Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 83-90 recibido 29-03-2017 revisado 16-05-2017 aceptado 29-05-2017

Exploring Classical and Linguistically Enriched Knowledge-based Methods for Sense Disambiguation of Verbs in Brazilian Portuguese News Texts

Exploración de Métodos basados en Conocimiento Clásicos y Lingüísticamente Enriquecidos para Desambiguación del Sentido de los Verbos en Textos de Noticias del Portugués Brasileño

Marco A. Sobrevilla Cabezudo, Thiago A. S. Pardo
Interinstitutional Center for Computational Linguistics (NILC)
University of São Paulo
São Carlos, SP, Brazil
msobrevillac@usp.br, taspardo@icmc.usp.br

Abstract: Word Sense Disambiguation (WSD) aims at determining the appropriate sense of a word in a given context. This task is challenging and highly relevant for the Natural Language Processing community. However, there are few works on Portuguese word sense disambiguation and some of these are domain oriented. In this paper, we report a study on general purpose WSD methods for verbs in Brazilian Portuguese. This study is divided into three steps: (1) the sense annotation of a corpus, (2) the exploration of classical WSD methods, and (3) the incorporation of linguistic knowledge to some of these classical methods. Among the contributions, we emphasize the free availability of the sense-annotated corpus and the use of a verb-focused repository to support classical methods in a new way.
Keywords: Word sense disambiguation, lexical semantics, verbnet.br

Resumen: La Desambiguación del Sentido de las Palabras (DSP) tiene como objetivo determinar el sentido más apropiado para una palabra en un contexto específico. Esta tarea es desafiante y altamente relevante para la comunidad de Procesamiento de Lenguaje Natural, mas existen pocos trabajos para el portugués y varios de ellos están orientados a dominios específicos. En este trabajo reportamos un nuevo estudio sobre métodos de DSP de propósito general para verbos en portugués brasileño. Este estudio se divide en tres etapas: (1) la anotación del sentido de verbos en un corpus, (2) la exploración de métodos clásicos de DSP, y (3) la incorporación de conocimiento lingüístico a algunos de estos métodos clásicos. Entre las contribuciones podemos enfatizar la libre disponibilidad del corpus anotado y el uso de un repositorio centrado en verbos para ayudar a métodos clásicos en una nueva forma.
Palabras clave: Desambiguación del sentido de las palabras, semántica léxica, verbnet.br

1 Introduction

Lexical Ambiguity (LA) is one of the most difficult problems to be solved in Semantics. It occurs when a word may express two or more senses in a determined context. For example, in the sentence "*O banco quebrou faz duas semanas*" (which could be "The bank failed two weeks ago" or "The seat fell apart two weeks ago", the verb "*quebrou*" might refer to the sense of "to fall apart" or "to fail". In this case, considering that we are talking about a financial institution, the most apropriate sense for the verb would be "to fail".

Word Sense Disambiguation (WSD) is the task that aims at identifying the correct sense of a word in its context of occurrence (Jurafsky and Martin, 2009). WSD is an important and useful task for several applications, as Sentiment Analysis, Machine Translation and Information Retrieval.

ISSN 1135-5948

WSD have been widely studied in English. Unfortunately, for Portuguese, there are few studies and most of these are focused on specific tasks, as machine translation (Specia, 2007) and geographical disambiguation (Machado et al., 2011). Only more recently, general purpose WSD methods have been studied for common nouns (Nóbrega and Pardo, 2014) and verbs (Travanca, 2013).

In this work, we investigate WSD methods for verbs in Brazilian Portuguese. Verbs are an important class and have a significant role in sentence structuring. One challenge in this research line is that verbs are the most difficult grammatical class to disambiguate, as some studies show (Mihalcea and Moldovan, 1999) (Agirre and Soroa, 2009). In general, verbs tend to be more polysemic than other grammartical classes. In this paper, we investigate general purpose WSD methods for verbs and the incorporation of linguistic knowledge in some methods, using a verb-focused repository, the VerbNet.Br (Scarton, 2013), which groups verbs into classes according to their syntactic and semantic behaviors, following Levin classes (Levin, 1993).

The adopted methodology in this work was composed by the following steps: (1) to sense annotate a corpus, (2) to explore some classical WSD methods, and (3) to incorporate linguistic knowledge to some of these classical methods. We evidence the difficulties of dealing with verbs and that incorporating linguistic knowledge may help.

This paper is organized in 5 sections. In Section 2, we present some related work. Section 3 shows the developed WSD methods and the incorporation of linguistic knowledge, while their evaluation is reported in Section 4. Finally, Section 5 presents some conclusions and future work.

2 Related Work

In this section, we briefly describe some previous WSD studies for Brazilian Portuguese.

The first one is a WSD method based on Inductive Logic Programming for the Machine Translation task (Specia, 2007). This method was focused on disambiguating ten highly polysemic English verbs to their respective Portuguese verbs. The author performed some experiments and showed that the proposed method outperformed the baseline method and other Machine Learning-based methods.

Another domain-oriented disambiguation method is presented in (Machado et al., 2011). The authors proposed a method to distinguish place names (geographical disambiguation) using an ontology as knowledge base, called OntoGazetter. This ontology contains place concepts. The results indicated that OntoGazetter positively contributes to geographical disambiguation.

The first research on general purpose WSD methods for Brazilian Portuguese is presented in (Nóbrega and Pardo, 2014). In this work, the authors focused on disambiguating nouns and explored some knowledge-based WSD methods. They used Princeton WordNet (Fellbaum, 1998) as sense repository and WordReference® as bilingual dictionary (before indexing the words to WordNet, it was necessary to translate them to English). Additionally, the authors developed a method using co-occurrence graphs, which proved useful in multi-document scenarios.

Another general purpose WSD method that focused on verbs for European Portuguese is presented in (Travanca, 2013). The author proposed two WSD methods, one using rules and other using machine learning. The sense repository was ViPer (Baptista, 2012), which contains syntactic and semantic information about verbs. The results showed that the baseline (the most frequent sense method) was difficult to be outperformed, but a combination of the methods got it

Finally, an exploratory study of several machine learning algorithms on an extension of the corpus analyzed in (Travanca, 2013) is presented in (Suissas, 2014). In this study, the author showed that the Naive Bayes algorithm outperformed the baseline (the most frequent sense method).

3 Methodology

In this work, following the previous approaches to WSD for Portuguese, we chose Princeton WordNet as sense repository. Three other reasons also motivated this: (1) this resource is widely used for WSD, (2) it is considered a linguistic ontology[1], and (3) some sense repositories for Portuguese are

[1]A linguistic ontology assumes that the concepts/senses are represented in a natural language - English, in this case.

still under development or have a lower coverage/accuracy.

In relation to the studied WSD methods, we selected only knowledge-based methods because they are more general purpose than other ones. We selected four methods, each one following a specific strategy: word overlapping, web search, graphs, and multi-document scenario.

In general, the studied WSD methods needed a previous step to get all possible synsets for each word (due to the multilingual nature of the task). This step consisted in: for each word, (1) to get all possible English translations using a bilingual dictionary, and (2) to retrieve all synsets for all translations. In this work, we used the online bilingual dictionary WordReference® to automatically get the translations. Additionally, all explored WSD method executed these other steps: (1) POS tagging (using MXPOST) (Aires et al., 2000), (2) stopword removal, (3) lemmatization of content words, and (4) retrieval of the context of the target word (the word to be disambiguated).

3.1 Sense Annotation of the Corpus

The CSTNews corpus[2] (Cardoso et al., 2011) was manually sense-annotated and used to test the WSD methods. This is a multi-document corpus composed of 140 news texts (in Brazilian Portuguese) grouped in 50 collections, where the texts in a collection are on the same topic.

This corpus has sense annotations for the most frequent nouns (Nóbrega and Pardo, 2014) and for all the verbs (Cabezudo et al., 2015), using Princeton WordNet as sense repository, as cited above. The selection of this corpus was motivated by the widespread coverage of topics and its previous use in other researches in this line.

In general, 5,082 verb instances were manually annotated in the corpus, which represent 844 different verbs and 1,047 synsets (senses). As the authors report, the corpus annotation achieved a 0.544 Kappa measure (Carletta, 1996), which is considered moderate (between 0.4 and 0.6, according to the literature), and a percent agreement of 38.5% and 56.09% for total and partial agreement, respectively. Given the difficulty of the task and the excessive sense refinement in WordNet, such numbers are considered satisfactory.

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

3.2 WSD Methods

The first method that we investigated was the traditional one proposed in (Lesk, 1986) (we simply refer to it by Lesk method). This method selects the sense of a word that has more common words with the words in its context window. For our work, we tested six variations for each target word: (G-T) comparing synset glosses with labels composed of possible translations in the word context; (S-T) comparing synset samples with labels composed of possible translations in the context; (GS-T) comparing synset glosses and samples with labels composed of possible translations in the context; (S-S) comparing synset samples with labels composed of the samples of all possible synsets for the context words; (G-G) comparing synset glosses with labels composed of the glosses of all possible synsets for the context words; and (GS2) comparing synset samples and glosses with labels composed of all possible synset samples and glosses for the context words. We also did some modifications in the size and balance of the context window. These modifications were motivated by a study presented in (Audibert, 2004), which says that verbs need unbalanced context windows. We used three window variations: 2-2, 1-2, and 1-3, where the first parameter represents the number of words at the left and the second one the number of words at the right of the target word.

The second one is a Web search-based method proposed in (Mihalcea and Moldovan, 1999) (referred by Mihalcea-Moldovan method). This method disambiguates a word in the context of other word. In our case, Mihalcea-Moldovan method selected the nearest content word for a target word as context word, then built one query for each synset of the target word and the possible translations of the context word. Finally, each query was posted in Bing® web search engine and the synset of the query with the best results was selected. In our case, the method tried to disambiguate a verb under focus with the nearest noun in the sentence. When there was more than one option of noun, we used two criteria to decide: using a randomly selected nearest noun in the sentence, or using the nearest noun at the right side of the verb.

The third one is a Graph-based method proposed in (Agirre and Soroa, 2009) (referred by Agirre-Soroa method). This method builds a semantic graph with all possible synsets of all content words in a sentence. Then, the PageRank algorithm (Brin and Page, 1998) is executed for each target word, giving priority to synsets of its context words in the sentence. For this work, the target words were the verbs and we tested two configurations: to disambiguate a verb using the context words in its sentence, and using the context words in its paragraph.

The last method is the one proposed in (Nóbrega and Pardo, 2014) (referred by Nobrega-Pardo method). This method is used in a multi-document scenario and assumes that all occurrences of a word in a text collection have the same sense. This works as follows: firstly, for each collection, the method creates a multi-document representation of the context words that co-occur with the target word in a pre-specified window; then, it selects the "n" most frequent context words and applies the Lesk method to disambiguate the target word. In this work, the window size and "n" had values of three and five, respectively, and the Lesk variations used were the G-T and S-T ones.

Besides the four WSD methods that we explored, we also tested two others as baselines. The first baseline was the Most Frequent Sense method (MFS), which is usually difficult to outperform in the area. For this work, the MFS method selected the first synset for a target word. The second baseline method was a random one. This method randomly selected a translation and then a synset for a target word.

3.3 Incorporating Linguistic knowledge

In this section, we will describe the incorporation of VerbNet.Br (Scarton, 2013) information into two WSD methods, one focused on the single document scenario (Lesk method) and one focused on the multi-document scenario (Nobrega-Pardo method). VerbNet.Br is a repository which groups verbs with similar syntactic/semantic behavior (Levin, 1993).

The basic assumption that we adopted was the following: if some verbs in a text belong to the same VerbNet.Br class, we may group their contexts to disambiguate them together. So, we defined two steps: (1) to group verbs (in clusters) according to VerbNet.Br classes, and (2) to enrich the context of the grouped verbs.

The idea of grouping verbs was motivated by the study presented in (Harris, 1954), which says that words in similar contexts tend to have similar senses. The way of grouping verbs was the use of a dominance criteria, which specifies that a greater quantity of verbs that belong to the same class indicates that they probably exhibit some relationship. In Figure 1, we may see an example in which all possible VerbNet.Br classes for each verb in a specified text are shown. As it may be seen, the VerbNet.Br class 1 ($VNClass_1$) includes most of the verbs (V_1, V_3, and V_5), and, therefore, this class might be considered a cluster. In this case, the other VerbNet classes would not form clusters because these would have only one verb ($VNClass_4$ and $VNClass_5$ in case of V_2, and $VNClass_3$ and $VNClass_7$ in case of V_4).

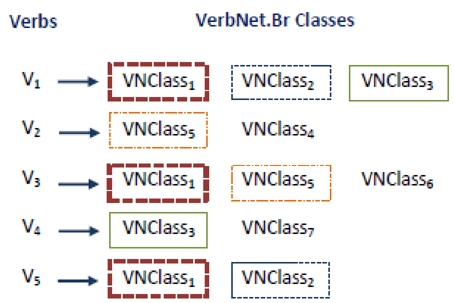

Figure 1: Possible VerbNet.Br classes for each verb

One problem in this step was that all possible VerbNet.Br classes were considered for each verb, introducing some noise. This was produced by the use of the lemma of the words instead of considering the syntactic/semantic behavior in the grouping step, that is how VerbNet.Br works.

To solve this problem, a refinement was performed using syntactic information. This information was obtained from the alignment between the output of PALAVRAS syntactical parser (Bick, 2000) and the Semantic Role Labeling system (SRL) proposed in (Alva-Manchego, 2013), using the model trained in (Hartmann, Duran, and Aluísio, 2016) to extract the necessary arguments (no adjuncts) to filter the VerbNet.Br classes. This alignment was necessary because VerbNet.Br only contains the arguments of the verbs. PALAVRAS produces full syntactic structures (without distinguishing between arguments and adjuncts), and the SRL identifies all semantic roles (without syntactic information), distinguishing among arguments and adjuncts.

In Figure 2, we may see the arguments and adjuncts of the verb "*reunir*" ("to meet", in English). Due to how VerbNet.Br was built, only the arguments/adjuncts after the verb were considered. Therefore, the structure obtained was "V AM-TMP AMP-PRP".

<ARG="AM-TMP">Em a quinta-feira**</ARG>**, **<ARG="A0">**a Mesa Diretora de o Senado**</ARG>** **<ARG="A0">**se**</ARG>** **<ARG="V">***reúne***</ARG>** **<ARG="AM-TMP">**a as 14 horas**</ARG>** **<ARG="AM-PRP">**para decidir se aceita a quarta representação contra o presidente de a Casa**</ARG>**.

Figure 2: Semantic Roles for the verb "*reunir*" ("to meet")

After this, PALAVRAS was executed and we did a process similar to the SRL to get the final syntactic structure to align. Finally, we did a mapping between the output of the SRL and PALAVRAS to get the relevant syntactic structure. Because VerbNet.Br only needs arguments, a filtering process was performed, eliminating the syntactic phrases related to adjuncts. In Figure 3, we may see the mapping between the output of SRL and PALAVRAS. In this case, the final syntactic structure for the verb "*reunir*" was simply "*V*", because "*PP*[*a*]" and "*PP*[*para*]" were related to adjuncts in the SRL.

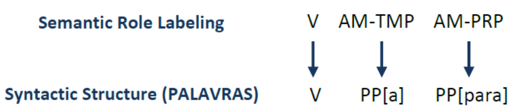

Figure 3: Mapping between the output of the Semantic Role Labeling system and the syntactic structure generated by PALAVRAS

At this point, we have to highlight that we considered some extra criteria related to include (or not) a verb in a cluster, exclusion of some VerbNet classes, and minimum number of verbs to form a cluster:

- Inclusion/exclusion of highly polysemic verbs: these verbs are called light verbs. For example, in "*fazer questão*" and "*fazer contas*", the verb "*fazer*"("to do") changes its sense ("to insist" in the first case, and "to count" in the second) according to the next word.
- Inclusion/exclusion of copula verbs: this kind of verbs is used for linking a topic to a comment.
- Exclusion of VerbNet class other-cos-53.2: this VerbNet class contains verbs that are not clearly related to other classes. Therefore, this class could bring noise in the clustering.
- Mininum number of verbs to form a cluster: we experimented with values in a range from two to nine.

In the second step (to enrich the context of the grouped verbs), we built the context for each target word in the verb cluster and then put together all the contexts. Finally, we selected the words that most co-occurred as context words and applied the WSD method to each target word in the cluster.

The two steps mentioned in the previous paragraphs were applied to each WSD method (Lesk and Nobrega-Pardo method), but the difference was that, in Lesk method, the grouping was performed considering the lemmas and the syntactic structures and, in Nobrega-Pardo method, the grouping was performed only using the lemmas because this method uses the heuristic of one sense per discourse, and the senses of the words are independent of syntactic structure.

In the case of the verb "*reunir*" ("*to meet*"), this was grouped with the verbs "*ocorrer*" ("*to happen*") and "*coordenar*" ("*to coordinate*"), and all of their individual contexts were grouped. In Figure 4, the co-occurrence graph for the cluster formed by "*reunir*", "*ocorrer*" and "*coordenar*" is presented, being "*representação*" ("representation") the most co-occurring word in the context. As mentioned before, the method selected the top "*n*" most co-occurring words as context of the cluster and then applied Lesk or Nobrega-Pardo method to determine the correct sense. In the graph, the method selected the word "*representação*" (most co-occurring) and "*líder*" and "*só*" (randomly selected) when the context size was three.

4 Evaluation and Results

The measures used in this evaluation were: Precision (P), which computes the number

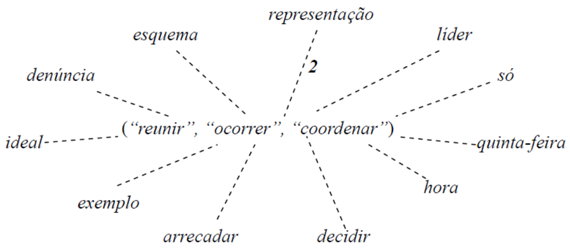

Figure 4: Co-occurrence graph of the cluster formed by "*reunir*", "*ocorrer*" and "*coordenar*"

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; Coverage (C), number of classified verbs over all the verbs in the corpus; and Accuracy (A), which is the same as (R), but using MFS method when no sense identification may be performed (Specia, 2007), as a back-off mechanism.

The classical WSD methods (described in Subsection 3.2) were evaluated in two tasks: All-words task, i.e., we evaluate all verbs in the corpus, and Lexical sample task, that consisted in evaluating a selection of verbs. The results for the All-words task, shown in Table 1, indicate that no method outperformed the MFS baseline, but all of them outperformed the Random baseline (Rnd). Analyzing the methods, we may note that Nobrega-Pardo method (NP) got the best results. This was due to the few sense variation for each word in the corpus. Mihalcea-Moldovan method (MM) got the worst performance. This is explained because the verb sense tends to be less stable in presence of different nouns. In relation to Coverage (C), we may note that no method reached 100 %. In case of MFS and Random methods, this occurred because some target verbs in corpus did not get translations from WordReference® and, therefore, did not get synsets from WordNet. In case of the other methods, the same problem occurred in target verbs and context words, causing lower results.

Method	P(%)	R(%)	C(%)	A(%)
MFS	49.91	47.01	94.20	-
Random (Rnd)	10.04	9.46	94.20	9.46
Lesk (L)	40.10	37.69	93.98	37.77
Mihalcea-Moldovan (MM)	17.21	14.43	83.87	19.44
Agirre-Soroa (AS)	28.45	26.80	94.20	26.80
NobregaPardo (NP)	40.33	37.97	94.14	38.00

Table 1: Results for the All-words task

We have to note that all results shown in Table 1 are the best results for each studied method. Thus, the best configuration for the Lesk (L) method was using the S-T variation and an unbalanced window with one word at the left of the target word and two words at the right. For Mihalcea-Moldovan method, the best result was obtained using the nearest noun at the right side of the target word. In relation to the Agirre-Soroa (AS) method, the use of paragraph as a context to disambiguate a verb yielded the best results. Finally, the best result for the Nobrega-Pardo method was obtained using the S-T variation and a window size of three.

The Lexical sample task was performed considering the twenty more polysemic verbs in the corpus. The verbs are shown in Table 2 with their Frequency (F) of occurrence and number of Senses (S) in the corpus.

The Precision measure was evaluated in order to compare the performance of all WSD methods over a well-defined sample. In general, Table 2 shows similar results to Table 1. One point to highlight was that Nobrega-Pardo method was positioned in the second place. This reflected the few verb sense variations and the dominance of a sense in the corpus. Lesk and Agirre-Soroa methods showed similar results in both tasks.

In Table 3, we may see the performance comparison of the best WSD method for verbs, i.e., Nobrega-Pardo method, with the same WSD method for nouns, which were evaluated in (Nóbrega and Pardo, 2014). The results show that the verb sense disambiguation task is in fact more difficult than the noun sense disambiguation, confirming what is cited in (Miller et al., 1990).

The results of the incorporation of Linguistic Knowledge (LK) from VerbNet.Br to the Lesk and Nobrega-Pardo methods are presented in Table 4. Both methods outperformed the original methods, but this difference was not statistically significant using the Wilcoxon test at the 95% confidence level. In the case of Lesk method, the best results were obtained when all highly ambiguous verbs and copula verbs were considered and the minimum number of elements by group was four. In the case of Nobrega-Pardo method, the best results were obtained when copula verbs were considered and the minimum number of elements by group was seven. Some of the problems that produced

Verb	F	S	MFS	Rnd	L	MM	AS	NP
ser ("to be")	450	14	88.11	8.59	69.32	27.40	58.37	72.69
ter ("to have")	143	10	75.82	5.88	62.75	5.44	5.23	67.97
fazer ("to do")	93	18	31.62	0.85	11.11	0.00	1.71	14.53
apresentar ("to present")	38	8	50.00	0.00	36.11	20.00	0.00	47.22
chegar ("to arrive")	55	12	29.09	3.64	23.64	20.41	27.27	23.64
receber ("to receive")	36	9	61.11	0.00	42.86	9.38	11.11	58.33
ficar ("to stay")	58	16	11.27	1.41	8.45	3.13	8.45	8.45
registrar ("to register")	27	8	3.85	3.85	7.69	20.00	15.38	3.85
deixar ("to leave")	49	16	19.61	1.96	13.73	2.00	7.84	19.61
cair ("to fall")	24	8	17.39	0.00	17.39	0.00	0.00	17.39
passar ("to pass")	44	15	38.30	2.13	23.40	2.56	8.51	29.79
fechar ("to close")	21	8	36.84	0.00	5.26	23.08	0.00	21.05
colocar ("to put")	20	8	63.16	5.26	31.58	6.25	52.63	21.05
encontrar ("to find")	24	10	12.50	4.17	4.17	4.17	4.17	0.00
levar ("to take")	31	13	9.09	0.00	3.03	0.00	6.06	0.00
vir ("to come")	18	8	30.00	5.00	30.00	0.00	0.00	15.00
estabelecer ("to establish")	12	7	8.33	8.33	16.67	9.09	16.67	8.33
marcar ("to mark")	12	7	0.00	0.00	9.09	10.00	36.36	0.00
dar ("to give")	22	14	13.21	3.77	9.43	4.00	0.00	7.55
tratar ("to treat")	9	7	11.11	11.11	22.22	11.11	22.22	0.00
Precision	-	-	30.52	3.30	22.39	8.90	14.10	21.82

Table 2: Results for the Lexical sample task

misclassifications were (1) the missing of syntactic frames in the VerbNet.Br classes, and (2) VerbNet.Br classes without syntactic filters, producing noise during verb grouping.

Method	P(%)	R(%)	C(%)	A(%)
NP-Verbs	40.33	37.97	94.14	38.00
NP-Nouns	49.56	43.90	88.59	43.90

Table 3: Comparative results of Nobrega-Pardo method for nouns and verbs

Method	P(%)	R(%)	C(%)	A(%)
Lesk+LK	40.28	37.87	94.00	37.95
NP+LK	41.02	38.48	93.80	38.52

Table 4: Results of Lesk and Nobrega-Pardo methods with Linguistic Knowledge (LK)

5 Conclusions and Future Work

In this work, we evaluated some classical WSD methods for verbs in Brazilian Portuguese and the performance variation when we incoporated linguistic knowledge (from VerbNet.Br) to two classical methods (one based on single document scenario and other on multi-document scenario). Another contribution of this work is the sense annotation of a corpus and its free availability.

Although the sense repository we used is in English (the Princeton WordNet), we believe that this did not compromise the performance of the WSD methods for Portuguese. However, there were some lexical gaps that we could notice. For example, the verb "pedalar" (a kind of drible in soccer) has no specific synset in Princeton WordNet. For these cases, the verb should be generalized (to drible).

One future work is to explore some voting schemes in ensemble methods to take advantages of the variability offered by the different WSD methods. Furthermore, we intend to incorporate selectional restrictions in the verb grouping step. Some studies mention that the semantics of the verb arguments may help in WSD.

Acknowledgments

To CAPES, FAPESP and Samsung Eletrônica da Amazônia Ltda., for supporting this work.

References

Agirre, E. and A. Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41.

Aires, R. V. X., S. M. Aluísio, D. C. S. Kuhn, M. L. B. Andreeta, O. N. Oliveira, and Jr. 2000. Combining multiple classifiers to improve part of speech tagging: A case study for Brazilian Portuguese. In *Proceedings of the Brazilian Artificial Intelligence Symposium*, pages 20–22.

Alva-Manchego, F. 2013. *Anotação Automática Semissupervisionada de Papéis Semânticos para o Português do Brasil.* MSc thesis, Instituto de

Ciências Matemáticas e de Computação, Universidade de São Paulo.

Audibert, L. 2004. Word sense disambiguation criteria: a systematic study. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Baptista, J. 2012. Viper: A lexicon-grammar of European Portuguese verbs. In J. Radimsky, editor, *Proceedings of the 31st International Conference on Lexis and Grammar*, pages 10–16.

Bick, E. 2000. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus.

Brin, S. and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, pages 107–117.

Cabezudo, M. A. S., E. Maziero, J. Souza, M. Dias, P. C. Cardoso, P. P. B. Filho, V. Agostini, F. A. Nóbrega, C. de Barros, A. D. Felippo, and T. A. Pardo. 2015. Anotação de sentidos de verbos em textos jornalísticos do corpus CSTNews. *Revista de Estudos da Linguagem*, 23(3):797–832.

Cardoso, P., E. Maziero, M. Castro Jorge, E. Seno, A. Di Felippo, L. Rino, M. Nunes, and T. Pardo. 2011. CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Harris, Z. 1954. Distributional structure. *Word*, 10(23):146–162.

Hartmann, N. S., M. S. Duran, and S. M. Aluísio. 2016. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 202–212.

Jurafsky, D. and J. H. Martin. 2009. *Speech and Language Processing*. Prentice-Hall.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26.

Levin, B. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Machado, I. M., R. O. de Alencar, R. de Oliveira Campos Junior, and C. A. Davis. 2011. An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, 17(4):267–279.

Mihalcea, R. and D. I. Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 152–158.

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

Nóbrega, F. A. A. and T. A. S. Pardo. 2014. General purpose word sense disambiguation methods for nouns in Portuguese. In *Proceedings of the 11th International Conference on Computational Processing of the Portuguese Language*, pages 94–101.

Scarton, C. E. 2013. *VerbNet.Br: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil*. MSc thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Specia, L. 2007. *Uma Abordagem Híbrida Relacional para a Desambiguação Lexical de Sentido na Tradução Automática*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Suissas, G. 2014. *Verb Sense Classification*. MSc thesis, Instituto Superior Técnico, Universidade de Lisboa.

Travanca, T. 2013. *Verb Sense Disambiguation*. MSc thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.