International Journal of Computational Linguistics and Applications, Vol. 6, No. 1, 2015, pp. 131–148
Received 27/02/2015, Accepted 23/05/2015, Final 18/06/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Exploratory Study of Word Sense Disambiguation Methods for Verbs in Brazilian Portuguese

MARCO ANTONIO SOBREVILLA CABEZUDO
THIAGO ALEXANDRE SALGUEIRO PARDO
Universidade de São Paulo, São Paulo, Brasil

ABSTRACT

Word Sense Disambiguation (WSD) aims at identifying the correct sense of a word in a given context. WSD is an important task for other applications as Machine Translation or Information Retrieval. For English, WSD has been widely studied, obtaining different performances. Analyzing by morphosyntactic class, Verb is the hardest class to be disambiguated. Verbs are an important class and help to the sentence construction. Studies show that the disambiguation of verbs brings improvements into other applications. For Portuguese, there are few studies about WSD and, recently, these have been focused on general purpose. In the present paper, we report an exploratory study of knowledge-based Word Sense Disambiguation methods for verbs in Brazilian Portuguese, using WordNet-Pr (for English) as sense repository; and a comparison with the results obtained for nouns. The results show that, both All-words and Lexical sample evaluation, no methods outperformed the baseline. However, the multi-document scenario helped the WSD task.

Keywords: Word sense disambiguation, knowledge-based methods, Brazilian Portuguese

1. INTRODUCTION

Semantics is a deep linguistic knowledge level [1] and it is a popular subject in the Natural Language Processing community.

One of the most important problems related to Semantics is the ambiguity and, specifically, Lexical Ambiguity. Lexical Ambiguity occurs when a word may express two or more senses in a determined context. Lexical Ambiguity may be expressed in various difficulty levels. For example, consider the following four sentences:

- *homemcontou o número de pessoasqueficaramferidas* ("The man counted the number of people who were injured.").
- *O jogador bateu na bola com força* ("The player kicked the ball strongly.").
- *lutadorbateu as botas* ("The fighter died." or "The fighter kickedthe bucket.").
- *banco quebrounasemanapassada* ("The seat broke last week."or "the bank failed last week.").

In the first example, the sense of the verb "*contar*" may beeasily identified (to determine the total number of a collection of items); in the second example, the sense of the verb "*bater*" is easily identified too (to kick something); but, in the third example, the sense of the same verb may be difficult to identify, it could mean"to die" if we consider the expression "*bater as botas*", or could mean "to kick something" if we consider only the verb "*bater*"; finally, in the last example, it is necessary that we have more context and world knowledge, therefore, it is hard to identify the sense of the verb "*quebrar*", that could mean "to break an artifact (like seat)" or "to failfinancially (financial institution)".

Word Sense Disambiguation (WSD) aims atidentifying the correct sense of a word within a given context using a pre-specified sense repository [2]. WSD is considered an important task of other applications, as Information Retrieval, Machine Translation and Sentiment Analysis.

For English, there are many studies about WSD using different approaches and techniques [3]. Recently, knowledge-based WSD methods have become very popular [4]. This is due to the increase of the need of General Purpose WSD methods, which are able to be integrated into any application and domain. Despite the increasing of this popularity, studies have shown that WSD is a hard task to be solved and the obtained performance is not high. Analyzing by morphosyntactic class, verb is the hardest class to be disambiguated. The main reason for this difficult in verbs is that WSD methods use, generally, surface information to disambiguate a word, but verbs need syntactic and semantic information to get better results.

According to [5], verbs are an important class and help to the sentence construction. Studies show that the disambiguation of verbs brings improvements into other applications as Semantic Role Labeling [6].

For Portuguese, there are few studies and some of these are domain-oriented [7] [8], and thiscannegatively influence other Natural Language Processing applications. Recently, general purpose WSD methods have been investigated with the purpose of integrating these in other applications. We can mention the studies proposed in [9] (focused on nouns) and [10] (focused on verbs).

In this paper, we present an exploratory study of knowledge-based WSD methods (specifically, based on word overlapping, web search, graphs and a method for disambiguation in multi-document scenario) for verbs in BrazilianPortuguese, using WordNet-Pr [11] as sense repository and WordReference®[1] as bilingual dictionary; and a comparison with the results obtained for nouns[9].

The results show that, in All-words evaluation, no methods outperformed the baseline and, in Lexical sample evaluation, the multi-document scenario helped to outperform the baseline. A comparison with the WSD for nouns was performed and the results agreed with the literature, i.e., WSD for verbs is more difficult than WSD for nouns.

The remainder of this paper is structured as follows: Section 2 introduces concepts related to WSD and an overview of the related works for Brazilian Portuguese; the adaption of WSD classic methods for Brazilian Portuguese and the implemented

[1] http://www.wordreference.com/

methods are presented in Section 3; Section 4 shows the performance of the different WSD methods and a comparison between results obtained for verbs and nouns; finally, there are concluding remarks and an outlook of future work in Section 5.

2. CONCEPTS AND RELATED WORK

As we mentioned, WSD aims at selecting the correct sense of a word within a given context using a pre-specified sense repository [2].Basically,WSD methods(a) receive a target word (to disambiguate), a context (words around the target word) and a sense repository (this can be dictionaries, thesaurus, ontologies or wordnets) as input, and (b)execute the automatic disambiguation, showingthe correct sense for the target word as output [1].

The WSD task may be seen in two ways: (1) disambiguating a limited sample of content words in a text, called Lexical sample task; and (2) disambiguating all content words included in a text, called All-words task.

According to the use of resources and techniques, WSD methods can be classified as knowledge-based, corpus-based and hybrid methods [2]. Knowledge-based methods use linguistic resources and similarity measures to disambiguate a wide range of words. This approach is useful for All-words task (because of the use of broad linguistic resources) but the performance obtained by these methods are not so good. Corpus-based methods use sense-annotated corpus to yield machine learning classifiers. This approach is useful for the Lexical sample task (because the number of words to disambiguated is limited by the corpussize) and the performance of these methods is better than the Knowledge-based methods. Finally, hybrid methods use techniques from knowledge and corpus-based methods.

For Portuguese, there are few studies and some of these are domain-oriented. This may negatively influence other Natural Language Processing applications. Recently, General Purpose WSD methods have been proposed. Below, we briefly show some of the main related works for Portuguese that support this investigation.

In [7], a WSD method based on Inductive Logic Programming for Machine Translation task is proposed. Inductive Logic Programming is characterized by using machine learning methods and propositional logic rules. This method was focused on the disambiguation of 10 English verbs with high polysemy (to ask, come, get, give, go, live, look, make, take, and tell) to their respective Portuguese verbs. The author performed some experiments and showed that the proposed method outperformed the most frequent translation method and other methods based on machine learning.

In [8], a geographical disambiguation method for disambiguating place names is presented. This method used an ontology created in this work, called OntoGazetter, as knowledge base. This ontology is composed by place concepts. The results showed that OntoGazetter positively contributes to geographical disambiguation.

The first research on general purpose WSD methods for Brazilian Portuguese is presented in [9]. The authors investigated Knowledge-based WSD methods for common nouns, using WordNet-Pr[10] as sense repository and WordReference® as bilingual dictionary (since the language was Portuguese). In this work, besides the investigation of WSD methods, the author proposed a WSD method based on co-occurrence graphs and a variation of Lesk algorithm [12] for multi-document scenario. The results showed that, although the method does not outperform the baseline (most frequent sense), it contributes to the Word Sense Disambiguation in a multi-document scenario.

In [11], two verb sense disambiguation methods for European Portuguese were developed, using ViPer[13] as sense repository. The proposed methods were based on rules, machine learning and, finally, a combination of the best results of both. The baseline was the most frequent sense method and this was difficult to be outperformed, thus, a combination of methods was performed to outperform the baseline.

3. WORD SENSE DISAMBIGUATION

3.1. *Previous considerations*

A previous step to implement the methods was the choice of the sense repository. For Portuguese, there are some sense repositories as WordNet-Br [14], OpenWordNet-Pt [15] and Onto-pt[16]. For this work, WordNet-Pr 3.0 was chosen (developed for English) as sense repository. This choice was made because of the following reasons: WordNet-Pr is the most used sense repository in the literature; WordNet-Pr is considered a linguistic ontology, thus, it includes concepts and words written in English; and some sense repositories for Portuguese are under development or have a lower coverage than WordNet-Pr.

Another issue to consider is the choice of the WSD methods, because of the need of general purpose WSD methods and its integration with other applications. We chose four Knowledge-based WSD methods, each one from a different approach: using word overlapping [12], web search [17], graphs and similarity measures[18], and, finally, a method that is used in multi-document scenario[9].

After the selection of the WSD methods, we had to adapt these methods (some of these developed for English, initially)for Portuguese, because synsets indexed in WordNet-Pr are written in English. The way to adapt these is the same as used in[9] and it is described as follows: to obtain all synsets for a Portuguese word, we, first obtain all English translations from a bilingual dictionary (in our case, WordReference®), and, then, we obtain all synsetsfor every English translation, using WordNet-Pr. In Figure 1, we can see how this was performed with the verb "*informar*":

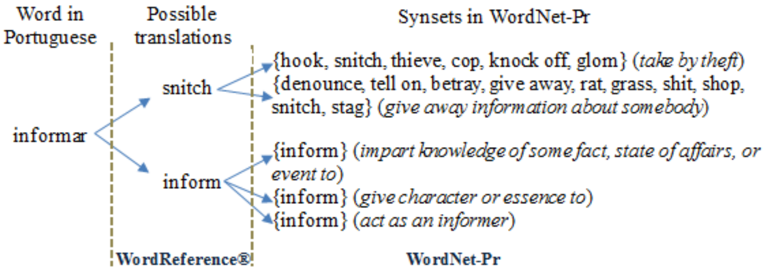
Word in Portuguese
Possible translations
Synsets in WordNet-Pr
informar
snitch
inform
{hook, snitch, thieve, cop, knock off, glom} (take by theft)
{denounce, tell on, betray, give away, rat, grass, shit, shop, snitch, stag} (give away information about somebody)
{inform} (impart knowledge of some fact, state of affairs, or event to)
{inform} (give character or essence to)
{inform} (act as an informer)
WordReference®
WordNet-Pr

Figure 1. *Method to obtain synsets \for a Brazilian Portuguese verb*

Besides getting synsets, all methods used the following pre-processing steps: (1) sentence splitting; (2) part-of-speech tagging with MXPOST tagger[19]; (3) removal of stopwords; (4) lemmatization of the content words; and (5) target words detection and context representation.

The following subsections describe the WSD Methods investigated in this work.

3.2. *Baseline methods*

In this work, we use 2 methods to compare with the implemented WSD methods. The first of these uses the most frequent sense (MFS) to determine the correct sense of a word. The MFS method uses a sense repository in which the indexed senses for a word are sorted by frequency and, then, it chooses the first sense.

For this work, the way it was adapted is described below: firstly, the MFS method chooses the first translation shown by WordReference® for a Brazilian verb (this is because the results shown by WordReference® are sorted by frequency), and, then, it chooses the first synset in the synset list shown by WordNet-Pr for the selected translation (this is because the results shown by WordNet-Pr are also sorted by frequency).

The second is a random, and blind, method that consists in, firstly, choosing a random translation for a Brazilian verb from the bilingual dictionary and, then, choosing a random synset from the synset list shown by WordNet-Pr for the selected translation.

3.3. *Word overlapping*

The most representative method from this approach is proposed in [12] (called Lesk for practical purposes). This method selects the sense of a word that has more common words with the words in its context window. For this approach, the configurations proposed in [9] were used. This method has six variations: (G-T) using synset glosses of the target word (word to be disambiguated) to compare with labels composed of possible word translations in the context; (S-T) using synset sample sentences of the target word to compare with labels composed of possible word translations in the context; (GS-T) using synset glosses and sample sentences of the target word to compare with labels composed of possible word translations in the context; (S-S) using only synset sample sentence of the target word to compare with labels composed of the sample sentences of all possible synsets for the context words; (G-G) using only synset glosses of the target word to compare with labels composed of the glosses of all possible synsets for the context words; and (GS2) using synset sample sentences and synset glosses of the target word to compare with labels composed of all possible synset sample sentences and glosses for the context words.

Besides these variations, we add other variations by modifying the length and the balance of the context window, this was done because literature says that verbs need unbalanced context windows, having a longer right side in the context window [20]. We use three window variations: (2-2) two words in the left and two words in the right; (1-2) one word in the left and two words in the right; (1-3) one word in the left and three words in the right; and (2-3) two words in the left and three words in the right.'

3.4. *Web search*

The Web Search-based method is the one proposed in [17] (called Mihalcea for practical purposes). This method constructs word pairs in order to disambiguate a word in the context of other word. This method works as follows: for a target word, the nearest random content word is used as context; then, the method obtains all synsets of the target word; then, queries are constructed, using the combination of every synset with the context, and posted on web search; finally, the synset included in the query with the best result is selected as sense for the target word.

For our case, a word pair consists of the verb under focus and the nearest noun in the sentence. Then, the results for every word pair combination are obtained from web, and, finally, the synset included in the word pair with the best result is selected. For this method, Microsoft Bing® was used for searching the web.

3.5. *Graphs*

The Graph-based WSD method is the one proposed in [18] (called Agirre Soroa for practical purposes). The authors in this work proposed 3 variations based on graphs that use PageRank algorithm [21] to rank the synsets. The first method creates a semantic graph with the synsets of all content words included in a sentence and then executes the PageRank algorithm over the generated graph to rank the synsets. Then, the method selects the highest scored synset for every content word. The second method uses the full WordNet graph and executes the PageRank algorithm over this. In this second method, PageRank algorithm is modified to give priority to synsets of all content words. The third method is similar to the second, but the difference is that this method gives priority to synsets of the context words, excepting the target word (and its synsets) and disambiguates one content word by execution (instead of the other methods that disambiguate all content words by execution). This has the assumption that the synset of the target word must be influenced by the synsets of the words around it. For our study, the last method is used because our focus is to disambiguate verbs only.

3.6. *Multi-document scenario*

The last WSD method is the proposed in [9] (called Nóbrega, for practical purposes). This method is used in multi-document scenario. This method uses a multi-document representation of context and assumes that all the occurrences of a word in a collection of texts have only one sense based in the corpus (one-sense per discourse heuristic). The multi-document representation of context for a word is built getting the “n” (in our case, we used 3 and 5 words) words that most co-occur with the word to disambiguate(target word) in a window of size “n”(assuming these words are the most related to the target word and help selecting relevant context words and the best synset). After the construction of the context window (obtained from the multi-document representation), the Lesk method is used to disambiguate the target word.

4. EVALUATION

4.1. *The corpus*

The CSTNews corpus[2] [22] [23] was used for evaluating the investigated WSD methods. This is a multi-document corpus composed of 140 texts, extracted from Brazilian news agencies, grouped in 50 collections, where texts of the same collection are about the same topic.

This corpus has sense-annotation for nouns [9] and verbs [24] using the WordNet-Pr as sense repository. In general, 5082 verb instances were annotated. These 5082 instances of verbs represent 844 different verbs with 1047 annotated synsets.

In agreement evaluation, the authors used the Kappa measure [25] and percent agreement among annotators. For percent agreement, the authors calculated the total agreement (when all annotators agreed for verb), partial agreement (when half of the annotators agreed, at least) and no-agreement. Due to the use of sense-repository for English and the use of a bilingual dictionary, the percent agreement was measured for translations, synsets and translation-synset pairs. In Table 1, we present the results of agreement evaluation.

According to the literature, the obtained Kappa values are considered moderate. We may see that the translation agreement shows the highest of the three evaluated items (Translation,

[2] Available at: http://www.icmc.usp.br/~taspardo/sucinto/cstnews.html

Synset and Translation-Synset). This happens because the selection of translations is a simpler task than synset selection. Analyzing the percent agreement, no-agreement is very low, the total agreement is higher and the partial agreement is the highest of the three. One of the reasons for this is that some verbs have a lot of senses and most of them look almost the same. Other reasons are the difficult for identifying participle verbs, complex predicates and the selection of a different English translation to annotate a verb.

Table 1. *Agreement measures computed in [24]*

	Kappa	Total (%)	Partial(%)	No-Agreement(%)
Translation	0.648	48.81	48.50	2.69
Synset	0.509	35.12	58.47	6.41
Translation-synset	0.474	31.73	61.29	6.98

4.2. *Comparison of WSD methods*

For evaluation, WSD methods (described from subsection 3.2 to 3.6) were tested in the CSTNews corpus. Two experiments were performed: the first experiment was to disambiguate all verbs included in the corpus, using all proposed WSD methods (all-words task); and the second experiment was to disambiguate 20 polysemic verbs in the corpus (Lexical sample task).

The measures used to evaluate all WSD methods were: Precision (P), which is the number of correctly classified verbs over the number of verbs classified by the method; Recall (R), number of correctly classified verbs over all verbs in the corpus (3); Coverage (C), number of classified verbs over all the verbs in the corpus; and (4) Accuracy (A), the same as (R), but using MFS method when no classification is found.

Results of All-words experiment are shown in Table 2. As one may see, no WSD method outperformed the MFS method, but all methods out-performed the Random method. The best method was the Nóbrega method, using three words as context and the S-T Lesk variation. The reason for this was the little verb sense variation in a collection of texts. We tested all variations of Lesk method (mentioned in Subsection 3.3) and the best configuration was using an unbalanced window (one word left and two words right) and the S-T variation. This confirms what the literature says: unbalanced windows help to Verb Sense Disambiguation. Mihalcea method got the worst results. One of the reasons for this is, as mentioned in [29], these senses of a verb change a lot in the presence of different nouns. Other reason is the lack of translations for its noun pair, so, it limits the quantity of verbs to disambiguate, and consequently, its coverage. AgirreSoroa method showed a reasonable result, in comparison with the other developed methods.

Table 2**.** *All-words experiment in CSTNews corpus*

	P (%)	R (%)	C (%)	A (%)
MFS	**49.91**	**47.01**	**94.20**	**-**
Random	10.04	9.46	94.20	-
Lesk-Verbs	40.10	37.69	93.98	37.77
Mihalcea	17.21	14.43	83.87	19.44
AgirreSoroa	28.45	26.80	94.20	26.80
Nóbrega-Verbs	**40.33**	**37.97**	**94.14**	**38.00**

Results of Lexical sample experiment are shown in Table 3.In this experiment, twenty random polysemic verbs (two or more senses in the corpus) were chosen and only the precision measure was computed in order to evaluate the performance of the methods (only the bests by approach). The numbers in bold indicate cases in that the methods performed as well as or better than the MFS method. In general, all WSD methods outperformed the random method. It is obvious, because the random method does not follow some heuristic to select a sense. Analyzing the other methods, it can be seen that Nóbrega method was the best method (P: 35.24%) but this did not outperform the MFS method (P: 36.97%).

One reason for this is the little variation of synsets for a sample word in a collection, i.e., some verbs were annotated in a collection using few synsets (see “F” column and “S” column in Table 3). Another reason is that, despite some verbs have high frequency, these have been annotated mostly with the same sense in a collection. This helps Nóbrega method, because, by using a window context based on the words that more co-occur in a multi-document scenario, it has a more consistent context and it is able to get a better result. Thus, if the Nóbrega method selects the majority sense for a verb, all verb instances will have the same sense, producing a high precision. The other methods (Lesk variation, Mihalcea and AgirreSoroa) presented results according to the All-words experiment, and the best of the three was Lesk variation, and the worst was Mihalcea.

4.3. *Comparison between morphosyntactic classes*

In Table 4, results of All-words experiment for nouns obtained in [9] are presented. Comparing the results of All-words experiment obtained for verbs (shown in Table 2) and nouns (shown in Table 4), it can be noted that verb is more difficult to disambiguate than noun. In the case of the Lesk method, noun senses disambiguation showed the best performance when this used balanced window (two words for the left and the right side). Unlike nouns, Verb Sense Disambiguation results were obtained when this used unbalanced window (one word for the left side and two words for the right side). Analyzing the content to compare, when this method used the content of the synset glosses, the noun sense disambiguation showed better results (Lesk), but when it used the content of the synset samples, the verb sense disambiguation showed better results. In the case of the Mihalcea method, the difference between verbs and nouns was greater. This occurred because noun senses are more stable in the presence of different verbs, unlike verb senses, which are less stable in the presence of different nouns. In the case of the Nóbrega methods, the best configuration for nouns (using the G-T variation) showed better performance than the best for verbs. This confirms that nouns have less meaning variation in the corpus than verbs [26].

5. FINAL REMARKS

In this work, the first exploratory study of classic WSD methods adapted for verbs in Brazilian Portuguese was presented. Due to the need of WSD methods that can be used in different contexts, knowledge-based WSD methods were chosen. The approaches for knowledge-based WSD methods were: word overlapping, web search, graphs and a method focused on multi-document scenario. Then, we used a journalistic corpus, which included various domains to guarantee the general use, to test these methods.

Table 3. *Lexical sample experiment in CSTNews corpus (F: Frequency; S: Number of synsets; MFS: Most frequent sense Method; R: Random Method; L: Lesk method; M: Mihalcea method; AS: AgirreSoroa method; N: Nóbrega method)*

Word	**F**	**S**	**MFS**	**R**	**L**	**M**	**AS**	**N**
Estragar (ruin)	2	2	0.00	0.00	0.00	0.00	0.00	0.00
Olhar (look)	2	2	0.00	0.00	0.00	0.00	0.00	0.00
Perceber (perceive)	2	2	0.00	0.00	0.00	0.00	0.00	0.00
Gostar (like)	2	2	0.00	100.00	0.00	0.00	0.00	0.00
Exibir (exhibit)	3	2	0.00	50.00	0.00	0.00	0.00	0.00
Resultar (result)	3	3	100.00	0.00	0.00	0.00	0.00	**100.00**
Pertencer (belong)	4	2	100.00	33.33	66.67	0.00	0.00	**100.00**
Voar (fly)	4	2	33.33	0.00	**33.33**	0.00	**33.33**	**33.33**
Entender (understand)	4	3	50.00	0.00	**50.00**	**50.00**	**50.00**	**50.00**
Descobrir (discover)	4	3	50.00	0.00	**50.00**	0.00	**50.00**	**50.00**
Destacar (feature)	5	2	0.00	0.00	0.00	0.00	0.00	0.00
Achar(find)	6	3	0.00	0.00	0.00	0.00	0.00	0.00
Recuperar (recover)	6	4	50.00	25.00	**50.00**	25.00	25.00	**50.00**
Retirar (withdraw)	9	3	100.00	0.00	**100.00**	16.67	0.00	**100.00**
Comandar (command)	12	4	33.33	22.22	**33.33**	28.57	22.22	**33.33**
Marcar (mark)	17	7	0.00	0.00	**20.00**	0.00	**20.00**	0.00
Entrar (enter)	21	4	62.50	0.00	**62.50**	28.57	0.00	50.00
Receber (receive)	36	9	77.78	0.00	50.00	14.29	0.00	55.56
Deixar (leave)	49	16	11.11	0.00	7.41	0.00	**11.11**	**11.11**
Informar (inform)	55	2	71.43	10.71	64.29	0.00	**71.43**	**71.43**
AvgPrecision	-	-	**36.97**	12.06	29.38	8.15	14.15	35.24

Two experiments were performed: All-words and Lexical sample task. Both All-words and Lexical sample tasks showed that no method outperformed the MFS method. However, considering the implemented methods, the Nóbrega method got the best results. One reason for this is the little variation of verb senses for the sample in a collection of texts. A third experiment was performed, aiming at comparing the performance between morphosyntactic classes (nouns and verbs). The results are consistent with what other studies claim that verbs are more difficult to disambiguate.

In spite of the fact that Wordnet-Pr is a wide resource used in WSD, the tested methods showed some problems with lexical gaps. For instance, the verb "*pedalar*" (action of doing a specific dribble) in the sentence "*O Robinhopedalou*" has not a respective synset in WordNet-Pr. To resolve this problem, a generalization of the Portuguese verbis necessary, using the verb "dribble" ("*driblar*", in Portuguese).

As we could see, there is still room for improvements in WSD for verbs. A future work is the use of repositories focused on verbs (and developed for Portuguese), which contain syntactic and semantic information that might be added to the methods to improve their performance.

Table 4. *All-words experiment for nouns*

Method	P (%)	R (%)	C (%)	A (%)
MFS	51.00	51.00	100.00	-
Lesk-Noun	42.20	41.20	91.10	41.20
Mihalcea	39.71	39.47	99.41	39.59
Nóbrega-Noun	49.56	43.90	88.59	43.90

REFERENCES

1. Jurafsky, D. & Martin, J. H. 2009. Speech and language processing: An introduction to natural language processing. *Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall.
2. Agirre, E. & Edmonds, P. 2006. Introduction. *Word Sense Disambiguation: Algorithms and Applications* (pp. 1-28). Springer.
3. Navigli, R. 2009. Word sense disambiguation: A survey. In *ACM Computational Survey*, 41, 1-69.
4. Gao, N., Zuo, W., Dai, Y. & Wei, L. 2014. Word sense disambiguation using WordNet semantic knowledge. In proceedings of the *Eighth International Conference on Intelligent Systems and Knowledge Engineering*, (pp. 147-156), Springer Berlin Heidelberg, China.

5. Fillmore, C. J. 1968. The case for case. In *Universals in Linguistic Theory* (pp. 1-89). New York.
6. Che, W. & Liu, T. 2010. Jointly modeling wsd and srl with markov logic. In proceedings of *23rd International Conference on Computational Linguistics*, (pp. 161-169), Association for Computational Linguistics, China.
7. Specia, L. 2007. Uma Abordagem Híbrida Relacional para a Desambiguação Lexical de Sentido na Tradução Automática. PhD thesis, Instituto de Ciências Matemáticase de Computação- ICMC-USP.Brazil.
8. Machado, I. M., de Alencar, R.O., de Oliveira Campos Junior, R. & Davis, C. A. 2011. An ontological gazetteer and its application for place name disambiguation in text. In *Journal of the Brazilian Computer Society*, 17, 267-279.
9. Nóbrega, F. A. A. & Pardo, T. A. S. 2014. General purpose word sense disambiguation methods for nouns in Portuguese. In proceedings of the *11th International Conference on Computational Processing of Portuguese* (pp. 94-101), Brazil.
10. Travanca, T. 2013. Verb sense disambiguation. MSc thesis. *Instituto Superior Técnico*. Universidade Técnica de Lisboa. Portugal.
11. Fellbaum, C. 1998. *WordNet, An Eletronic Lexical Database*. MIT Press.
12. Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In proceedings of *5th Annual International Conference on Systems Documentation* (pp. 24-26), Association for Computing Machinery, USA.
13. Baptista, J. 2012. ViPEr: A lexicon-grammar of European Portuguese verbs. In proceedings of the *31st International Conference on Lexis and Grammar* (pp. 10-16), Czech Republic.
14. Dias da Silva, B. C. 2005. A construção da base da wordnet.br : Conquistas e desafios. In proceedings of *XXV Congresso da Sociedade Brasileira de Computação*.
15. Paiva, V., Rademaker, A. & Melo, G. 2012. OpenWordNet-PT: An open Brazilian Wordnet for reasoning. In proceedings of *COLING 2012: Demonstration Papers* (pp. 353-360), India.
16. Gonçalo Oliveira, H., Antón, L. P. & Gomes, P. 2012. Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. In *Natural Language Processing and Information Systems, Proceedings of 17th International Conference on*

Applications of Natural Language to Information Systems (pp. 210-215), The Netherlands,

17. Mihalcea, R. & Moldovan, D. I. 1999. A method for word sense disambiguation of unrestricted text. In proceedings of *37th Annual Meeting of the Association for Computational Linguistics* (pp. 152-158), USA.
18. Agirre, E. & Soroa, A. 2009. Personalizing pagerank for word sense disambiguation. In proceedings of the *12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 33-41), *Association for Computational Linguistics*.
19. Ratnaparkhi, A. 1996. A maximum entropy model for part-of-speechtagging. In proceedings of the *First Empirical Methods in NLP Conference* (pp. 133-142), Association for Computational Linguistics.
20. Vasilescu, F., Langlais, P. & Lapalme, G. 2004. Evaluating variants of the lesk approach for disambiguating words. In proceedings of *Language Resources and Evaluation (LREC 2004)* (pp. 633-636), Portugal.
21. Brin, S. & Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. In proceedings of *17th International World-Wide Web Conference* (WWW 1998).
22. Aleixo, P. & Pardo, T. A. S. 2008. CSTNews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-documentstructuretheory). Relatório Técnico 326, Instituto de Ciências Matemáticas e de Computação., Universidade de São Paulo. Brazil.
23. Cardoso, P. C. F., Maziero, E. G., Jorge, M. L. R. C., Seno, E. M. R., Felippo, A. D., Rino, L. H. M., Das Graças, M. V. N. & Pardo, T. A. S. 2008. CSTNews – a discourse-annotated corpus for single and multi-document summarization of News texts in Brazilian portuguese. In *proceedings of III Workshop "A RST e os Estudos do Texto,"* (pp. 88-105), Sociedade Brasileira de Computação, Brazil.
24. Sobrevilla Cabezudo, M. A., Maziero, E.G., Souza, J. W. C., Dias, M. S., Cardoso, P. C. F., Balage Filho, P. P., Agostini, V., Nóbrega, F. A. A., Barros, C. D., Di Felippo, A. & Pardo, T. A. S. 2014. Anotação de Sentidos de Verbos em Notícias Jornalísticas em Português do Brasil. In proceedings of the *XII Encontro de Linguística de Corpus*-ELC, Brazil.
25. Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22, 249-254.

26. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J. 1990. Introduction to Wordnet: An on-line lexical database. In *International Journal of Lexicography*, 3, 235-244.

ACKNOWLEDGMENT

Part of the results presented in this paper were obtained through research on a project titled "Semantic Processing of Texts in Brazilian Portuguese", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No.8.248/91.

MARCO ANTONIO SOBREVILLA CABEZUDO
NÚCLEO INTERINSTITUCIONAL
DE LINGUÍSTICA COMPUTACIONAL,
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO,
UNIVERSIDADE DE SÃO PAULO
AV. TRABALHADOR SÃO-CARLENSE,
400 - CENTRO CEP: 13566-590 –
SÃO CARLOS – SÃO PAULO, BRASIL.
E-MAIL: <MARCOSBC@ICMC.USP.BR>

THIAGO ALEXANDRE SALGUEIRO PARDO
NÚCLEO INTERINSTITUCIONAL
DE LINGUÍSTICA COMPUTACIONAL,
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO,
UNIVERSIDADE DE SÃO PAULO,
AV. TRABALHADOR SÃO-CARLENSE,
400 - CENTRO CEP: 13566-590 –
SÃO CARLOS – SÃO PAULO, BRASIL.
E-MAIL: <TASPARDO@ICMC.USP.BR>