more or less substitutable in the target language context. We consider that it would be interesting to integrate target language information in the CLLS decision process for selecting *best* translations. Given that MT is one of the envisaged applications for this type of task, but the use of a full-blown MT system would probably mask system capabilities at a lexical level, a possibility would be to exploit the CLLS system suggestions in a simplified MT task such as *word translation* (Vickrey et al., 2005) or *lexical selection* (Apidianaki, 2009), or in an MT evaluation context. This would permit to estimate the usefulness of the system suggestions in a specific application setting.

References

Marianna Apidianaki and Yifan He. 2010. An algorithm for cross-lingual sense clustering tested in a MT evaluation setting. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT-10)*, pages 219–226, Paris, France.

Marianna Apidianaki. 2008. Translation-oriented sense induction based on parallel corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-08)*, pages 3269–3275, Marrakech, Morocco.

Marianna Apidianaki. 2009. Data-driven Semantic Analysis for Multilingual WSD and Lexical Selection in Translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, pages 77–85, Athens, Greece.

Wilker Aziz and Lucia Specia. 2010. USP*wlv* and WLV*usp*: Combining Dictionaries and Contextual Information for Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 117–122, Uppsala, Sweden.

Pierpaolo Basile and Giovanni Semeraro. 2010. UBA: Using Automatic Translation and Wikipedia for Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 242–247, Uppsala, Sweden.

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of the Joint EMNLP-CoNLL Conference*, pages 61–72, Prague, Czech Republic.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 33–40, Prague, Czech Republic.

Andrew Chesterman. 1998. *Contrastive Functional Analysis*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Mona Diab. 2003. *Word sense disambiguation within a multilingual framework*. Ph.D. dissertation, University of Maryland.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

Zelig Harris. 1954. Distributional structure. *Word*, 10:146–162.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL Workshop on Word Sense Disambiguation:*

Recent Successes and Future Directions, pages 54–60, Philadelphia.

Nancy Ide. 1999. Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34(1-2):223–234.

Philip Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the ACL-2007 Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 Task 3: Cross-lingual Word Sense Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 15–20, Uppsala, Sweden.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond*, 43(2):139–159.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL 2010*, pages 9–14, Uppsala, Sweden.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Roberto Navigli. 2009. Word Sense Disambiguation: a Survey. *ACM Computing Surveys*, 41(2):1–69.

Hwee Tou Ng and Yee Seng Chan. 2007. SemEval-2007 Task 11: English lexical sample task via English-Chinese parallel text. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 54–58, Prague, Czech Republic.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 455–462, Sapporo, Japan.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Philip Resnik and David Yarowsky. 2000. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering*, 5(3):113–133.

Philip Resnik. 2007. WSD in NLP applications. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, pages 299–337, Dordrecht. Springer.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Ravi Sinha, Diana McCarthy, and Rada Mihalcea. 2009. SemEval-2010 Task 2: Cross-Lingual Lexical Substitution. In *Proceedings of the NAACL-HLT Workshop SEW-2009 - Semantic Evaluations: Recent Achievements and Future Directions*, pages 76–81, Boulder, Colorado.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-Sense Disambiguation for Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 771–778, Vancouver, Canada.

Richard Wicentowski, Maria Kelly, and Rachel Lee. 2010. SWAT: Cross-Lingual Lexical Substitution using Local Context Matching, Bilingual Dictionaries and Machine Translation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2), ACL*, pages 123–128, Uppsala, Sweden.