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

Marco Antonio Sobrevilla Cabezudo, Thiago Alexandre Salgueiro Pardo

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
{marcosbc, taspardo}@icmc.usp.br

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, Sentiment Analysis or Information Retrieval. For English, WSD has been widely studied, obtaining different performances. Verbs are more important morphosyntactic class and help in sentence construction, but, in an analysis of morphosyntactic classes, verb is the harder class to be disambiguated. For Portuguese, there are few studies about this and, only recently, focused on general purpose. In the present paper, we report an exploratory study of Word Sense Disambiguation methods based on knowledge 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 baseline is generally difficult to outperform, but there is still room for improvements in WSD for verbs.

Keywords: Word Sense Disambiguation, based on knowledge, Brazilian Portuguese

1 Introduction

Currently, the amount of information on internet has increased a lot. This is caused by the explosion of web 2.0, which produces a lot of information in unstructured format. Because of information increase and the need of more intelligent, and complex, processing and understanding of this information, deeper linguistic knowledge is necessary. Semantics is a high linguistic knowledge level [1] and it's a popular subject in the Natural Language Processing community. One of the more important problems for computing related to Semantics is the ambiguity and, specifically, lexical ambiguity. Lexical Ambiguity occurs when a word can express two or more senses in a determined context.

Lexical Ambiguity can be expressed in various difficulty levels. For example, consider the following four sentences:

– *O homem contou o número de pessoas que ficaram feridas* (“The man counted the number of people who were injured.”).

– *O jogador bateu na bola com força* (“The player kicked the ball strongly.”).
– *O lutador bateu as botas* (“The fighter died.” or “The fighter beat the boots.”).
– *O banco quebrou a semana passada* (“The seat broke last week.” or “the bank went under last week.”).

In the first example, the sense of verb “*contar*” can be identified easily (to determine the total number of a collection of items); in the second example, the sense of verb “*bater*” is easily identified too (to kick something), but in the third example, the sense of the same verb can be difficult to identify (to die), because it is necessary that we know the context (given by “*as botas*” in "*bater as botas*"). 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 verb “*quebrar*”, that could mean “break” or fail, go under.

Word Sense Disambiguation (WSD) aims at identifying the correct sense of a word within a given context using a pre-specified sense repository [2]. WSD is considered an important sub-task of other tasks, as Semantic Role Labeling [3]; and applications, as Information Retrieval [4], Machine Translation [5] and Sentiment Analysis [6].

For English, there is a lot of studies about WSD, using different approaches and techniques [7]. Recently, knowledge-based WSD methods have become very popular [8]. It is due to the increase of the need of WSD methods for general purpose and the increase of the coverage of sense repositories. Although these, studies have shown that WSD is a hard task to be resolved and the performance obtained is not high. Analyzing by morphosyntactic class, WSD methods show different performances, being the verb the worst class. One of the reasons for this is that verbs need more syntactic and semantic information to get better results. Verb is one of the most important morphosyntactic classes. This class is useful for sentence construction.

For Portuguese, there are few studies and some of these are domain-oriented [9] [10], and this can negatively 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 [11] (focused on nouns) and [12] (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 Brazilian Portuguese, using WordNet-Pr [13] as sense repository and WordReference®[1] as bilingual dictionary; and a comparison with the results obtained for nouns [11].

The results show that baseline is generally difficult to outperform, but there is still room for improvements in WSD for verbs.

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 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 to future work in Section 5.

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

2 Concepts and Related Works

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 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; execute the automatic disambiguation, and show the correct sense for the target word as output [1].

WSD task can 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 in knowledge-based, corpus-based and hybrid methods [2].

Knowledge-based methods use linguistic resources and some 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 this 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. The reason for this is due to the size of corpus would have to be so broad to create an enough classifier that covers all words. Although this approach covers a limited sample of words, this has a better performance than Knowledge-based methods.

Finally, hybrid methods use techniques from knowledge-based and corpus-based methods.

For Portuguese, there are few studies and some of these are domain-oriented. This can 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 [9], a WSD method based on Inductive Logic Programming for Automatic 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 shows that the proposed method outperformed the most frequent translation method and other methods based on machine learning.

In [10], a geographical disambiguation method, specifically, for disambiguating place names is presented. This method used an ontology created in [10], called OntoGazetter, as knowledge base. This ontology is composed by place concepts. The results showed that OntoGazetter contributes to geographical disambiguation positively.

The first research about general purpose WSD methods in Brazilian Portuguese is presented in [14]. The author researched Knowledge-based WSD methods for common nouns, using WordNet-Pr [13] as sense repository and WordReference® as bilingual dictionary (since the language was Portuguese). In this work, besides the research of WSD methods, the author proposed a WSD method based on co-occurrence graphs and a variation of Lesk algorithm for multi-document scenario [15]. 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 [12], two verb sense disambiguation methods for European Portuguese are developed, using ViPer [16] 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 it was difficult to outperforming, 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 [17], OpenWordNet-Pt [18] and Onto-pt [19]. 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 as a linguistic ontology, thus, it includes all 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, everyone from a different technique: using word overlapping [15], web search [20], graphs and similarity measures [21], and finally, a method that is used in multi-document scenario [11].

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 [11] and it is described as below: 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 synsets for every English translation, using WordNet-Pr. In Figure 1, we can see how this was performed with the verb "*informar*":

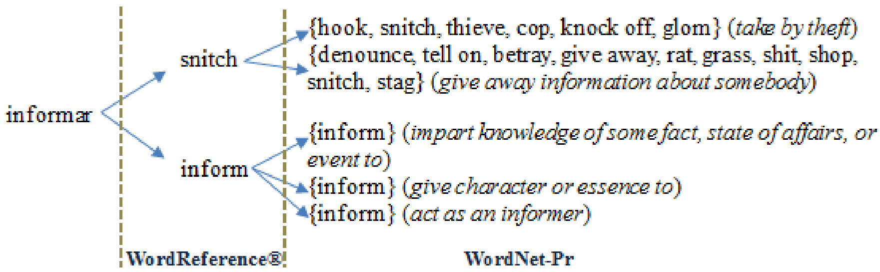

Fig. 1. Method to obtain synsets for a Brazilian Portuguese verb

Besides getting synsets, all methods used the following pre-processing steps: (1) sentence splitting; (2) part-of-speech tagging with MXPOST tagger [22]; (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 developed 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. 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 as it was adapted is described below: first, 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, finally, 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 sorted by frequency).

The second is a random, and blind, method that consists in, first, choosing a random translation for a Brazilian verb from bilingual dictionary and, finally, 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 [15] (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 [11] were used. This method has six variations: (G-T) using synset glosses of 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 target word to compare with labels composed of possible word translations in the context; (GS-T) using synset glosses and sample sentences of target word to compare with labels composed of possible word translations in the context; (S-S) using only synset sample sentence of 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 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 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 modifying the length and the balance of the context window; it was done due to the literature says that verbs need unbalanced context windows, having the right side in the context window longer than left side [23]. We use three window variations: (2-2) two words for left and two words for right; (1-2) one word for left and two words for right; and (1-3) one word for left and three words for right.

3.4 Web Search

The Web Search-based method is the one proposed in [20] (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 below: For a target word, the nearest 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 results 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 best results 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 [21] (called AgirreSoroa for practical purposes). The authors in this work propose 3 methods based on graphs that use PageRank algorithm [24] to rank the synsets. The first method creates a semantic graph with all words included in context and then executes PageRank algorithm over the generated graph to rank the synsets. Finally, the method selects the highest scoring synsets. The second method uses the full WordNet graph and executes the PageRank algorithm over it. In this method, PageRank algorithm is modified to give priority to synsets of all context words. The third method is the same as the second, but the difference is that this method gives priority to synsets of the context words, excepting the target word (and its synsets). This has the assumption that the synset of the target word must be influenced by the synsets of the words around it. For this study, the last method is used because the results in the work show that this method has good results for verbs.

3.6 Multi-document scenario

The last WSD method is the proposed in [11] (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 our corpus evidence (one-sense per discourse heuristic). The multi-document representation of context for a word is built getting the “n” 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 (obtaining from the multi-document representation), Lesk method is used to disambiguate the target word.

4 Evaluation

4.1 The Corpus

The CSTNews[2][25] [26] corpus was used for evaluating the proposed WSD methods. This is a multi-document corpus composed of 140 texts, extracted from Brazilian news agencies, grouped in 50 collections, where texts of a same collection are about the same topic.

This corpus has sense-annotation for nouns [14] and verbs [27] 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 [28] and percent agreement between annotators. For percent agreement, the authors calculated the total agreement (when all annotators agreed for an instance 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 obtained results, the Kappa values are considered moderate. Analyzing the percent agreement, we can see that partial agreement is the major of the three. One of the reasons for this is that some verbs have a lot of senses and most of them are almost equals. 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 [26]

	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 polysemous verbs in the corpus (Lexical sample task).

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

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 it can see, no WSD method outperforms the MFS method, but all methods outperform Random method. The best method was Lesk method. We tested all variations of Lesk method (mentioned in Subsection 3.3) and the best configuration was using a balanced window (two words left and two words right) and the S-T variation (Lesk-S-T- 2-2 in the Table 2). In Table 2, only the 2 best results for Lesk method variations are shown. Mihalcea method gets 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 has a reasonable result, in comparison with the other developed methods. For Nóbrega methods, the best configuration was using three words in the context and the S-T Lesk variation (Nóbrega-S-T-3 in Table 2). Although the performance of this configuration was not bad (compared to other methods), it was not better than the best Lesk variation.

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-G-T-2-2	37.34	35.15	94.14	35.15
Lesk-S-T-2-2	**38.56**	**36.30**	**94.14**	**36.30**
Mihalcea	17.21	14.43	83.87	19.44
AgirreSoroa	30.13	28.38	94.20	28.38
Nóbrega-G-T-3	32.19	30.30	94.14	30.34
Nóbrega-S-T-3	35.09	33.04	94.14	33.08
Nóbrega-G-T-5	29.06	27.38	94.20	27.38
Nóbrega-S-T-5	32.53	30.64	94.20	30.64

Results of Lexical sample experiment are shown in Table 3. In this experiment, twenty random polysemous verbs (two or more senses indexed in the corpus) were chosen and only the precision measure was computed to evaluate the performance of the methods. The numbers in bold indicate cases 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-

S-T-3 method was the best method (P: 41.07%) and outperformed 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. 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 which 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 [11] 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 senses are more difficult to disambiguate than noun senses. In the case of the Lesk method, verb and noun disambiguation have the best performance when they use balanced window (two words for the left and the right side). Analyzing the content to compare, it can be see that when it uses the content of the synset glosses, the noun sense disambiguation shows better results (Lesk-G-T-2-2), but when it uses the content of the synset samples, the verb sense disambiguation show better results (Lesk-S-T-2-2). In case of the Mihalcea method, the difference between verbs and nouns is greater. This occurs 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 (Nóbrega-3 G-T) has better performance than the best for verbs (Nóbrega-3 S-T). This affirms that nouns have less variation of meanings in the corpus than verbs [29].

5 Final Remarks

In this work, the first exploratory study of WSD classic 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 use a journalistic corpus, which included various domains to guarantee the general use, to test these methods.

Two experiments were performed: the first one, All-words task, showed that no method outperformed the MFS method. The second one, Lexical sample task, presented that Nóbrega method outperformed the MFS method. One reason to 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). The results are consistent with what other studies claim that verbs have greater difficulty in disambiguating the sense.

Table 3. Lexical sample experiment in CSTNews corpus

Word	Freq	Synsets	MFS	Random	Lesk S-T 2-2	Mihalcea	AgirreSoroa	Nóbrega-S-T-3
estragar	2	2	0.00	0.00	0.00	0.00	0.00	0.00
olhar	2	2	0.00	0.00	0.00	0.00	0.00	0.00
perceber	2	2	0.00	0.00	0.00	0.00	0.00	0.00
gostar	2	2	0.00	100.00	0.00	0.00	0.00	**100.00**
Exibir	3	2	0.00	50.00	0.00	0.00	0.00	0.00
resultar	3	3	100.00	0.00	0.00	0.00	0.00	**100.00**
pertencer	4	2	100.00	33.33	66.67	0.00	0.00	**100.00**
voar	4	2	33.33	0.00	**33.33**	0.00	**33.33**	**33.33**
entender	4	3	50.00	0.00	**50.00**	**50.00**	**50.00**	**50.00**
descobrir	4	3	50.00	0.00	**50.00**	0.00	**50.00**	0.00
destacar	5	2	0.00	0.00	0.00	0.00	0.00	0.00
achar	6	3	0.00	0.00	0.00	0.00	0.00	**66.67**
recuperar	6	4	50.00	25.00	**50.00**	25.00	25.00	**50.00**
retirar	9	3	100.00	0.00	83.33	16.67	0.00	**100.00**
comandar	12	4	33.33	22.22	**33.33**	28.57	22.22	**33.33**
marcar	17	7	0.00	0.00	**20.00**	0.00	**20.00**	0.00
entrar	21	4	62.50	0.00	**62.50**	28.57	0.00	50.00
receber	36	9	77.78	0.00	44.44	14.29	0.00	55.56
deixar	49	16	11.11	0.00	7.41	0.00	**11.11**	**11.11**
informar	55	2	71.43	10.71	60.71	0.00	**71.43**	**71.43**
Avg Precision	-	-	36.97	12.06	28.09	8.15	14.15	**41.07**

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

Table 4. All-words experiment for nouns

Method	P (%)	R (%)	C (%)	A (%)
MFS	51.00	51.00	100.00	-
Lesk-2-2-G-T	42.20	41.20	91.10	41.20
Mihalcea	39.71	39.47	99.41	39.59
Nóbrega-3-G-T	49.56	43.90	88.59	43.90
Nóbrega-5-G-T	46.87	41.80	87.65	41.80

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 for adding this information to the developed methods to improve their performance.

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.this work.

References

1. Jurafsky, D., Martin, J. H.: Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Prentice-Hall. (2009).
2. Agirre, E., Edmonds, P.: Introduction. In: Word Sense Disambiguation: Algorithms and Applications. pp 1-28. Springer. (2006).
3. Che, W., Liu, T.: Jointly modeling wsd and srl with markov logic. In: Proceedings of 23rd International Conference on Computational Linguistics. China. pp. 161-169. Association for Computational Linguistics. (2010).
4. Zhong, Z., Ng, H. T.: Word Sense Disambiguation Improves Information Retrieval. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Republic of Korea. pp. 273-282. Association for Computational Linguistics. (2012).
5. Chan, S. Y., Ng, T. H., Chiang, D.: Word Sense Disambiguation Improves Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Czech Republic. pp. 33-40. Association for Computational Linguistics. (2007).
6. Akkaya, C., Wiebe, J., Mihalcea, R.: Subjectivity Word Sense Disambiguation. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 190-199. Association for Computational Linguistics. (2009).
7. Navigli, R.: Word sense disambiguation: A survey. In: ACM Computational. Survey. Vol 41. pp. 1-69. ACM. (2009).
8. Gao, N., Zuo, W., Dai, Y., Lv, Wei.: Word Sense Disambiguation Using WordNet Semantic Knowledge. In: Proceedings of the Eighth International Conference on Intelligent Sys-

tems and Knowledge Engineering. China. pp. 147-156. Springer Berlin Heidelberg. (2014).

9. Specia, L.: Uma Abordagem Híbrida Relacional para a Desambiguação Lexical de Sentido na Tradução Automática. PhD thesis, Instituto de Ciências Matemáticas e de Computação-ICMC-USP. Brazil. (2007).
10. Machado, I.M., de Alencar, R.O., de Oliveira Campos Junior, R., Davis, C.A.: An ontological gazetteer and its application for place name disambiguation in text. In: Journal of the Brazilian Computer Society, vol. 17. pp. 267-279. (2011).
11. Nóbrega, F.A.A., Pardo, T.A.S.: General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. In: Proceedings of the PROPOR 2014 PhD and MSc/MA Dissertation Contest / 11st International Conference on Computational Processing of Portuguese. Brazil. pp. 94-101. (2014).
12. Travanca, T.: Verb Sense Disambiguation. MSc thesis. Instituto Superior Técnico. Universidade Técnica de Lisboa. Portugal. (2013).
13. Fellbaum, C.: WordNet An Eletronic Lexical Database. MIT Press. (1998).
14. Nóbrega, F. A. A.: Desambiguação Lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento. Msc thesis. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Brazil. (2013).
15. Lesk, M.: 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. USA. pp. 24-26. Association for Computing Machinery. (1986).
16. Baptista, J.: ViPEr: A Lexicon-Grammar of European Portuguese Verbs. In: Proceedings of the 31st International Conference on Lexis and Grammar. Czech Republic. pp 10-16. (2012).
17. Dias da Silva, B. C.: A construção da base da wordnet.br : Conquistas e desafios. In: Proceedings of XXV Congresso da Sociedade Brasileira de Computação. (2005).
18. Paiva, V., Rademaker, A., Melo, G.: OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In: Proceedings of COLING 2012: Demonstration Papers. India. pp. 353–360. The COLING 2012 Organizing Committee. (2012).
19. Gonçalo Oliveira, H., Antón, L. P., Gomes, P.: Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. In: Natural Language Processing and Information Systems, Proceedings of 17h International Conference on Applications of Natural Language to Information Systems. The Netherlands. pp. 210-215. Springer. (2012).
20. Mihalcea, R., Moldovan, D.I.: A method for word sense disambiguation of unrestricted text. In: Proceedings of 37th Annual Meeting of the Association for Computational Linguistics. USA. pp. 152-158. Association for Computational Linguistics. (1999).
21. Agirre, E., Soroa, A.: 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. (2009).
22. Ratnaparkhi, A.: A maximum entropy model for part-of-speech tagging. In: Proceedings of the First Empirical Methods in NLP Conference. pp 133-142. Association for Computational Linguistics. (1996).
23. Vasilescu, F., Langlais, P., Lapalme, G.: Evaluating variants of the Lesk approach for disambiguating words. In: Proceedings of Language Resources and Evaluation (LREC 2004). Portugal. pp. 633-636. (2004).
24. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of 17th International World-Wide Web Conference (WWW 1998). (1998).

25. Aleixo, P., Pardo, T. A. S.: CSTNews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory). Relatório Técnico 326, Instituto de Ciências Matemáticas e de Computação. Universidade de São Paulo. Brazil. (2008).
26. 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.: 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". Brazil. pp. 88-105. Sociedade Brasileira de Computação. (2008).
27. 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.: Anotação de Sentidos de Verbos em Notícias Jornalísticas em Português do Brasil. In: Proceedings of the XII Encontro de Linguística de Corpus - ELC. Brazil. (2014).
28. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics. Vol. 22. pp. 249-254. (1996).
29. Miller, G. A., Beckwith, R, Fellbaum, C., Gross, D., Miller, K. J.: Introduction to Wordnet: An on-line lexical database. In: International Journal of Lexicography. vol. 3. pp. 235-244. (1990).