General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese*

Fernando Antônio Asevedo Nóbrega and Thiago Alexandre Salgueiro Pardo

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
{fasevedo,taspardo}@icmc.usp.br

Abstract. Word Sense Disambiguation (WSD) aims at determining the appropriate sense of a word in a particular context. Although it is a highly relevant task for Natural Language Processing, there are few works for Portuguese, which are tailored to specific applications, such as translation and information retrieval. In this work, we report our investigation of some general purpose WSD methods for nouns in Portuguese, tackling two additional challenges: using Princeton Wordnet (for English) as the sense repository and applying/customizing a WSD method for multi-document applications, which, to the best of our knowledge, has not been addressed before. In this paper, we also report our efforts on building a sense annotated corpus (for nouns, only), which was used for evaluating the investigated WSD methods.

1 Introduction

Word Sense Disambiguation (WSD) aims at determining the appropriate sense for a word in a particular context [1]. For example, in the sentence "the bank has borrowed money in a variety of ways", the word "bank" might be associated to multiple senses, as: sloping land; a financial institution (which is the correct sense); a supply held in reserve for future use; etc. For many applications, it is essential to know the correct word senses in order to produce good results. Consider, for instance, machine translation: it is necessary to know the sense for correctly translating the word to the target language. Therefore, WSD is an important area, dealing with a phenomenon of semantic nature and, as such, is as difficult as useful. However, for the Portuguese language, there are only a few WSD works, which are tailored to specific applications. For example, [14] shows a disambiguation method for ten highly ambiguous verbs in English, aiming at machine translation applications. [9], in turn, presents a WSD method for ambiguous geographical words, as "São Paulo", which, in Portuguese, may refer to a city, a state, a soccer team or even a saint.

In this work, we are interested in investigating general purpose WSD methods for Portuguese. As a departure point, we opted for dealing only with nouns, which

* The corresponding MSc dissertation is available at www.teses.usp.br/teses/disponiveis/55/55134/tde-28082013-145948

are usually the most frequent open morphosyntactic class found in texts, and, as [13] claims, the disambiguation of nouns is enough to positively impact some applications.

In particular, we dealt with two specific challenges: a multilingual challenge, since we use Princeton Wordnet [12] as sense repository, whose synsets were considered as the possible senses that the words in Portuguese might be associated to; and a multi-document challenge, in that we are also interested on checking how multiple documents may contribute to the WSD task. The former decision of using Princeton Wordnet as sense repository is due to (i) its widespread use in the area for WSD and also for other applications, (ii) it has been manually produced , and (iii) the current partial development state of most of the similar resources for Portuguese. The decision of also dealing with multiple documents comes from the current high demand of multi-document processing tasks, as information retrieval and extraction and multi-document summarization.

Our main research hypotheses that guided our work are that it is possible to achieve good accuracy with general purpose WSD methods, that the English sense repository suffices for representing most of the senses found in Portuguese texts, and that multiple documents may present more context information for improving WSD results. We have used a corpus of news texts written in Brazilian Portuguese to conduct this investigation. We have investigated and made the necessary adaptations to three classical methods, namely, the heuristic method that always select the more frequent sense for the words, the traditional Lesk algorithm [8] and some of its variations, and the web-based proposal of Mihalcea and Moldovan [10]. For multi-document WSD, we explore a graph-based multi-document representation, inspired by the work of Agirre and Soroa [2].

In general, our methods adopt the following procedures: considering a sentence in Portuguese, we initially find the appropriate translation to English (it is a necessary step considering that we use Princeton Wordnet); and, then, given the possible synsets in Wordnet for the translations, the synset that represents the best sense for the word must be selected. In this work, we use the online bilingual dictionary WordReference® to automatically perform the translations (other tools have also been tested, as Google Translate®, but there were limitations of use and license). Our results show that the Mihalcea and Moldovan approach is better to disambiguate highly ambiguous words. This method also performs well for all the words, together with the classical (heuristic) method and our graph-based multi-document proposal.

This paper is organized in 6 sections. In section 2 we briefly present some related work. The corpus and the sense annotation task are described in Section 3. Section 4 shows the developed WSD methods, while their evaluation is described in Section 5. Section 6 presents some final remarks.

2 Basic Concepts and Related Work

WSD usually considers three main elements: 1) the target word to be disambiguated; 2) its context (usually, words surrounding the target word); and 3)

a sense repository, as dictionaries, thesaurus, ontologies and wordnets. Some methods try to disambiguate all the words in a text ("all words" task), while some try to disambiguate only a group of words ("lexical sample" task). The methods may also be classified as corpus-based or knowledge-based. The formers use annotated corpora for producing machine learning classifiers. Generally, they perform the lexical sample task. The knowledge-based methods use linguistic resources and similarity measurements to be able to deal with a wider range of words. In this work, as we aim at general purpose methods, knowledge-based approaches are more appropriate. In what follows, we briefly introduce the main knowledge-based works on which we base our investigation.

Lesk [8] presents what is considered the most classical method for WSD. It is a simple method based on machine-readable dictionaries. It assumes that the best sense is the one whose definition in the dictionary is the most similar to the labels of the target word context (by comparing their words). Some authors, as [7], [3] and [15], present variations to the methods, considering wordnet synsets and their glosses and examples for performing the comparisons.

Mihalcea and Moldovan [10] propose an unrestricted method for WSD, based on the use of the web. The authors assume that the correct sense should be the one that occurs more frequently with the target word context in the web. In this method, the context is a single word, usually the closest word according to the observed syntactic pattern. For example, to disambiguate a noun, the context is the nearby verb.

Agirre and Soroa [2] propose a graph-based method. The authors use the PageRank algorithm [4] in a graph with words and synsets as nodes. The authors initially use the hierarchical structure of Wordnet to initialize the graph. In the next step, edges are created among words and their respective synsets. Finally, PageRank is applied in the graph to rank the synsets, with the best one being chosen as the correct sense.

3 The CSTNews corpus

We have used the CSTNews[1] corpus [5] for testing the WSD methods. It has 140 news texts grouped by topic into 50 clusters. Each cluster has 2 or 3 texts. Since the corpus was not annotated with senses, this annotation was carried out as follows.

Given the difficulty of the sense annotation task and the limited time and human resources to annotate the full corpus, we opted to annotate the 10% most frequent nouns in each cluster. Each cluster was manually tagged by groups of two or three human annotators. For each new cluster to annotate, the annotation groups were mixed, in order to avoid any annotation bias. The task was carried out by 10 annotators in 1-hour daily meetings, during five weeks. To assist the participants, an easy-to-use annotation tool was built, called NASP[2] (in Portuguese, *NILC – Anotador de Sentidos para o Português*).

[1] Available at www.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html
[2] Available at www.icmc.usp.br/pessoas/taspardo/sucinto/files/NASP.zip

In general, 4366 words were annotated, with 519 different translations and 575 distinct synsets. The number of distinct synsets annotated per word in the corpus ranged from 1 to 5. However, when limiting the counting of different synsets for the clusters, the number ranged from 1 to 3. In general, 93% of the words happened with only one sense in their clusters; 6% had 2 senses; and 1% had 3 senses. This information suggests that to use multi-document information may assist the WSD in this context.

We analyzed the sense annotation task complexity by counting the number of available synsets for the annotation of each word. Figure 1 shows the results. One may see that it happens that some words have more than 50 possible synsets, which clearly shows how difficult the annotation task is. On the other extreme, some words have zero synsets. This happen for two reasons: our synset search approach could not find appropriate translations for some words; and/or there were no good synsets in Wordnet for the intended sense. For example, the word *licenciamento* ("licensing", in English) has no translation in WordReference® dictionary; on the other hand, the word *desabrigado* has two translation options ("unsheltered" and"unprotected") in the dictionary, but no corresponding synset in Wordnet (for noun category).

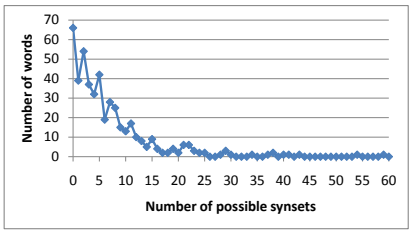

Fig. 1: Number of possible synsets per word in CSTNews noun annotation task

Table 1: Sense annotation task agreement in CSTNews

	Kappa	Percent agreement (%)		
		Total	Partial	Null
Translation	0.85	82.87	11.08	6.05
Synset	0.72	62.22	22.42	15.36
Translation + Synset	0.69	61.21	24.43	14.36

The annotation agreement was measured by both percent agreement counts and kappa [6]. For percent agreement, we computed the total agreement (when the annotators fully agreed), partial agreement (when the majority of the annotators agreed) and null agreement (for the remaining cases). Given the multilingual characteristic of our work, the agreement was measured for three elements: the chosen translations, the selected synsets, and the translation-synset pairs. Differently from percent agreement, the Kappa measure discounts the agreement by chance. Its values range from 0 to 1, with 1 indicating perfect agreement. Several researches consider that a kappa value above 0.6 is enough to have a reliable annotated corpus (although such value depends on the subjectivity and difficulty of the performed annotation). Table 1 shows the obtained agreement values. The rows show the evaluated elements, while the columns exhibit the agreement values. As expected, translation was easier than synset selection, but directly influences it. Once a bad translation is chosen, it is hard to find a good synset. Again, as expected, the agreement is lower when we consider the perfect

matching of translations and synsets among annotators. In general, the obtained values indicate that the annotation is reliable and may be used for the intended purpose.

4 Developed Methods

We tested four WSD methods. Besides translating and recovering synsets, all methods use the following pre-processing steps: (1) sentence splitting; (2) part-of-speech tagging; (3) removal of stopwords; (4) lemmatization of the remaining words; and (5) target words detection and context representation. Initially, we implemented the heuristic (baseline) method, frequently used in the literature, which assigns to the target word the most frequent sense in Wordnet (usually the first in the list), considering only the most frequent translation (the first one too) in the bilingual dictionary.

Our second method is an adaptation of Lesk algorithm [8] based on the work of [3]. Our algorithm has six variations: (G-T) using synset Glosses to compare with labels composed of possible word Translations in the context; (S-T) using synset sample Sentences to compare with labels composed of possible word Translations in the context; (GS-T) using synset Glosses and sample Sentences to compare with labels composed of possible word Translations in the context; (S-S) only synset sample Sentences to compare with labels composed of the sample Sentences of all possible synsets for the context words; (GS2) synset sample Sentences and synset Glosses to compare with labels composed of all possible synset sample Sentences and Glosses for the context words.

The third method is the WSD algorithm of Mihalcea and Moldovan [10] (which we will simple refer by Mihalcea method). In this method, we build word pairs for posting queries in the web. A word pair consists of the noun under focus and the nearest verb in the sentence. We use Microsoft Bing®for searching the web.

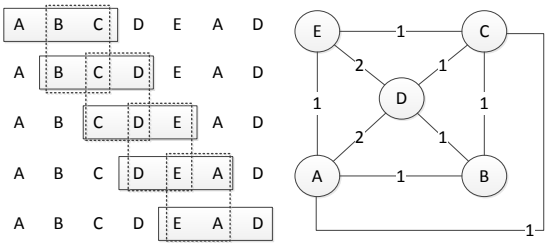

Fig. 2: Example of Multi-document Word Co-occurrence Network

The fourth method uses the best variation of the Lesk algorithm (G-T) for the multi-document WSD. This method uses a multi-document representation of context and assumes that all the occurrences of a word in a cluster have only one sense based in our corpus evidence. Furthermore, this method makes a second assumption: finding the most related words (which co-occurred the most) with the target word in its cluster helps selecting relevant context words and

suggesting the best synset. We use a graph, which we call Multi-document Word Co-occurrence Network (MWCN) (adapting from [11]) to represent the multi-document context. Thus, each node represents a single word in the cluster, and the edges indicate the frequency of occurrence of the corresponding words in a moving window of size N (which is a parameter of the algorithm), restricting that the same word pair in the same windown contributes only once to the edge weight. Figure 2 shows a MWCN for a sequence of hypothetical words A,B,C,D,E,A,D, using windows with three words (N=3). One may see that the words in the same windows (solid squares) are connected in the graph. The edge shows the number of windows in which the corresponding letters occurred together. For example, the edge A-D has weight 2 (since the words occurred in the fourth and fifth windows). It is important to note that window overlaps (indicated by dashed squares) contribute only once to the graph.

In the disambiguation process, the algorithm applies the G-T method using the N most related words with the target word in the MWCN. For example, for disambiguating the word D in the MWCN in Figure 2, the context words were A, E and either B or C. In our experiments, we test MWCN with windows sizes N=3 (MWCN3) and N=5 (MWCN5).

5 Evaluation

The WSD methods were evaluated on two tasks: the "all words" task; and the "lexical sample" task – in our case, the 20 most ambiguous words in the corpus. While the former task is necessary for measuring the coverage of the methods, the latter gives an idea of the robustness when dealing with difficult words.

In the all words task, the evaluation was measured by four metrics: (P) Precision – number of correct classifications over the number of words classified by the method; (R) Recall – number of correct classifications over the total number of tagged words in the corpus; (C) Coverage - number of words classified (correctly or not) by the method over the total number of tagged words in the corpus; and (A) Accuracy – equal to R, but using the heuristic method when no classification is found.

Table 2 shows the obtained results. The rows show the methods and the columns show the average values for the metrics. We show only the best configuration for the Lesk method (G-T). One may see that the MWCN3 method achieved 49.56% precision, and the Mihalcea method achieved 99.41% coverage. The heuristic method achieved the highest accuracy (51%). This may be explained by the fact that, in the CSTNews corpus, the majority of the words were annotated with the most frequent synset and this is a good scenario for this method. However, we did not found statistical difference in the precision and coverage values for the MWCN3 and heuristic methods.

Besides the heuristic method, Mihalcea method shows the best coverage. This is explained by the shorter context window (only one word). The MWCN3 e MWCN5 methods were better than Lesk [8] results for precision and coverage. It is important to note that the MWCN3 and MWCN5, for multi-document pur-

poses, are better for WSD in this scenery. It is also interesting to notice that most of the accuracy values are similar to recall values, which indicates that the heuristic method do not find the correct senses for words that were not classified by others methods.

Table 2: All-word evaluation

Method	P(%)	R(%)	C(%)	A(%)
Heuristic	51.00	51.00	100	51.00
G-T	42.20	41.20	91.10	41.20
Mihalcea	39.71	39.47	**99.41**	39.59
MWCN3	**49.56**	**43.90**	88.59	43.90
MWCN5	46.87	41.80	87.65	41.80

Table 3: Lexical-sample evaluation

Word	Heuristic	S-T	Mihalcea	MWCN(3,5)
ano	90.50	86.30	**94.83**	47.22
hora	50.00	**50.00**	**50.00**	0.00
local	30.00	**30.00**	**33.33**	17.65
vez	0.00	**10.50**	**0.00**	**0.00**
>= Heuristic	-	12	**13**	8
Avg precision	27.88	28.46	**32.37**	19.10

In the lexical sample task, we used only the precision metric in order to evaluate the quality of the methods. Table 3 shows the evaluation results. The rows show some ambiguous words (with varying precision values) and the columns show each method (again, we show only the best configurations for Lesk method). For example, Mihalcea method achieved a 94.83% precision for the word *ano* ("year"). The bold values indicate cases that the methods performed as well as or better than the heuristic method. In general, Mihalcea method was the best method for dealing with highly ambiguous words. The MWCN3 and MWCN5 are presented together because they produced the same results. The last rows show the number of times that the methods were better or equal than the heuristic method. The last row shows the average precision of the methods for all the 20 words, with the Mihalcea method being the best one, in overall.

6 Final Remarks

As far as we know, this is the first investigation of general purpose WSD methods for nouns in Portuguese. We evaluated some classical methods and also proposed a new one that takes into account the available multi-document information. Another contribution of this work is the annotation of a corpus with noun senses, which is freely available for use.

Although Princeton Wordnet is a valuable resource (given its widespread use), our methods suffer with some lexical gaps. For instance, the word *caipirinha* (which is a typical drink in Brazil) has no specific synset in Wordnet. Instead, for cases like this, more generic synsets must be adopted (as the "drink" synset). The opposite – the specification – also happens. While in Portuguese we have the word *dedo* (does not mattering if it refers to hands or feet), in English it is necessary to decide among the specific words "finger" (if one talks about hands) or "toe" (about feet). For automatic WSD methods, this is a challenge that remains for future work.

Acknowledgements. To FAPESP and CNPq, for supporting this work.

References

1. Agirre, E., Edmonds, P.: Introduction. In: Word Sense Disambiguation: Algorithms and Applications. Springer (2006) 1–28
2. Agirre, E., Soroa, A.: Personalizing pagerank for word sense disambiguation. In: Proceedings of 12th Conference of the European Chapter of the ACL. (2009) 33–41
3. Banerjee, S.: Adapting the Lesk algorithm for word sense disambiguation to wordnet. Master's thesis, Department of Computer Science, University of Minnesota (2002)
4. 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)
5. Cardoso, P.C.F., Maziero, E.G., Jorge, M.L.R.C., Seno, E.M.R., Di Felippo, A., Rino, L.H.M., Nunes, M.d.G.V., Pardo, T.A.S.: CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In: Anais do III Workshop "A RST e os Estudos do Texto", Cuiabá, MT, Brasil, Sociedade Brasileira de Computação (2011) 88–105
6. Carletta, J.: Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics **22** (1996) 249–254
7. Kilgarriff, A., England, B., Rosenzweig, J.: English senseval: Report and results. In: Proceedings of 2nd International Conference on Language Resources and Evaluation. (2000) 1239–1244
8. 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, New York, NY, USA, Association for Computing Machinery (1986) 24–26
9. 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. Journal of the Brazilian Computer Society **17** (2011) 267–279
10. 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, College Park, Maryland, USA, Association for Computational Linguistics (1999) 152–158
11. Mihalcea, R., Radev, D.: Graph-based natural language processing and information retrieval. Cambridge University Press (2011)
12. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38** (1995) 39–41
13. Plaza, L., Diaz, A.: Using semantic graphs and word sense disambiguation techniques to improve text summarization. In: Proceedings of Procesamiento del Lenguaje Natural. Volume 47. (2011) 97–105
14. 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 (2007)
15. Vasilescu, F., Langlais, P., Lapalme, G.: Evaluating variants of the Lesk approach for disambiguating words. In: Proceedings of Language Resources and Evaluation (LREC 2004), Lisbon, Portugal (2004) 633–636