UNAL-NLP: Cross-Lingual Phrase Sense Disambiguation with Syntactic Dependency Trees

Emilio Silva-Schlenker
Departamento de Lingüística
Universidad Nacional de Colombia
Departamento de Ingeniería de Sistemas
Universidad de los Andes,
Bogotá D.C., Colombia
esilvas@unal.edu.co

Sergio Jimenez and Julia Baquero
Universidad Nacional de Colombia,
Bogotá D.C., Colombia
sgjimenezv@unal.edu.co
jmbauqerov@unal.edu.co

Abstract

In this paper we describe our participation in the SemEval 2014, Task 5, consisting of the construction of a *translation assistance system* that translates L1 fragments, written in L2 context, to their correct L2 translation. Our approach consists of a bilingual parallel corpus, a system of syntactic features extraction and a statistical memory-based classification algorithm. Our system ranked 4th and 6th among the 10 participating systems that used the English-Spanish data set.

1 Introduction

An L2 writing assistant is a tool intended for language learners who need to improve their writing skills. This tool lets them write a text in L2, but fall back to their native L1 whenever they are not sure about a certain word or expression. In these cases, the assistant automatically translates this text for them (van Gompel et al., 2014).

Although at first glance this may be seen as a classification problem, it might be better fulfilled by a cross-lingual word sense disambiguation (WSD) approach, which takes context into account by means of contextual features used in a machine learning setting. The main differences between this and previous approaches to cross-lingual WSD are the bilingual nature of the input sentences (see section 2.3) and the annotation of target phrases, rather than single words.

The remainder of this article is organized as follows. Section 2 describes the proposed method. A description of the system we submitted, the obtained results and an error analysis are discussed in section 3. In section 4 we present a brief discussion about the results. Finally, in section 5 we make some concluding remarks.

2 Method Description

The core of the proposed system uses techniques from memory-based classification to find the most appropriate translation of a target phrase in a given context. It receives an input as in (1) and yields an output as in (2).

(1) No creo que ella *is coming*.
(2) No creo que ella *venga*.

It does so on the basis of a syntactic selection of context features, a large bilingual parallel corpus and a classifier built using the Tilburg Memory-Based Learner, TiMBL (Daelemans et al., 2010).

The proposed system consists of several stages. First, a large bilingual corpus is aligned at word and phrase level. Next, an index is built by each phrase in the L1 side of the corpus to retrieve efficiently the occurrences of a particular L1 phrase in the aligned corpus along with their translations and contexts in L2 (subsection 2.1). Second, the relevant contexts for each L1 phrase in the test set (example sentences) are retrieved from the corpus and a set of syntactic features are extracted from each sentence (subsection 2.2). Third, a special two-stage process is used to extract the same features from the sentences in the test set to deal with the fact that these sentences were written in two languages (subsection 2.3). Finally, each target phrase is translated using the IBL algorithm and the translations were incorporated in the original test sentences (subsection 2.4).

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 743–747,
Dublin, Ireland, August 23-24 2014.

Input sentence	Parallel example sentences	
No creo que las necesidades afectivas de las personas estén necesariamente *linked* al matrimonio.	He said Boyd already *linked* him to Brendan.	Dijo que Boyd ya le había *relacionado* con Brendan.
	The three things are inextricably *linked*, and I have the formula right here.	Las tres cosas están estrechamente *vinculadas*, y tengo la fórmula aquí.

Table 1: An input sentence and 2 example sentences from Linguee.com.

2.1 Parallel Corpus Selection and Preparation

As no training corpus was given prior to developing this system, finding and processing the most suitable corpus for this task was paramount. As the purpose of this system is to help language students, the corpus needs to account for simple yet correct everyday speech.

In an initial stage of development we opted to use the 70-million sentences OpenSubtitles.org corpus compiled by the Opus Project (Tiedemann, 2012), which includes many informal everyday utterances, at the expense of a less accurate translation quality[1]. Although the use of this training corpus yielded over 95% of recall on the trial corpus given by the task organizers, only 80% of the trial sentences had enough (>100) training examples in order to produce a quality translation. To solve this issue, an ad-hoc corpus compilation mechanism was created by using the Linguee.com. Thus, a set of parallel example sentences is retrieved from Linguee.com by querying all the L1 target phrases from the evaluation data (see an example in Table 1).

The corpus preparation procedure consisted of several steps. The first step was to clean the corpus with the Moses cleaning script (Koehn et al., 2007). Next, the corpus was tokenized and PoS-tagged using FreeLing (Padró and Stanilovsky, 2012) (HMM tagger was used). After that, the corpus was word-aligned using Giza++ (Och and Ney, 2003) over Moses (Koehn et al., 2007). The resulting alignment was then combined with the tagged version of the corpus. Finally, a phrase index was built using a SMT phrase extraction algorithm (Ling et al., 2010) including for each phrase pointers to all its occurrences in the corpus for further retrieval.

[1]The EPPS corpus (Lambert et al., 2005) was very useful as a training corpus in the developing stages of this system. It was however not used in the final system training.

2.2 Syntactic Feature Extraction

The syntactic tags feature is a novel feature we are introducing for the CLWSD problem (Lefever and Hoste, 2013). They are linearizations of syntactic dependency trees. These trees were built by Freeling's Txala Parser (Lloberes et al., 2010) and were introduced as individual tags in a sentence analysis by parsing the tree and mapping its leaves with their corresponding order in the source sentence. Then, each leaf's label and parent number was extracted. For the root, the special parent tag 'S' was used.

The WSD literature commonly distinguishes between local and global context features (Martinez and Agirre, 2001). The former are extracted from the neighboring words and the latter are extracted from words of the whole context provided using some heuristic to select relevant. Unlike global features, the relevance of the surrounding words is not put into question or are weighted by the degree of relevance according to their position in the sentence and lexicographic distance from the target phrase (van Gompel, 2010). There is a linguistic explanation as to why surrounding words play a significant role in determining the target's translation. Often, these words have a direct dependency relation with the target. Indeed, physical closeness is an approximation of syntactic relatedness. What we propose in this paper is that the relevance of the context words for determining a correct translation is proportional to their syntactic relatedness to the target, rather than their physical closeness in the sentence. Unlike Martinez et al. (2002), what we propose here is to use syntax as a feature selector, rather than as a feature itself.

Instead of defining a local and a global set of relevant words, we selected a single set of relevant words according to their syntactic relation to the target phrase. This set consisted of all the children of the target words, and the parents of the main target words. The main target words are the subset

	0	1	2	3	4	5	6
Forms	Las	tres	cosas	están	estrechamente	vinculadas	.
Lemmas	el	3	cosa	estar	estrechamente	vincular	.
PoS Tags	DA0FP0	Z	NCFP000	VAIP3P0	RG	VMP00PF	Fp
Syn Tags	espec:1	espec:2	subj:3	co-v:7	espec:5	att:3	?:7

Table 2: Tagging of the sentence *"Las tres cosas están estrechamente vinculadas."*

of words with the highest number of (nested) children within the target phrase. Table 3 features the rules used for selecting the relevant words.

This Feature Extraction method uses the dependency labels as a means of selecting only relevant examples. Take for instance the example sentences in Table 1. Given that the target word is an attribute, the subject is included as a relevant feature, as per the last rule in Table 3. Any example sentence in which there is no subject as the sibling of the target word (as is the case for the first example sentence in Table 1) will have an empty feature, which increases its likelihood of not being included in the training set of this sentence.

2.3 Test Data Pre-processing

The test data for this task is composed of bilingual input sentences, making it impossible to obtain a correct tagging or parsing. To overcome this issue, a two-stage process wherein the first stage obtains translations for the L1 portions was performed. These plausible translations are obtained by TiMBL using as features the neighboring words of the target phrases. Once the sentences are in a single language (L2) they are tagged and parsed syntactically. Finally, the second stage consists in applying the same feature selection algorithm proposed in subsection 2.2.

2.4 Translation Selection

The processing of each sentence consists of several steps. In the first step, the L1 target phrase is searched for in the phrase index Given an L1 phrase, a binary search algorithm iterates through the phrase index and returns an array of pointers[2] to the corpus. Then, a multi-threaded subroutine reads the word-aligned bilingual corpus and extracts all the referenced sentences. Thus, for each input sentence, a set of example bilingual word-aligned sentences is extracted from the corpus. Relevant features are extracted according to a syntactic analysis as explained in subsection 2.2, and written to text files in the C4.5 format. The features extracted from the example sentences, as well as the L2 translations of the target phrases in each sentence, are used as the training set for TiMBL, while the features extracted from the input sentence are used as its (singleton) test set.

The L2 translations of each target phrase in the example sentences are used as the classes for the training set, in order to turn a bilingual disambiguation problem into a machine learning classification problem. TiMBL learns how to classify the training feature vectors into their corresponding classes and then predicts the class for the test set feature vector, i.e. its most likely translation using an IBL algorithm (Aha et al., 1991), which is a variation of the k-nearest neighbor classifier.

3 System Submissions

We submitted three result sets for the English-Spanish language pair. Two of them were submitted for the 'Best' evaluation type, and the other one was submitted for the 'out-of-five' evaluation type. The difference between these two evaluation types is that out-of-five evaluation expects up to five different translations for every target phrase, while 'best' only accepts one. The evaluation metrics include accuracy and recall, and also a word-based special type of accuracy, which takes into account partially correct translations.

Of the two runs submitted in the 'Best' evaluation type, **Run1-best** (see table 4) used our proposed syntactic feature extraction method, while **Run2-best** used a regular 2-word window around the target phrase. For the **Run1-oof** we combined the two methods mentioned above with different values of k.

3.1 Results

The test data consisted in 500 sentences written in Spanish, with target English phrases. The official results obtained by our runs are shown in Table 4.

Our control run, **Run2-best**, yielded slightly

[2]Given that line breaks are just regular characters, what is actually referenced in the phrase index are byte offsets.

Case	Rule	Example
One of the *target words* is a subject.	Include any **sibling** which is an auxiliary or modal verb.	*Our cat* **quiere** comerse la ensalada.
The parent of one of the main *target words* is a coordinative conjunction.	Include its closest **sibling**.	No quería ni *eat*, ni **dormir**.
The parent of one of the main *target words* is a relative pronoun.	Include its **grandparent**.	No **creo** que ella *is coming*.
One of the *target words* is an attribute.	Include any **sibling** which is subject.	Mis **tías** están *very tired*.

Table 3: Relevant word selection rules.

better results than our experimental run, **Run1-best**. This means that our method of syntactic features extraction did not improve translation quality.

3.2 Error Analysis

By analyzing our results, we detected three groups of recurrent errors. The first group of errors is related to verb morphology, in which a single English verbal form corresponds to many Spanish verbal forms. In these cases, our system often outputs an infinitive form or a past participle instead of a finite verb.

The second group of errors we detected comprises incomplete translations. In these cases, a single word in English has a multiword Spanish translation, but our system often outputs a single-word translation.

The third group of errors are related to English words with multiple possible parts of speech, as 'flood', which can be a noun but also a verb. Our system tends to output nouns instead of verbs and vice versa.

4 Discussion

There are two main reasons as to why the syntactic feature extraction method did not work. The first reason is related to the nature of the task; the second is related to the scope of the method.

The fact that this task involved analyzing sentences partly written in two languages made syntactic analysis extremely difficult as dependencies span all over the bilingual sentence. The best solution we found for this was to divide the operation of the system in two stages, where the first one did not involve syntactic dependencies and provided a working translation, and the second one used this first translation to perform a syntactic analysis and then rerun the classification step. This, however, favored error propagation. Although translation quality did improve between the two stages, there were many cases in which a bad initial translation involved a bad syntactic analysis, which in turn resulted in a bad final translation.

A more sophisticated version of his method was initially developed for the English-Spanish language pair and involved several language-specific rules. However, we decided to make this method language-independent, so we simplified it to its actual version. This simplified version uses syntactic dependencies as feature selectors, but the features themselves are regular lemma/PoS combinations, which is not always the best feature choice.

5 Conclusion

Syntactic dependency relations are an important means of analyzing the internal structure of a sentence and can successfully be used to improve the feature selection process in WSD. However, syntactic parsing is far away from optimal in Spanish, a fortiori if it involves sentences written in two languages. For this kind of task, perhaps a statistical language model of L2 would have yielded better results.

Acknowledgments

We would like to specially thank Professor Silvia Takahashi of Universidad de los Andes for her continued advice and support. We would also like to thank Pedro Rodríguez for his development of the Linguee crawler, María De-Arteaga, Alejandro Riveros and David Hoyos for their useful suggestions in the conception and development of this project, and the rest of the UNAL-NLP team for

Run	Recall	Accuracy	Word Accuracy	Rank (runs)	Rank (systems)
Run1-best	0.993	0.721	0.794	5	2
Run2-best	0.993	0.733	0.809	4	2
Run1-oof	0.993	0.823	0.880	6	3

Table 4: Official results.

their interest and encouragement. Many thanks to Jay C. Soper for proof-reading this article.

References

David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Walter Daelemans, Jakub Zavrel, Ko Van der Sloot, and Antal Van den Bosch. 2010. Timbl: Tilburg memory-based learner. reference guide. ILK Research Group, Tilburg University.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, and Richard Zens. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, page 177–180.

Patrik Lambert, Adrià Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285, December.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. In *Second joint conference on lexical and computational semantics*, volume 2, page 158–166.

Wang Ling, Tiago Luís, João Graça, Luısa Coheur, and Isabel Trancoso. 2010. Towards a general and extensible phrase-extraction algorithm. In *IWSLT'10: International Workshop on Spoken Language Translation*, page 313–320.

Marina Lloberes, Irene Castellón, and Lluís Padró. 2010. Spanish FreeLing dependency grammar. In *LREC*, volume 10, page 693–699.

David Martinez and Eneko Agirre. 2001. Decision lists for english and basque. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, page 115–118.

David Martínez, Eneko Agirre, and Lluís Màrquez. 2002. Syntactic features for high precision word sense disambiguation. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, page 1–7.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2473–2479, Istambul, Turkey, May.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Language Resources and Evaluation Conference*, page 2214–2218, Istambul, Turkey, May.

Maarten van Gompel, Iris Hendrickx, Antal van den Bosh, Els Lefever, and Véronique Hoste. 2014. Semeval-2014 task 5: L2 writing assistant. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, Dublin, Ireland, August.

Maarten van Gompel. 2010. UvT-WSD1: a cross-lingual word sense disambiguation system. In *Proceedings of the 5th international workshop on semantic evaluation*, page 238–241, Uppsala, Sweden.