An automatic approach to create a sense tagged corpus for word sense disambiguation in machine translation

Lucia Specia
NILC/ICMC - Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400
São Carlos – SP, Brazil, 13560-970
lspecia@icmc.usp.br

Syllas Freitas de Oliveira Neto
DCAT - Uniara
Rua Carlos Gomes, 1338
Araraquara – SP, Brazil, 14801-340
syllasneto@terra.com.br

Maria das Graças Volpe Nunes
NILC/ICMC - Universidade de São Paulo
Av. do Trabalhador São-Carlense, 400
São Carlos – SP, Brazil, 13560-970
gracan@icmc.usp.br

Mark Stevenson
Department of Computer Science -
University of Sheffield
Regent Court, 211 Portobello Street
Sheffield, UK, S1 4DP
M.Stevenson@dcs.shef.ac.uk

Abstract

In this paper we describe a simple approach to the automatic creation of a sense tagged corpus intended for multilingual word sense disambiguation (WSD). The approach is based on English-Portuguese parallel corpora and a set of straightforward heuristics. In experiments with two corpora containing some verbs, a preliminary evaluation showed that, regardless of its simplicity, the proposed approach is quite promising. Besides the word senses tags, the resulting corpus provides other kinds of useful information for WSD, such as POS-tags. We plan to employ the corpus created in a supervised machine learning process in order to build a WSD model for machine translation.

1 Introduction

Word Sense Disambiguation (WSD) is an area of language processing concerned with the task of identifying the correct sense of an ambiguous word given its context. A “sense”, here, stands for one of the possible (related or unrelated) meanings of the word, all belonging to the same part of speech. For example, the noun “bank” has at least two unrelated meanings (homography): “financial institution” and “land along the side of a river, lake, etc.”. On the other hand, the verb “to run” has at least two possible related meanings (polysemy): “to move quickly” and “to travel”.

Although WSD can be thought of as an independent task, its importance is more easily realized if we consider its application to a specific task, such as Information Retrieval (IR) and Machine Translation (MT). In IR, WSD must identify the sense of the query words in order to retrieve only the documents related to the sense intended by the user. In MT, WSD must identify the correct translation for the ambiguous words, that is, solve the ambiguity between two languages. According to this concept of ambiguity between languages (Hutchins and Sommer, 1992), different senses of a word in the source language can be translated by different words in the target language; and a non-ambiguous word in the source language can have two or more possible translations in the target language. So, in this context, “sense” means, in fact, “translation”.

Since the earliest research in MT, which represents the focus of this paper, sense ambiguity has been thought of as one of the most (or the most) important problems of this area (Bar-Hillel, 1960). Nowadays, despite the great advances in WSD, this problem is considered by many authors as the main barrier to the progress in MT (Wilks, 1997). We recently investigated this problem considering MT from English to Portuguese. Our study (Specia and Nunes, 2004) has shown that the current MT systems do not appropriately handle the sense ambiguity problem and that this is one of the main causes of the very low quality resulting translations.

The various approaches that have been proposed to WSD are generally aimed at monolingual contexts, considering mainly the English language. Recent approaches have focused on the use of corpus-based and machine learning techniques in order to avoid the massive work of codifying linguistic knowledge. These approaches have shown good results, in terms of accuracy and coverage, especially those that follow the supervised learning. Indeed, according to evaluation exercises, such as SENSEVAL (Edmonds & Cotton, 2001), the supervised approaches are superior to the unsupervised ones. When considering MT, it is very difficult to imagine the effectiveness of an unsupervised approach, since the possible senses (translations) need to be previously determined, as stressed by Wilks & Stevenson (1997). However, it is worth noticing that the supervised approaches are dependent on a special kind of corpus in order to carry out the learning process, namely, a sense tagged corpus. The non-existence or inadequacy of such a corpus and the difficulty of its creation is the main drawback of those approaches. For some applications, mainly the monolingual (English) ones, there are already some functional sense tagged corpora, for example, SEMCOR (Miller et al., 1994) and DSO (Ng & Lee, 1996). For multilingual applications, in contrast, there are few corpora and the problem is even bigger due to the language-dependency issue: one corpus created to one pair of languages is of no use to a different pair. For the language pair English-Portuguese, addressed in this work, there is not a noteworthy corpus available. Certainly, the creation of an expressive sense tagged corpus would be a great step towards achieving effective WSD between this pair of languages. However, such a task would require a great deal of effort and time, if we consider the manual tagging of a large number of texts.

Considering these issues in the context of the wider purpose of developing a supervised WSD model for English-Portuguese MT, the aim of this work was to automatically create a sense tagged corpus specifically for this pair of languages. For that, we followed an approach based on parallel corpora and very simple heuristics. A preliminary evaluation of the resulting corpus has shown that it is a very promising approach.

In the rest of this paper we first discuss some alternative approaches for the creation of a sense tagged corpus (Section 2). We then present our approach, the resulting corpus and a preliminary evaluation (Section 3). Finally, we conclude with some remarks about this work and the next steps (Section 4).

2 Related work

The automatic creation of sense tagged corpus is one of the best strategies to minimize the knowledge acquisition bottleneck inherent to supervised approaches. However, it is still quite little explored (Agirre and Martínez, 2004). As a consequence, only a few related works can be cited. Not all of these works employ parallel corpora, and some of them consider a monolingual context.

Without considering parallel corpora and aiming at the creation of a monolingual tagged sense corpus, Agirre and Martínez (2004) employ a method based on Wordnet (Miller et al., 1990) monosemous relatives of the ambiguous words to generate sense tagged samples for these words. The relatives are the non-ambiguous synonyms of the ambiguous words. For each ambiguous word, the authors perform searches in corpora or in the web, using as queries the non-ambiguous synonyms, along with the context of the ambiguous word in a corpus. Their hypotheses is that, for each sense of an ambiguous word, if it possible to find a non-ambiguous synonyms of such sense, so the recovered samples containing that synonyms must be very similar to the sense under consideration. These samples could then be used to train a supervised model for that word sense.

Some examples of the use of parallel corpora in order to create a sense tagged corpus are the approaches of Diab & Resnik (2002) and Dinh (2002). Only the latter is designed for multilingual WSD. It is important to note that there are other works that use parallel corpora to train WSD models; however, they do not create a sense tagged corpus. Instead, they use statistical techniques to learn directly from the parallel corpus (Brown et al., 1991, for example).

Diab & Resnik apply a word-aligned bilingual parallel corpus and an inventory of senses to one of the languages (L1) with the intention of creating a sense tagged corpus to that language and, as a consequence, to the other language (L2). The parallel corpus is generated by means of a MT system. The correspondence between both languages is accomplished by automatic sentence and word alignment systems.

The identification of the word senses in L1 is carried out by firstly grouping all the words of this language that are possible translations of one L2 word. Next, a semantic similarity measure points to the most appropriate sense of the word, among those of the initial inventory, according to its similarity to the other words of the group (calculated considering their senses). Since the translations corresponding to each word in L1 are already known, the other language (L2) could be also tagged, considering the same senses.

Regardless of the advantages of a completely automatic process, it is worth commenting that both the MT and the alignment systems are not highly accurate, so that their errors can be propagated to the whole process, conveying possible problems to the resultant corpus.

Dihn (2002) builds a training corpus for WSD in MT from English to Vietnamese. The author employs a parallel manually sentence aligned corpus between these languages and a class-based word alignment method to find the correspondences between words and, at the same time, assigns each word a semantic class. The classes used are the ones of the LDOCE (*Longman Lexicon of Contemporary English*). Hence, the result is the English words annotated with the LDOCE semantic classes, which could also be undertaken to the Vietnamese words, according to the word alignment already established. So, although the approach uses a parallel corpus, it is not concerned with interlingual translation in the same sense we are. Actually, it is concerned with WSD in the source language (English, in this case).

In the following, we describe our approach for the creation of a bilingual sense tagged corpus, considering the ambiguity between the languages, as emphasized by Hutchins and Sommer (1992).

3 Creating the sense tagged corpus

A direct strategy for creating a sense tagged corpus for WSD in MT would be to use parallel corpora and identify correspondences between each word pair (in the source and target languages). This approach seems to be very practical, since the senses in MT are, in fact, the translations, thus it is not necessary (nor suitable) to disambiguate firstly in the source language and then in the target language.

When it is possible to apply an accurate word alignment method to the language pair under examination, the creation of the sense tagged corpus from parallel corpora can be quite simple. In such cases, a parallel corpus can also be directly used for training the MT system (Lee, 2002, for example). However, word alignment methods hardly present a satisfactory performance, especially in corpora of real translations, where correspondences are often not one-to-one. For English-Portuguese parallel texts, in particular, the performance of the most relevant alignment methods is still very low. The method with the best precision-recall relation, as reported by Caseli et al. (2004), has a precision of 57% and a recall of 61%.

So, in this work, we do not employ word alignment methods, since they certainly would convey serious errors to the tagged corpus. Another reason for not using word alignment methods is that we consider the disambiguation of only a few words, namely, some highly ambiguous verbs. So, a more specific strategy may allow more accurate results than the strategy adopted by the word alignment methods. The set of verbs, the rationale for choosing it, and the tagging process are described as follows.

3.1 Scope

As a prime delimitation of our work, we are concerned about only highly ambiguous verbs, whose ambiguity causes serious problems in the resulting translation. After the corpus study mentioned above (Specia and Nunes, 2004), we selected the seven most problematic English verbs, according to their frequency in a corpus, their number of possible translations to Portuguese, the inadequate handling provided by the MT systems and the harmful repercussion of this mishandling in the resultant translations. The following verbs were selected: to go, to get, to make, to take, to come, to look and to give. The average number of possible translations[1] of these verbs is 162.

This set of verbs constitutes the initial scope of this work; consequently, our corpus shall consist of sentences containing these verbs. In future work, we plan to expand this initial scope to a much larger set of words.

3.2 Source sentences

The original untagged corpus consists of sentences collected from two sources: the corpus Compara (Frankenberg-Garcia and Santos, 2003) and the corpus Europarl (Koehn, 2002). Compara comprises fiction books, originally in Portuguese or in English, and their manually elaborated translation to English or Portuguese, in a one-to-one correspondence. Although there are books translated to and from Portuguese from Brazil and Portugal, only the Brazilian ones were considered. Europarl comprises English and Portuguese (from Portugal)[2] versions of texts of the European Parliament. The sentences in both languages were automatically segmented and aligned, resulting in several non one-to-one correspondences. In order to use this corpus, we manually correct those cases.

Using two specific concordancers (one provided by the Compara and one specially developed), we select all the sentences from both corpora containing at least one of the seven verbs. The numbers of sentences for each verb in English (E) and Portuguese (P), as well as the total number of words, are shown in Table 1. The number of sentences in both languages is the same, so, only the total number (E and P) is shown.

Verb	Sentences from Compara (E and P)	Sentences from Europarl (E and P)	Total
go	2,000	46,848	48,848
get	1,662	15,542	17,204
make	1,590	94,426	96,016
take	1,530	84,480	86,010
come	1,688	28,748	30,436
look	1,474	15,734	17,208
give	1,108	49,946	51,054
Total	11,052	335,724	**346,776**
E Words	133,712	6,228,239	**6,361,951**
P Words	120,754	6,371,370	**6,492,124**

Table 1: Number of sentences and words

It is important to say that these are preliminary experiments; therefore, we plan to extend our corpus, incorporating sentences from other genres and domains.

3.3 Pre-processing

Several steps of pre-processing were carried out in

[1] Including those related to phrasal verbs for which the translation consists of only one Portuguese word. Expressions and multiword translations will be tackled in a future stage of this work.

[2] We considered Portuguese from both origins (separately) because we intend to analyze the influence of this feature in our WSD model.

order to transform the corpus into an appropriate format:

1. Tokenization of sentences in both languages;
2. POS tagging of sentences in both languages, using MXPOST (Ratnaparkhi, 1996);
3. Lemmatization of verbs and expressions including verbs of the Portuguese sentences;
4. XML annotation of sentences in both languages, using the Hofland's XML schema (Hofland, 1996).

Every verb, from each one of the two corpora, was handled separately. This decision is related to our WSD model purpose under development: we consider verbs separately because we intend to create a WSD model for each verb, and we consider corpora separately because we plan to compare the WSD models generated for different gender and domain corpora (also for different types of Portuguese).

The results of these pre-processing steps are 28 files, each one corresponding to one verb of one of the corpus and language, being all the sentences annotated according to the XML schema (which includes language, corpus, verb and sentence identifiers), all the words POS-tagged and the Portuguese verbs lemmatized.

3.4 Sense identification and tagging

Given that the corpus is in the appropriate format, the next step was the identification of the correct sense, that is, the translation of the verbs in question, followed by the annotation of those verbs in the English sentence with their corresponding translation. For that, the following assumptions were considered (assuming a parallel corpus correctly aligned at the level of the sentence):

- Since every English sentence has, in our parallel corpus, only one sentence[3] that represents its translation in Portuguese, the translation of the verb can only be found in that set of sentences;
- Every English verb has a pre-defined set of possible translations, including those referring to phrasal verbs, and this set can be defined by means of the translations given by common dictionaries;
- Phrasal verbs have specific translations; so whenever a verb occurs in such constructions, only the translations of the complete phrasal verb must be considered as possible;
- If there are two or more possible translations for an English verb, the more similar to the position of the English verb is the position of the translation, the more likely it is the correct one.

To define the set of possible translation for each verb (in fact, for its lemma), two English-Portuguese dictionaries were consulted: Houaiss (1982 edition) and Collins Gem (2001 edition). We only considered single words as translations; multiword translations will be tackled in a future phase. We also do not intend to automatically handle complex constructions, such as idiomatic expressions.

The occurrence of two or more possible translations for a single verb in the same sentence is quite frequent, since a sentence can have several verbs and the word used to translate some of the other verbs sometimes belongs to the set of possible translations of the verb under consideration. Besides, a sentence can have more than one occurrence of the same verb.

Given these assumptions, the translation (target-word) of an English verb (source-word) was identified through the following heuristics: for each occurrence of the source-word (i.e. a sentence), we sought for one of its possible target-words in the corresponding Portuguese sentence, based on its list of possible translations. We looked only at the lemmas of the words annotated as verbs, giving preference to the translations in a more similar position to the position of the source-word. If there is more than one occurrence of the same verb in a sentence, all of them are tagged individually.

For example, for the pair of sentences from the parallel corpus shown in Figure 1 (without any annotation), considering the verb "come" (position = 8), the system is going to correctly identify that the translation of the verb is "vir", the lemma of "vindo" (position = 9). It worth noticing that, according to the set of translations defined to the verb "come", two other translations in the sentence could be selected: "sair" (position = 2) and "ir (lemma of "for") (position = 6). However, the heuristic of the most similar position allows avoiding these wrong selections.

"I'd rather do without whatever I came for."
"Prefiro sair sem o que for que tenha vindo comprar."

Figure 1: An example of parallel sentences

Applied to our parallel corpus, our approach was able to determine a translation for 87% of the verbs of Compara and 70% of Europarl's. This could be considered a measure of the coverage of our approach in automatically tagging a corpus.

As we have predicted, the translations of some occurrences of the verb were not found. This can be due to four main reasons: (a) the incompleteness of our list of possible translations; (b) the fact that we did not

[3] Some sentences have more than one sentence as their translation, since the sentences can be segmented differently across the languages. However, every sentence or set of sentences in one language has one and only one corresponding sentence or set of sentences in the other language.

consider translations of expressions and translations realized as multiwords; (c) problems with the tools used in the pre-processing steps; and (d) problems with the parallel sentences.

The incompleteness of our list of translations is a consequence of the incompleteness of the dictionaries used (and of the dictionaries in general). In a subsequent stage of this work, we intend to consult more dictionaries, including dictionaries specialized in phrasal verbs, but we are aware that to develop a complete list is almost impossible, because of the dynamic nature of the language. Thus, some translations very possibly will remain out of our list, since they represent very new, rare or slang terms, not in dictionaries. The problems derived from the pre-processing tools could include wrong POS-tags and wrong lemmas. However, they are quite rare, since both the tools have been reported to perform well (over 95%). Finally, since both the source parallel corpora do not comprise literal translations, there are cases of omission and addition of words, as well as other changes in the translated sentences. More importantly, in Europarl there are several errors derived from the automatic segmentation and alignment process. Despite of our correction to assure the one-to-one correspondence, there are cases in which this correspondence is inadequate. In all of these cases, it is reasonable that the system does not select any translation as possible, avoid selecting a wrong one. This is the why the coverage of the system to Europarl is much lower then the coverage to Compara.

Before using the resulting corpus as training material to learn a WSD model according to a supervised process, we need to verify the accuracy of this corpus, since it will be the main source of information for that process.

3.5 Evaluation

In order to accomplish a proper evaluation of the sense tagged corpus, we should analyse the correctness of each identified translation, according to the translation in the corresponding parallel sentence. Although this process would be facilitated by the existence of the parallel translations, it still would require a great amount of time, which is just what we are trying to avoid with our automatic approach.

For this reason, we first accomplished a preliminary evaluation of our approach, taking a sample of the tagged corpus and manually analyzing it. If the accuracy shows to be too low, we will have to change our approach and will not have employed unnecessary efforts in its evaluation. Thus, we randomly selected 20 tagged occurrences of each verb, in both corpora (amounting to 280 sentences), and looked into their corresponding original pairs of sentences. We believe that this number of sentences is enough to give us an idea of the accuracy of our system in this first stage. The results are shown in Table 2.

Compara verb	**% correct**	**EPC verb**	**% correct**
go	80	go	55
get	70	get	85
make	90	make	80
take	80	take	85
come	75	come	70
look	95	look	100
give	90	give	95
Average	82.86%	**Average**	81.43%

Table 2: Accuracy of the sense tagging process

The accuracy measure used is the one traditionally exploited in WSD (sometimes called “precision”). It computes the proportion of senses correctly identified by the system over the total of senses identified.

As we can note in the Table 2, the results show that our approach is able to identify, in average, the correct senses of 82.86% of the verbs from Compara, and 81.43% from Europarl. Again, the results for Compara are slightly better, since there are fewer problems in the parallel sentences.

The tagging errors in both corpora are, in general, consequences of the four problems listed in the last section. If our list of possible translations is incomplete for a verb and so does not include the correct translation of a given verb occurrence, but there is one of the possible translations of the verb in the sentence, referring to another verb, this translation is (wrongly) chosen for the verb in consideration. The same occurs when there is an expression in the source sentence (without a corresponding translation). Some few errors are due to problems with the tools used in the pre-processing. However, most of the errors in the sense identification and tagging of the verbs are due to characteristics of the original parallel sentences. First of all, as we mentioned, they are not literal translations. For example, if the human translator does not translate, or changes, the part of the sentence that contains the verb, it is not possible for the system to identify the correct translation. When there is an intersection among the possible translations of the verbs in a sentence, a wrong choice can be made. Besides, some sentences of both corpora, mainly Europarl, have a huge number of words (for example, 179 words), making the problems even more difficult to solve. There were also some errors caused by the alignment of the original sentences, as mentioned above.

It is important to remember that the verbs considered are highly ambiguous and of very general use. So, it is inevitable that they have possible translations in common with other verbs. The previous sentence alignment certainly reduced the number of possible translations of each verb occurrence. However, after that alignment, each Portuguese sentence still presented, in average, three possible translations for the correspondent English verb occurrence (in some cases, there were eight possible translations).

To avoid the problem of selecting the wrong translation when the actual one is not in the sentence, a rationale strategy that we will investigate is to add more constraints to the search process, for example, considering the POS-tags of some words in the context of the sentences.

Despite the mentioned problems, the results are promising, considering that we employed only very simple heuristics. As the next steps, we intend to include new heuristics and evaluate the system again, analyzing more sentences, so that we can really get approximate measure of the sense tagging process.

4 Conclusion

In this paper we presented an approach to create a multilingual sense tagged corpus aimed at MT, based on parallel corpora. Our approach employs simple heuristics and language processing tools. The results of a preliminary evaluation, considering two parallel corpora and subset of verbs, showed that it is a quite promising approach, since most of the tagging errors are mainly due to problems in the original parallel corpora or relate to language construction that hardly could be handled by an automatic process. Nevertheless, the addition of more sophisticated heuristics could help avoid tagging in those cases, reducing the coverage of the system, but increasing its accuracy. The non-tagged samples could then be separately tackled. Although the tagging process would not be completely automatic, the manual effort would be significantly minimized. The investigation of such heuristics, as well as a more comprehensive evaluation of the system, is our next step.

It is worth noticing that the corpus created using our approach provides, in addition to the sense tags, other kinds of useful information for WSD: POS-tags and the neighbour words in the sentence. Therefore, through a feature extraction module, these information, as well as information about collocations and other sorts of co-occurrences, could be used to train a supervised machine learning algorithm in order to generate a WSD model, which is our main goal.

References

E. Agirre and D. Martínez. 2004. Unsupervised WSD Based on Automatically Retrieved Examples: The Importance of Bias. In *Proceedings of the Conference on Empirical Methods in NLP*.

Y. Bar-Hillel. 1960. Automatic Translation of Languages. *Advances in Computers*. Academic Press, New York.

P. F. Brown, S. A. Della Pietra, V.J Della Pietra, R. L. Mercer. 1991. Word Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting of ALC*, pages 264-270.

H. M. Caseli, A. M. P. Silva and M. G. V. Nunes. 2004. Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Texts. In *Proceedings of the 7th* SBIA, pages 184-193, LNAI 3171. Sao Luiz.

M. Diab and P. Resnik. 2002. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the 40th Anniversary Meeting of the ACL*. Philadelphia.

D. Dinh. 2002. Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation. In *Proceedings of Workshop on Machine Translation in Asia*, pages 26-32.

P. Edmonds and S. Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of the 2nd International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1-5.

A. Frankenberg-Garcia and D. Santos. 2003. Introducing COMPARA: the Portuguese-English Parallel Corpus. *Corpora in translator education*, pages 71-87. Manchester.

K. Hofland. 1996. A program for aligning English and Norwegian sentences. *Research in Humanities Computing*, pages 165-178. Oxford University Press, Oxford.

W. J. Hutchins and H. L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press, Great Britain.

P. Koehn. 2002. *Europarl: A Multilingual Corpus for Evaluation of Machine Translation*. (www.isi.edu/~koehn/publications/europarl).

H. Lee. 2002. Classification Approach to Word Selection in Machine Translation. In *Proceedings of AMTA'2002*, pages 114-123. Berlin.

G. A. Miller, R.T. Beckwith, C. D. Fellbaum, D. Gross and K. Miller. 1990. Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4), pages 235-244.

A. Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Empirical Methods in NLP Conference*. University of Pennsylvania.

L. Specia and M.G.V. Nunes. 2004. *A Ambigüidade Lexical de Sentido na Tradução do Inglês Para o Português – Um Recorte de Verbos Problemáticos*. Technical Report NILC-TR-04-01. São Carlos.

Y. Wilks. 1997. Senses and Texts. *Computers and the Humanities*, 31(2).

Y. Wilks and M. Stevenson. 1997. Sense Tagging: Semantic Tagging with a Lexicon. In *Proceedings of the SIGLEX Workshop 'Tagging Text with*

Lexical Semantics: What, why and how”.
Washington.