Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation

Lucia Specia, Maria das Graças Volpe Nunes

ICMC – University of São Paulo Av. do Trabalhador São-Carlense, 400 São Carlos, 13560-970, Brazil {lspecia,gracan}@icmc.usp.br

Abstract

We describe an approach to the automatic creation of a sense tagged corpus intended to train a word sense disambiguation (WSD) system for English-Portuguese machine translation. The approach uses parallel corpora, translation dictionaries and a set of straightforward heuristics. In an evaluation with nine corpora containing 10 ambiguous verbs, the approach achieved an average precision of 94%, compared with 58% when a state of the art statistical alignment tool was used. The resulting corpus consists of 113,802 instances tagged with the senses (i.e., translations) of the 10 verbs. Besides the word-sense tags, this corpus provides other useful information, such as POS-tags, and can be readily used as input to supervised machine learning algorithms in order to build WSD models for machine translation.

1 Introduction

Word Sense Disambiguation (WSD) is concerned with the identification of the sense of an ambiguous word in a given context, that is, one among its possible meanings. For example, the noun *pen* has at least two unrelated meanings: *writing device* and *enclosure*. The verb *to run*, in turn, has at least two possible related meanings: *to move quickly* and *to go*.

Although WSD can be thought of as an independent task, its importance is more straightforwardly realized when it is used in an application, such as Information Retrieval or Machine Translation (MT) (Wilks & Stevenson, 1998). In MT, which is the focus of this paper, WSD can be used to identify the most appropriate translation for a source language word when the target language offers more than one option with different meanings, but the same part-of-speech. However, there is not always a direct relation between the number of possible senses and translations of a word; different senses of a word in the source language can be translated by the same target word, and a non-ambiguous source word can have two or more possible translations (Hutchins & Somers, 1992). In

Mark Stevenson

DCS – University of Sheffield Regent Court, 211 Portobello Street Sheffield, S1 4DP, UK M.Stevenson@dcs.shef.ac.uk

this context, thus, "sense" means, in fact, "translation". For example, assuming the translation from English to Portuguese, explored in this work, bank can be translated as banco (financial institution or seat) or margem (land along the side of a river). Financial institution and land along the side of a river are both senses of the English word bank, however, the seat sense is valid only in the translation.

Sense ambiguity has been recognized as one of the most important problems in MT (Bar-Hillel, 1960). Nowadays, despite the great advances in WSD, this problem is still considered a serious barrier to the progress in MT. The problem was recently investigated for English-Portuguese MT (Specia, 2005). The study showed that the current MT systems do not handle sense ambiguity appropriately and that this is one of the reasons for the unsatisfactory translations.

The various approaches to WSD are generally aimed at monolingual contexts. Recent approaches have focused on the use of corpus-based and machine learning techniques in order to avoid the massive effort required to codify linguistic knowledge. These approaches have shown good results, especially those using supervised learning (Edmonds & Cotton, 2001). However, supervised approaches are dependent on a sense tagged corpus. The lack or inadequacy of such corpora is one of the main drawbacks of those approaches.

For monolingual applications, there are some available sense tagged corpora, such as SemCor (Miller et al., 1994). However, for multilingual applications there are only few corpora for certain languages. For English-Portuguese, in particular, there are no available corpora. The creation of an expressive corpus would represent an important step towards achieving effective WSD between this pair of languages. Certainly, automating this process would avoid the effort required to carry out manual tagging.

Although a good strategy, the automatic creation of sense tagged corpora is still little explored. Some approaches aimed at the creation of English tagged sense corpora include the work of Agirre & Martínez (2004), who exploited Wordnet relations and monolingual corpora, and Diab & Resnik (2002), who made use of bilingual parallel corpora and word alignment methods. Dinh (2002) also explored bilingual parallel corpora and word alignment methods to create an English-Vietnamese sense tagged corpus.

Given the large amount of multilingual machine readable texts currently available, identifying the correspondent word pairs in the source and target languages of parallel corpora seems to be a very practical strategy to automatically create sense tagged data. Parallel corpora are also good knowledge sources to directly carry out the sense disambiguation, especially for MT purposes. In fact, parallel corpora have been explored in several ways for MT since (Brown et al., 1991). They have also been used for monolingual WSD (Dagan & Itai, 1994, Ide et al., 2002; Ng et al., 2003).

Most of these works rely on the existence of accurate word alignment methods. However, current word alignment methods do not present a satisfactory performance, when applied to English-Portuguese. Indeed, experiments with several alignment methods on English-Portuguese reported a precision of 57% and a recall of 61% for the best method (Caseli et al., 2004).

Considering these issues in the context of our ultimate goal of building a WSD system for English-Portuguese MT, we developed a hybrid approach, mixing linguistic and statistical knowledge, to automatically create a sense tagged corpus. The approach makes use of parallel corpora, bilingual dictionaries, and a set of simple heuristics. We experimented with nine parallel corpora containing 10 ambiguous verbs, and compared the results to those produced by the word alignment tool GIZA++ (Och & Ney, 2003).

In the remaining of this paper, we first present our approach, including its scope, the parallel corpora explored, and the sense tagging process (Section 2). We then present the evaluation of the approach, discussing its results (Sections 3 and 4), and conclude with some remarks and future work (Section 5).

2 The sense tagging approach

2.1 Scope

This work focuses on verbs; these represent difficult cases for WSD and, once disambiguated, can help to disambiguate other words in the sentence, especially their arguments. In this stage, we are dealing with seven frequent and highly ambiguous verbs identified as very problematic to MT systems according to a previous study (Specia, 2005). We also consider other three frequent but not so ambiguous verbs. These three verbs were selected in order to analyze the effect of polysemy level on our method. The complete list of verbs, along with their number of possible translations², is given in Table 1.

Possible translations are single words, including synonyms, and phrasal verb usages. Phrasal verb senses are considered because the occurrence of a verb followed by a preposition / particle does not necessarily indicate a phrasal verb. Multiword translations are not considered for these experiments and will be tackled in future work. The average number of translations for the seven highly ambiguous verbs (come, get, give, go look, make and

take) is 203. The average for the three other verbs (ask, live and tell) is 19.

Verb	# translations	Verb	# translations
come	226	make	239
get	242	take	331
give	128	ask	16
go	197	live	15
look	63	tell	28

Table 1: Verbs and its possible translations

2.2 Parallel corpora

The original untagged corpus, consisting of English sentences containing the 10 verbs along with their manually translated Portuguese sentences, was collected from nine sources, including a mixture of genres and domains, as shown in Table 2. Europarl (Koehn, 2002) comprises bilingual versions of the European Parliament texts. Compara (Frankenberg-Garcia & Santos, 2003) comprises fiction books. Messages contains input / output messages used by Linux software³. Bible contains versions of the Christian Bible. Red Badge is the novel The Red Badge of Courage, by Stephen Crane. PHP consists of the user manual to the PHP programming language⁴. ALCA comprises bilingual versions of documents from Free Trade Area of the America⁵. NYT comprises some on-line daily news of the New York Times newspaper. Finally, CPR consists of 65 abstracts of Computer Science thesis from the University of São Paulo.

All these corpora were already sentence aligned. Sentences in a many-to-one or one-to-many relationship with sentences in the translation were grouped together to form a "unit". So, the number of units is the same for both languages. Using specific concordancers, we selected the sentences from these corpora containing one of the 10 verbs. The number of resulting units (in one language), and English (E) and Portuguese (P) words for each corpus are illustrated in Table 2.

Corpus	# units	# E words	# P words
Europarl	167,339	6,193,904	6,299,686
Compara	19,706	518,710	475,679
Messages	16,844	385,539	394,095
Bible	15,189	474,459	443,349
Red Badge	823	15,172	12,555
PHP	226	7,964	6,342
ALCA	191	7,478	7,386
NYT	47	1,585	1,575
CPR	41	1,339	1,381
Total	220,406	7,606,150	7,642,048

Table 2: Numbers of sentences and words

The proportion of units for each verb varies from corpus to corpus. The smallest corpora did not contain any occur-

 $^{^{\}rm l}$ According to the frequency list of the British National Corpus (Burnard, 2000).

² According to the DIC Prático Michaelis® machine readable English-Portuguese dictionary, version 5.1.

³ www.gnome.org

⁴ www.php.net/download-docs.php

⁵ www.ftaa-alca.org/alca_p.as

⁶ www.nytimes.com

rences of some verbs.

2.3 Pre-processing

Some pre-processing steps were carried out to filter units and to transform the corpus into an adequate format:

- 1. English units were lemmatized using the Minipar parser (Lin, 1993).
- 2. Unit pairs containing English idioms involving one of the 10 verbs were eliminated.
- 3. POS tag the units in both languages, using the Mxpost tagger (Ratnaparkhi, 1996).
- 4. Portuguese verbs and verbal expressions were lemmatized (Feltrim, 2004).
- 5. Pairs of units for which the English verb under consideration has no valid verb tag in the English unit were eliminated; likewise, when the Portuguese unit has no word with a verb tag.

Units containing idioms were eliminated to avoid tagging errors, since idiom translations are usually non-literal. For that, we created a list of idioms containing the verbs based on the on-line version of the Cambridge Dictionary of Idioms⁷.

The filter of the fifth step intended to isolate cases referring to tagger and concordancer problems, as well as to avoid errors due to modified translations, that is, when the verb in the English unit was paraphrased by words other than verbs.

The units from each of the corpora were handled separately, since we intend to analyze the genre / domain influence in our WSD model. The outputs of the preprocessing steps are English and Portuguese filtered units, being all words POS tagged, and English words and Portuguese verbs lemmatized. The total number of sentences was 206.913.

2.4 Sense identification

In order to identify the translation of each verb occurrence, the following assumptions were made:

- Given a sentence aligned parallel corpus, the translation of the verb in an English unit can be found in its corresponding Portuguese unit.
- Every English verb has a pre-defined set of possible translations, including those referring to phrasal verbs, and this set can be extracted from bilingual dictionaries.
- Phrasal verbs have specific translations; so, if a verb occurs in such constructions, the translations of the complete construction should be considered first. Some verb plus particle / preposition constructions may also be used as non-phrasal verbs. In this case, the translations of the verb itself should be also considered.
- Translations have different probabilities of being used in a given corpus, and these probabilities can be identified through a statistical co-occurrence analysis of the corpus.
- 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 in its respective unit, the more likely it is the correct one.

Based on these assumptions, a sense tagging process was created, relying in the following resources and heuristics.

2.4.1 Resources

To define the set of possible single-word translations for each verb, we used machine readable versions of bilingual dictionaries. We used the same dictionaries to identify a list of phrasal verbs and their translations. We consulted the on-line version of the Cambridge Dictionary of Phrasal Verbs⁸ in order to create lists of separable and inseparable phrasal verbs, that is, phrasal verbs that can and can not have words between the verb and the particle. We consulted occurrences of each construction in the British National Corpus to elaborate a list of verbs plus particles / prepositions that can be used both as phrasal verb and as non-phrasal verb.

The NATools package (Simões & Almeida, 2003) was used to produce a list of translation probabilities. NATools uses statistical techniques to create bilingual dictionaries from sentence aligned parallel corpora. It generates bidirectional lists of at most 20 possible translations for all the words in the parallel corpus, along with their probabilities. Although the tool does not make use of any language-dependent resource, we pre-processed the parallel corpora in order to improve the produced dictionaries. Processing the units for all verbs in a given corpus together, we performed the following steps:

- 1. POS tag units in both languages.
- 2. Lemmatize English (Lin, 1993) and Portuguese verbs (Feltrim, 2004).
- Eliminate the unit pairs containing idioms in the English version, using the list of idioms previously mentioned
- 4. Remove stop words, punctuation, and other symbols from units in both languages.

In Table 3 we illustrate the list of translation probabilities produced by NATools for the verb *to give*, in the Compara corpus.

Translation	Prob.	Translation	Prob.
ceder_v	0.0117	lançar_v	0.0131
devolver_v	0.0053	pergunta	0.0063
\(null\)	0.1520	entregar_v	0.0252
renunciar_v	0.0055	provocar_v	0.0077
desistir_v	0.0225	fazer_v	0.0309
soltar_v	0.0060	dar_v	0.5783
deixar_v	0.0065	ser_v	0.0230
receber_v	0.0079		

Table 3: Translation probabilities for to give

In general, the lists produced contain mostly verbs appropriate as translations (bold face in Table 3), but also some

⁷ http://dictionary.cambridge.org/default.asp?dict=I

⁸ http://dictionary.cambridge.org/default.asp?dict=P

verbs that are not possible translation according to our dictionary (other words with a $_{\nu}$ tag), words with other POS, and a null translation probability, that is, the probability of the verb not being translated. Since we assume that at least one possible translation of the verb is in the Portuguese unit, we normalized the resulting list to eliminate the null translation probability.

The lists produced do not include all the possible translation belonging to our dictionaries, because many of them may not occur in the corpus, or may occur with a very low frequency. For those translations, we assigned a zero probability.

Since the probabilities vary from corpus to corpus, the translation probabilities were generated individually for each corpus.

2.4.2 Heuristics

Given the assumptions and the resources created, we defined a set of heuristics to find, in the Portuguese unit (PU), the most adequate translation for each occurrence of the verb in an English unit (EU). The general procedure is shown in Figure 1. In detail, the heuristics comprises the following steps:

- Identify inseparable phrasal verbs in the EU, annotating the unit when they occur. We compare the lemmas of the words tagged as verbs and the following 1-5 words to the list of inseparable phrasal verbs.
- 2. Identify, in the remaining EUs, separable phrasal verb, annotating the unit when they occur. Again, we compare the lemmas of the words tagged as verbs and the following 1-8 words to our list of separable phrasal verbs, allowing 2-3 words between the verb and the particle. We assume the remaining EUs do not contain any phrasal verb.
- 3. Identify the absolute positions of the verb / phrasal verb in the EU, ignoring punctuation signals and other symbols.
- 4. In the verb lemmas of the PU, search for all possible translations of the verb, consulting specific dictionaries for inseparable, separable, or non-phrasal verbs. Three possible situations arise:
 - a. No translation is found go to step 5.
 - Only one translation is found go to step 6.
 - c. Two or more translations are found go to step 7.
- 5. If the occurrence is a non-phrasal verb, finalize the process, considering that no adequate translation was found. Otherwise, first verify if the verb plus particle / preposition can be used as non-phrasal verb. If yes, go back to the step 4, now looking for possible translations of the verb in the non-phrasal dictionary. If it can not be used as a non-phrasal verb, finalize the process, considering that no adequate translation was found.
- 6. Select the only possible translation and use it to annotate the EU.
- Identify the absolute positions of each translation in the PU and assign a position weight (PosW) to the translation, penalizing translations in distant

positions from the position of the EU verb, according to the following:

$$PosW = 1 - \left(\frac{|EUposition - PUposition|}{10}\right)$$

8. Verify the translation probability for each of the possible translation, calculating the final translation weight (TraW) as follows:

$$TraW = PosW + probability$$

9. Choose the translation with the highest weight (TraW) to annotate the EU.

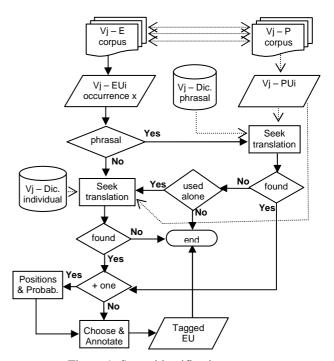


Figure 1: Sense identification process

The position plus probability weighting schema adopted in the case of more than one possible translation was empirically defined after experimenting with different schemas. As an example of its use, consider the pair of sentences shown in Figure 2, for *to come* (EU position = 7). The system correctly identifies the translation as *vir*, the lemma of *vindo* (PU position = 9, PosW = 0.8, probability = 0.432, TraW = 1.232), although there are two more possible translations in the sentence, according to our list of possible translations: *sair* (PU position = 2, PosW = 0.5, probability = 0.053, TraW = 0.553) and *ir* (lemma of *for*) (PU position = 6, PosW = 0.9, probability = 0.04, TraW = 0.94). If we had considered only the position of the words, without the weighting schema, the system would have chosen the wrong translation: *ir*.

- "I'd rather leave without whatever I came for."
- "Prefiro sair sem o que for que tenha vindo buscar."

Figure 2: Example of parallel sentences

It is worth noticing that the word position plays the most important role in this example. The probabilities generally take effect when the possible translations are close to each other.

3 Evaluation and discussion

Our approach determined a translation for 55% of all verbs (113,802 units) in the nine corpora (Table 2). Similar identification percentages were observed among verbs and corpora. The lack of identification for the remaining occurrences was due to three main reasons: (a) we do not consider multi-word translations; (b) errors from the tools used in the pre-processing steps, especially POS tagging errors; and (c) modified translations, including cases of omission and addition of words.

Although the coverage of our approach in automatically tagging a corpus can be considered low, it is important to mention that we give preference to the precision of the sense tagging to the detriment of wide coverage. Our intention is to use this corpus to train a WSD model and we therefore require data to be as accurate as possible.

In order to estimate the precision of the sense tagging process, we randomly selected 30 tagged EU from each corpus, for each verb, including units without phrasal verbs and with both kinds of phrasal verbs. We grouped the five smallest corpora (Miscellaneous) for this evaluation. The total number of evaluated units was 1,500. The precision for each corpus and verb is shown in Table 4.

Verb	Europarl	Compara	Messages	Bible	Misc.
come	80%	84%	95%	90%	91%
get	93%	87%	100%	95%	82%
give	97%	95%	95%	97%	93%
go	90%	90%	95%	85%	95%
look	100%	98%	95%	90%	100%
make	87%	86%	100%	93%	97%
take	80%	88%	91%	90%	93%
ask	100%	98%	100%	100%	100%
live	100%	100%	100%	100%	100%
tell	100%	94%	100%	100%	96%
Ave.	93%	92%	97%	94%	95%

Table 4: Precision of the sense tagging process

On average, our approach was able to identify the correct senses of 94.2% of the analyzed units. It achieved a very high average precision (99.2%) for the less ambiguous verbs (the three last in Table 4). Of the seven highly ambiguous verbs, to look and to give have lower numbers of possible senses than the rest, and for them the system also achieved a very high average precision (96%). For the remaining five verbs, the system achieved an average precision of 90.3%. Therefore, although there is no direct relation between the number of senses and the precision, the precision was generally lower for the most ambiguous verbs

The tagging errors were consequences of the problems mentioned above, regarding the coverage of the system, but were also due to limitations of our heuristics. The distribution of the errors sources for each corpus is shown in Table 5.

Corpus	Idiom / slang	Modified translation	Tagger error	Heuristics
Europarl	6%	66%	8%	20%
Compara	8%	71%	0%	21%
Messages	0%	100%	0%	0%
Bible	6%	74%	10%	10%
Mics.	10%	69%	16%	5%

Table 5: Tagging error sources

Most of the errors were due to modified translations, including omissions and paraphrases (such as active voice sentences being translated by different verbs in a passive voice). In fact, with exception of the technical corpora (Messages and PHP), the translations were far from literal. In those cases, as in the case of idioms or slang usages, the actual translation was not in the sentence, or was written using words that were not in the dictionary, but the system found other possible translation, corresponding to other verb. Tagger errors refer to the incorrect tagging of the verbs with any other POS. In this case, the system also pointed out other possible translations in the PU. Errors due to the choices made by our heuristics are also related to the other mentioned errors. For example, considering the position of the words as the main evidence can be an inappropriate strategy when translations are modified by the inclusion or omission of words.

It is important to remember that some units are very long (for example, 180 words), containing many possible translations. In fact, an EU can have many verbs and the words used to translate other verbs may also be translations of the verb under consideration. The sentence alignment certainly reduced the number of possible translations, however, even after that process, the average number of possible translations in a PU, in all corpora and for all verbs, was 1.5. If we consider only the seven most ambiguous verbs, the average was 2.4 (from 1 to 15 possible translations in a PU).

4 Comparison with an alternative approach

We compared the precision of the system to the precision of the GIZA++ word alignment package (Och & Ney, 2003). Every pre-processed corpus was individually submitted to GIZA++ (the five smallest corpora were grouped in order to provide enough data for the statistical processing). We then analyzed the alignment produced for the verbs using the same sentences used to evaluate our system. The average precision for each corpus is shown in Table 6.

We considered as correct alignments all those including the verb translation, even if they were not one-to-one, that is, if they included other words. As shown in Table 6, the precision of the alignment produced by GIZA++ is considerably lower than the precision of our system. Unsurprisingly, the difference between the performances of the two approaches is statistically significant (p < 0.05, Wilcoxon Signed Ranks Test). Since statistical evidence is the only information used by GIZA++, it was not successful in identifying non-frequent translations. More-

over, it rarely found the correct alignment in the case of modified translations.

Corpus	Precision
Europarl	51%
Compara	61%
Messages	70%
Bible	42%
Miscellaneous	66%

Table 6: Precision of the GIZA++ word alignment

It is important to note that in this analysis we considered only the cases for which our system had proposed a possible translation. As previously mentioned, filters were used to avoid tagging errors. In order to find out GIZA++ outputs for those cases that were not tagged by our system, we analyzed 10 cases, for every verb and corpus, amounting to 500 parallel units. In average (all verbs and corpora), only 1% of these non-tagged units corresponded to GIZA++ null alignments for the verb. In 29% of the cases GIZA++ produced a correct alignment; while in 70%, the alignment pointed was incorrect. Although we analyzed the pre-processed corpora, again, in most of the cases, the incorrect GIZA++ alignments were due to modified translations. In those cases, the actual translation was not in the sentence, but the alignment system indicated a non-null alignment, since it does not include any linguistic knowledge about possible translations.

This comparison shows that the precision of our approach is, indeed, superior to those of the most relevant current word-alignment methods. It also shows that the use of the dictionaries avoided many tagging errors. Moreover, though our approach uses statistical information as one of the clues during the tagging process, it will still work if that information is not available. As a consequence, the performance for very small corpora will not be severely affected. So, we believe that the precision achieved by our system is satisfactory and that the resulting instances are thus appropriate to be used as a training corpus to produce WSD models.

5 Conclusion

We presented an approach to create a sense tagged corpus aimed at MT, based on parallel corpora, linguistic knowledge and statistical evidence. The results of an evaluation using a subset of nine parallel corpora and 10 verbs showed that the approach is effective, achieving an average precision of 94%. Most of the tagging errors were related to characteristics of the corpora: non-literal translations and use of language constructions that are very difficult to process automatically (idioms, e.g.). Nevertheless, the use of filters and elaborated heuristics avoided many errors, reducing the coverage of the system, but increasing its precision.

The resultant corpus of 113,802 instances provides, in addition to the sense tags, other kinds of useful information: POS-tags, lemmas and the neighbour words. This corpus will be used to train a supervised machine learning algorithm in order to produce a WSD model.

Although applied to a small set of words, the approach

can be extended to wider contexts. Besides the parallel corpora, the required resources can be extracted from machine readable sources. In addition, the evaluation reported here was carried out on difficult cases, and thus the results on other lexical items are likely to be as good, if not better, than those reported.

In future work, we will experiment with different weighting schemes, in order to explore more deeply the statistical analysis of the parallel corpora. We plan to consider as possible translations also those indicated by the statistical analysis, but which are not included in the bilingual dictionaries. With this, we hope to minimize the dependence on the knowledge resources and allow unusual, but valid, translations to be identified.

References

- (Agirre & Martínez, 2004) E. Agirre, D. Martínez. Unsupervised WSD Based on Automatically Retrieved Examples: The Importance of Bias. In *Proceedings of the Conference on Empirical Methods in NLP*, pp. 25-32, 2004.
- (Bar-Hillel, 1960) Y. Bar-Hillel. Automatic Translation of Languages. *Advances in Computers*. Academic Press, New York, 1960.
- (Brown et al., 1991) P.F. Brown, S.A. Della Pietra, V.J Della Pietra, R.L. Mercer. Word Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting of ALC*, pp. 264-270, 1991.
- (Burnard, 2000) L. Burnard. Reference Guide for the British National Corpus. Oxford University Press, 2000.
- (Caseli et al., 2004) H.M. Caseli, A.M.P. Silva, M.G.V. Nunes. Evaluation of Methods for Sentence and Lexical Alignment of Brazilian Portuguese and English Parallel Texts. In *Proceedings of the 7th* SBIA, Sao Luiz, pp. 184-193, 2004.
- (Dagan & Itai, 1994) I. Dagan, A. Itai. Word Sense Disambiguation Using a Second Language Monolingual Corpus. Computational Linguistics, 20:563-596, 1994.
- (Diab & Resnik, 2002) M. Diab, P. Resnik. An Unsupervised Method for Word Sense Tagging using Parallel Corpora. In *Proceedings of the 40th Anniversary Meeting of the ACL*, Philadelphia, 2002.
- (Dinh, 2002) D. Dinh. Building a training corpus for word sense disambiguation in the English-to-Vietnamese Machine Translation. In *Proceedings of Workshop on Machine Translation in Asia*, pp. 26-32, 2002.
- (Edmonds & Cotton, 2001) P. Edmonds, S. Cotton. SENSEVAL-2: Overview. In Proceedings of the 2nd Workshop on Evaluating Word Sense Disambiguation Systems, pp. 1-5, 2001.
- (Feltrim, 2004) V.D. Feltrim. Uma abordagem baseada em córpus e em sistemas de crítica para a construção de ambientes Web de auxílio à escrita acadêmica em português. Tese de Doutorado, Universidade de São Paulo, São Carlos, 2004.
- (Frankenberg-Garcia & Santos, 2003) A. Frankenberg-Garcia, D. Santos. Introducing COMPARA: the Portuguese-English Parallel Corpus. Corpora in translator education, pp. 71-87, 2003.
- (Hutchins & Somers, 1992) W.J. Hutchins, H.L. Somers. An Introduction to Machine Translation. Academic Press, UK, 1992.
- (Ide et al., 2002) N. Ide, T. Erjavec, D. Tufis. Sense Discrimination with Parallel Corpora. In *Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Philadelphia, pp. 56-60, 2002.
- (Koehn, 2002) P. Koehn. Europarl: A Multilingual Corpus for Evaluation of Machine Translation, 2002.,

- (www.isi.edu/~koehn/publications/europarl).
- (Lin, 1993) D. Lin. Principle based parsing without overgeneration. In *Proceedings of the 31st Annual Meeting of the ACL*, Columbus, pp. 112-120, 1993.
- (Miller et al., 1994) G.A. Miller, M. Chorodow, S. Landes, C. Leacock, R.G. Thomas. Using a Semantic Concordancer for Sense Identification. In *Proceedings of the ARPA Human Language Technology Workshop ACL*, Washington, pp. 240-243, 1994.
- (Ng et al., 2003) H.T. Ng, B. Wang, Y.S. Chan. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proceedings of the ALC-2003*, Sapporo, pp. 455-462, 2003.
- (Och & Ney, 2003) F.J. Och, H. Ney. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1):19-51, 2003.
- (Ratnaparkhi, 1996) A. Ratnaparkhi. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Conference on Empirical Methods in NLP*, Pennsylvania, 1996.
- (Simões & Almeida, 2003) A.M. Simões, J.J. Almeida. NA-Tools - A Statistical Word Aligner Workbench. In Proceedings da Sociedade Española para el Procesamiento del Lenguaje Natural, Madrid, 2003.
- (Specia, 2005) L. Specia. A Hybrid Model for Word Sense Disambiguation in English-Portuguese Machine Translation. In *Proceedings of the 8th Research Colloquium of the UK Special-interest Group in Computational Linguistics*, Manchester, pp. 71-78, 2005.
- (Wilks & Stevenson, 1998) Y. Wilks, M. Stevenson. The Grammar of Sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering*, 4(2):135-144, 1998.