A SUPERVISED MACHINE LEARNING METHOD FOR WORD SENSE DISAMBIGUATION OF PORTUGUESE NOUNS

Marcos Zampieri
University of Wolverhampton
University of Algarve

Abstract

Word Sense Disambiguation (WSD) is vital in many Natural Language Processing (NLP) applications. This work aims to explore supervised machine learning techniques for the disambiguation of Portuguese nouns. For the comparison of different WSD algorithms and techniques, a selection of ambiguous words from a Portuguese academic vocabulary was taken and a catalogue of word senses was established for each of them. A training corpus of real occurrences of each word in context was collected, providing manually annotated contextual data for each sense of the ambiguous word. The corpus was processed and features were extracted using Python and the Natural Language Tool Kit (NLTK) and later classified using Naïve Bayes, Maximum Entropy and Decision Tree Algorithms.

Key-Words: Word Sense Disambiguation, Machine Learning, Automatic Disambiguation, Portuguese

1. Word Sense Disambiguation Task

Many words have more than one meaning in natural language, and each one of them is determined by its context. For example, the Portuguese word *apêndice* (*appendix* in English) is defined in commonly used dictionaries such as Houaiss for Brazilian Portuguese and Porto Editora for European Portuguese as:

1. (book part) A separate part at the end of a book or magazine which gives additional information to the readers;
2. (body part) A small tube-shape part which is joined to the intestines;

Both senses of *apêndice*, book part or body part in their respective context, are easy to recognize for any Portuguese native or competent speaker. However, for computational applications, this distinction is not always trivial and can generate problems in language processing.

The automated process of recognizing word senses in context is known Word Sense Disambiguation (WSD). The interest in WSD has increased in recent years in an attempt to improve the performance in several language processing tasks that need to deal with ambiguous words. However, its need had already been detected in early NLP applications (Stevenson and Wilks, 2003).

1.1. Applications

The WSD task is an important component of several NLP systems, such as Machine Translation (MT), Question Answering (QA) Information Retrieval (IR), Information Extraction (IE) and speech processing applications. Researchers in MT have concentrated efforts on WSD since the earliest NLP applications (Stevenson, 2003). MT researchers identified that their results would be considerably better by using WSD methods to disambiguate words in automatic translation for various pairs of languages, such as English and Portuguese (Specia, 2007).

In Question Answering applications, WSD is useful to link question words to answer words. When users pose questions to a QA system it is very likely that the question will contain at least one ambiguous word. Therefore it is necessary for the system to decode the question with the correct sense of the ambiguous word according to context in order to search for the correct answer.

Information Retrieval is another language processing application that benefits from WSD. Most of the words used to execute queries in IR systems have more than one meaning and therefore when performing a query the system may retrieve documents which are not relevant to the search (Kulkarni et. al., 2007).

In speech recognition systems, a WSD module can be used to increase systems´ performance by distinguishing senses for homophone and homograph words according to their context. A speech synthesis application may also benefit from WSD and generate more accurate and natural pronunciation as discussed by Yarowsky (1996) and (1997).

1.2. Sample Application WSD for IR – REAP.PT

The initial idea of this work emerged from an application described in the previous section: Information Retrieval. More precisely, a disambiguation module is considered necessary to increase performance of an IR engine developed to be part of a computer-aided language learning (CALL) software, the REAP.PT (Marujo, 2009).

REAP.PT is the Portuguese version of REAP and it is currently in development by an interdisciplinary research group in Portugal. REAP was originally developed for English at Carnegie Mellon University (CMU) (Collins-Thompson and Callan, 2004) and aims to improve language skills and vocabulary learning, by retrieving texts from the Internet according to specific criteria and the students´ preferences. After retrieval, the texts are presented to the students along with reading-practice exercises to help them acquire new vocabulary or new contexts for words already known.

The REAP research team at CMU is currently carrying out experiments in WSD for the English language (Kulkarni et.al., 2007) but so far, the WSD module has not been integrated. As the REAP is currently being developed for Portuguese, the aim of this experiments is to replicate the first experiments described for English for the disambiguation of Portuguese nouns, in so doing, this work will constitute the foundations for the development of a wider disambiguation module for use in the Portuguese version of REAP.

2. Background Information

The first need for disambiguation emerged from MT systems, thus the very first approaches to WSD considered the task as part of the analysis module of MT systems and WSD was therefore not considered as a topic of study on its own. The first attempts to address lexical ambiguity as an autonomous task occurred in the 1980´s. Among the first studies of automatic disambiguation on its own, Hirst (1997) aimed to provide an abstract semantic representation of the entire input text, making it possible to distinguish senses of ambiguous words in the text. Even though conceptually lexical ambiguity could be resolved by semantic representation, further studies have shown that this kind of approach aims too high due to what is described in the literature as the knowledge acquisition bottleneck.

Again in the 1980´s, dictionary publishers started to develop electronic versions of their dictionaries, and this solved one of the bottlenecks of the early WSD approaches, the coverage of lexicons. Lesk (1996) was one of the first researchers that tried to disambiguate Machine Readable Dictionaries (MRD) definitions using algorithms His algorithm became well-known among WSD researchers.

The Lesk algorithm is based on the assumption that words in a given neighbourhood will tend to share a common topic, and therefore it aims to disambiguate words in short phrases. Given an ambiguous word, the dictionary definition of each of its senses is compared to the definitions of every other word in the sentence. The algorithm assigns the word sense whose definition shares the largest number of words in common with the definitions of the other words. The algorithm begins a new process for each new word and does not use the senses it previously assigned.

Given the level of polysemy in MRD, these resources were later not considered ideal for WSD, which lead researchers to explore other sources of knowledge, especially corpora. One of the pioneer studies on corpus usage in WSD was detailed in the paper by Ng and Lee (1996). In this approach, called “exemplar-based learning”, the word sense was assigned to the sense of the most similar example already seen by the system. This approach is considered to be a supervised learning approach which requires previously disambiguated training text.

Algorithms for WSD can rely on rules to assert the correct sense of a word however this kind of approach is not as widely used in state-of-the-art applications as it was in early approaches. The use of statistical methods and machine learning techniques has significantly increased in the last few years, not only in WSD but in NLP applications as a whole.

2.1. Machine Learning in WSD

In the last decade, the NLP community has observed a research paradigm shift from rule-based approaches to statistical and machine learning approaches.

This change was observed in a wide range of applications in NLP such as Machine Translation, pre-processing tasks as POS tagging (Marquez, Padro and Rodriguez, 1999) and the task described here, Word Sense Disambiguation. A reasonable number of tools for NLP researchers developed in the past few years contain plug-ins and integration to ML toolkits, such as the Natural Language Toolkit (NLTK) (Bird, Klein and Loper, 2009) which will be used for the experiments described here.

Machine Learning involves a computer algorithm learning from data. Based on a set of predefined features, algorithms identify patterns in data and can therefore infer predictions. Several works describe the use of machine learning algorithms to the word sense disambiguation task, such as Kulkarni et. al. (2007), as well as Yarowsky (1996).

A wide range of features can be used in WSD. In particular, *collocational* features that specify words which can appear in specific locations before and after the target word. Usually, this is set to a pre-defined window of two, sometimes three words on each side. Binary features are also used to define the presence or absence of a word in the sentence and therefore provide more intuition to the context. Syntactic information about words as well as information about the POS of neighbouring words, namely POS bi-grams, can also be employed to increase results.

For these experiments, three major groups of features were used: Label Feature, Neighbouring Words and Key Words. None of these features depend on external linguistic resources. The idea behind this implementation was to extract all the information necessary for classification direct from the raw data without using any additional information.

3. Methods and Preliminary Results

The experiments proposed in this work begin with the selection of ambiguous words from a Portuguese academic vocabulary, the Portuguese Academic Word List P-AWL (Baptista et. al., 2010). The selection and sense distinctions for these words were made according to linguistic criteria discussed further.

Based on a random sample of 100 occurrences of each word extracted from corpus, the Most Frequent Sense (MFS) baseline was established. Along with the MFS baseline, these sample occurrences were used to calculate a Kappa coefficient in order to measure inter-annotator agreement.

Finally, corpus data was prepared for supervised classification into two stages, namely: tokenization and feature extraction and then classified using the algorithms Naive Bayes, Maximum Entropy and Decision Tree in their current implementation in Python NLTK.

3.1. Word Selection

As the main motivation for this research is restricted to academic texts, the selection of the vocabulary used in this task was based on academic vocabulary. For this selection the Portuguese Academic Word List (Baptista et. al., 2010), P-AWL, was used as the first resource. P-AWL contains 1823 words in different POS categories and the list was scrutinized word by word to compile a list of 13 ambiguous nouns with two major senses that are clearly distinguishable for any Portuguese native or competent speaker. It also took into account the possible differences of word sense between the Brazilian and European Portuguese.

Ambiguity between two words with two different POS, were not considered for these experiments, once they can be are resolved at the level of the POS tagging. State-of-the-art results for POS taggers are currently above 95%. For the sample application in this work, the REAP.PT, WSD will be in future integrated in an NLP chain (Mamede et. al., 2010), which features a POS tagger that reports above 97% precision.

Word	**Count**	**Word**	**Count**
Apêndice	85	Foco	704
Arquivo	928	Garantia	1925
Comissão	4209	Geração	2618
Crédito	4360	Imagem	9114
Cultura	5670	Regime	2736
Essência	521	Volume	4198
Etiqueta	311	**Overall**	**78909**

Table 1: Frequency of words in the corpus.

After this step the word *apêndice* was disregarded because it occurred only 85 times. A minimum amount of 100 examples was established in order to have enough training and testing examples for the classifier. In this case, even though *apêndice* and its English equivalent *appendix* is a classic example for disambiguation and a frequent word in everyday vocabulary, the NILC corpus containing journalistic texts does not contain enough occurrences of the word.

3.2. Training and Text Corpora

The NILC corpus available at Linguateca[1] was used to collect the examples for the training corpus. Queries in the database were made for each word in the wordlist in order to capture all of the occurrences in the corpus. The queries were made using simple regular expressions in the corpus interface.

[1] www.linguateca.pt

For the selection of the corpus sentences, not all the occurrences of the words were considered and only a predefined set of forms were taken into account. Nouns in their basic form (no diminutives, augmentatives or superlatives) were selected.

Linguistic criteria were established to ensure that all the examples or instances, for training and testing had the necessary length, clear context and distribution to be classified by the algorithms. For example short sentences provide usually less information for automatic disambiguation. The position of the target word is also important and in the case of the word appearing in the beginning or in the end of the sentence, features such as neighbouring words will have less information to assert the correct sense of a word.

3.3. Baseline

The baseline for this task was established using the notion of Most Frequent Sense (MFS). MFS baseline is a simplistic approach in which for a set of occurrences containing the ambiguous word, the most frequent sense is assigned.

To calculate the MFS baseline for this set of words, a step was done manually by looking at the occurrences extracted from the corpus and classifying each of them into one of the established senses. The number of occurrences for the establishment of the baseline was set to 100, and they were sorted to avoid any bias.

Word	MFS	Word	MFS
Arquivo	0.69	Foco	0.69
Comissão	0.97	Garantia	0.81
Crédito	0.80	Geração	0.77
Cultura	0.86	Imagem	0.69
Essência	0.78	Regime	1.00
Etiqueta	0.82	Volume	0.68

Table 2: MFS Baseline

For these 12 words, results varied according to each word in a range from 0.68 to 1.00. Two of the words presented a very high baseline result, *regime* and *comissão*. They were disregarded for the final experiments, mainly because of the lack of examples to constitute the minority class. Another reason is that for these cases, when assigning only the most frequent sense the system will already have a very high accuracy, recall and precision, which renders disambiguation virtually useless.

3.4. Inter-Annotator Agreement

WSD is more than a laborious computational challenge of automatically asserting the right sense of a given word in a context. WSD proves to be also a matter of disagreement when it comes to what are the actual senses of the words. Commonly-used dictionaries often present a very high level of polysemy due to very fine sense distinctions.

The idea of inter-annotator agreement is to measure how well native and competent speakers can agree on a given meaning of word in context. Comparison between annotators regarding sense distinction provides information on how difficult established senses are to distinguish.

In this task the number of senses established was restricted to those clearly distinguishable for any Portuguese native or competent speaker, and only two major senses or classes were established. The experiment consisted of providing the description of the senses along with 100 random occurrences of that given word in context extracted from the corpus. These occurrences were given to different Portuguese native speakers, who were asked to assign only one sense. Inter-annotator agreement is usually calculated using the Cohen´s Kappa index:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

In the formula presented above, P(A) represents the proportion of times that the annotators agreed and P(E) represents the proportion of times annotators were likely to agree by chance. In the following table, the scores obtained by Kappa are shown:

Word	Kappa	Word	Kappa
Arquivo	0.627	Foco	0.776
Crédito	0.731	Garantia	0.657
Cultura	0.896	Geração	0.821
Essência	0.836	Imagem	0.552
Etiqueta	0.776	Volume	0.493

Table 3: Kappa Results for Inter-Annotator Agreement

3.5. Feature extraction

For this work, three major groups of features were used: Label Feature, Neighbouring Words and Key Words. The label feature represents the first position in a concordance line. The sentences extracted from the NILC corpus available at Linguateca were retrieved from journalistic texts and some occurrences are indentified by the name of the newspaper section that they belong: Economy, Politics, Sports, etc. There is also a code identifying the specific location of the sentence in the newspaper, and therefore a pattern can be inferred by this information.

The idea to use this feature was inspired by the work of Koeling, McCarthy and Carrol (2007). These researchers claimed that for some domains, the simple presence of an indication of the domain of the text is enough for a classifier to assert the correct class of ambiguous words.

The neighbouring words are features that look at a certain window to the left and to the right of the index word, defined by the range parameter. This feature gives particularly good results when applied to previously processed data removing what is commonly described in the literature as stop words.

The key words feature gives good results for the algorithms. They were extracted after analysis of the data as well as the number of occurrences of the given word in each of the senses. A Boolean value, true or false, was attributed to the presence or absence of the given word in the sentence. There was no standard number of key words used as features for each word. In each case, words were added or subtracted to help improve the results.

4. Evaluation

Evaluation plays an important role to determine the accuracy of any learning method. Machine learning methods need data for training and test. There are several ways of doing it and the most common is to split data into two sets, training set and test set. Although this distribution is commonly used for large datasets, it presents a challenge for smaller datasets and it might lead to problem of representativeness of the training or testing data.

To avoid inaccuracy of results due to data splitting, a statistical technique called cross-validation can be applied. In cross-validation, a fixed number (n) of folds or partitions of the data are assigned, and it is referred to as n-fold cross-validation. For these experiments, the method of n-fold cross validation is used divided in three sets, each set containing 33% of the total data, therefore a three-fold cross validation.

4.1. Metrics

Accuracy is the easiest and most common way of reporting the performance of machine learning methods. However, for some classification tasks, especially those involving highly imbalanced data, more precise metrics should be adopted in order to evaluate results more clearly.

When classifying skewed and highly imbalanced data, accuracy is usually very high and it does not reflect exactly the performance of the classifier. In these cases, evaluation should be concerned with the minority class and assign it as a positive class, and all other classes as a negative class. For these reasons, precision and recall have an important role in the evaluation of classifiers, because they can measure how precise and how complete the classification is on the positive class. For this reason, precision (P) and recall (R) scores are reported along with accuracy. The importance of each of these two measures depends on the nature of the application and usually f-measure is used as a single measure to compare classifiers.

5. Results

5.1. Accuracy Results

For an overview of the best classification results, table 4 presents the best accuracy obtained for each of the ten words along with the respective algorithms. The MFS baseline is presented once again to provide the baseline for the classifiers.

At first glance, the best methods performed above the baseline in terms of accuracy for all the cases, except the word *foco*, for which the results were four per cent below the baseline. As explained in the previous section, accuracy is not the only measure that needs to be taken into account when analyzing the performance of classifiers, especially in unbalanced data such as the one used for these experiments.

Word	**MFS**	**Naive Bayes**	**Maxent**	**Decision Tree**
Arquivo	0.69	0.85	0.79	0.68
Crédito	0.80	0.91	0.96	0.45
Cultura	0.86	0.96	0.99	0.65
Essência	0.78	0.78	0.93	0.85
Etiqueta	0.82	0.93	0.92	0.94
Foco	0.69	0.65	0.65	0.61
Garantia	0.81	0.96	0.97	0.65
Geração	0.77	0.73	0.93	0.69
Imagem	0.69	0.67	0.69	0.72
Volume	0.68	0.81	0.80	0.73
Average	**0.76**	**0.82**	**0.86**	**0.70**

Table 4: Accuracy Results for All Algorithms

Regarding the algorithms, Maximum Entropy performed better in six out of the ten cases and was considered the best classifier for this task in terms of accuracy. Naïve Bayes was the best classifier for three words and its performance for most of the words is very close to the best results that were obtained using Maxent.

Decision Tree was the best method for two words, *etiqueta* and *imagem*, however its overall performance for the ten words is significantly below the other two algorithms and also below the average baseline.

Another important point is that the best accuracy results were obtained for the word *cultura*, which also obtained the best results for the Kappa coefficient on inter-annotator agreement. The second and the third best results for Kappa, *essência* and *geração* also obtained good accuracy results using Maximum Entropy, providing evidence of agreement between annotators and the performance of disambiguation methods.

5.2. Precision, Recall and F-Measure

In order to compare the three methods in terms of precision, recall and f-measure, table 5 presents the results calculated based on the average scores for the ten words disambiguated regarding its minority class:

Method	Precision	Recall	F-Measure
Naive Bayes	0.81	0.77	0.78
Maximum Entropy	0.87	0.70	0.75
Decision Tree	0.75	0.69	0.62

Table 5: Average Precision, Recall and F-Measure.

In the overall results it is once again possible to see how Decision Tree performs below the other two methods, providing another evidence of its inadequacy for the model proposed. Maximum Entropy was the method with the best precision score, 0.87. It was also the method which performed better in terms of overall accuracy, obtaining the best accuracy results in six out of ten words, however it is not the method who presents the highest average f-measure results. The best method in f-measure was Naïve Bayes, presenting also the best numbers in terms of recall.

6. Conclusions

The results presented here constitute an encouraging perspective for other machine learning approaches to WSD as well as other tasks in NLP. This is mainly because the corpus data used for training and testing is untagged (POS tags, syntactic and semantic parsers were not used). Secondly, it is encouraging because the amount of data collected for the experiments was not significantly large as most other applications using machine learning, which proves that it is possible to perform automatic disambiguation using medium sized corpora. This can be particularly useful for resource-poor languages that have fewer linguistic resources available than Portuguese.

Some other conclusions can be drawn from these experiments and they can be applied in further WSD and NLP research. The first is that the assumption that domain information is an important feature to perform automatic disambiguation is true for the set of words and corpora used in this work. Therefore, it corroborates the conclusions of Koeling, McCarthy and Carroll (2007).

Another important point is that the results obtained by the Kappa coefficient on inter-annotator agreement, seems to indicate how well classifiers will perform disambiguation for a given word. When the agreement on word senses is high, it is more likely that the senses will have strong distinctive features that will provide evidence for the algorithms to disambiguate it.

This work was the first step towards the design of a WSD module for REAP.PT. Further experiments should be carried out to show the importance of WSD within the framework of CALL software, not only from a technological perspective, but also from a pedagogical point of view, as in the experiments described by Kulkarni, et. al. (2007).

Acknowledgment

The work here presented was supervised by Dr. Constantin Orasan at the University of Wolverhampton and Dr. Jorge Baptista from the University of Algarve. It was funded by an Erasmus Mundus scholarship offered by the European Union Education and Training Commission, EMMC 2008-0083 at the Erasmus Mundus Masters in NLP & HLT.

References

BAPTISTA, J.; COSTA, N.; GUERRA, J.; ZAMPIERI, M.; CABRAL, M.; MAMEDE, N. (2010) in T.A.S. Pardo et. al (Editors) "P-AWL: Academic Word List for Portuguese", PROPOR2010, LNAI 6001, p. 120-123.

BIRD, S; KLEIN, E; LOPER, E. (2009) Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit, O´Reilly Media.

COLLINS-THOMPSON, K.; CALLAN, J. (2004) Information retrieval for language tutoring: An overview of the REAP project (poster description). Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK.

FLORIAN, R.; CUCERZAN, S.; SCHAFER, C.; YAROWSKI, D. (2002). Combining classifiers for word sense disambiguation. Journal of Natural Language Engineering, 8(4):327–342.

HIRST, G. (1987) Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press.

KOELING, R.; MCCARTHY, D.; CARROLL, J. (2007) Text categorization for improved priors of word meaning. In Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics (Cicling 2007), pages 241.252, Mexico City, Mexico.

KULKARNI, A.; HEILMAN, M.; ESKENAZI, M.; CALLAN, J. (2008) Word Sense Disambiguation for Vocabulary Learning. Ninth International Conference on Intelligent Tutoring Systems.

LESK, M. (1986) Automatic sense disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In: Proceedings of ACM SIGDOC Conference, p. 25-26. Toronto, Canada.

NG, H; LEE, H. (1996) Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In Proceedings of the 34th Meeting of the Association for Computational Linguistics (ACL-96), pages 40-47, Santa Cruz, CA.

MAMEDE, N.; BAPTISTA, J.; VAZ, P.; HAGEGE, C. (2010) Nomenclature of chunks and dependencies in Portuguese XIP grammar (v. 2.1.). Internal Report. Lisboa: L2F/INESD-ID Lisboa.

MARQUEZ, L.; PADRO, L.; RODRIGUEZ, H. (1999) A machine learning approach to POS tagging. Machine Learning, 39(1), 59 – 91.

MARUJO, L. (2009) REAP.PT – REAP em Português, Master Thesis, Instituto Superior Técnico (IST), Lisboa.

SANTOS, D. (2000) "O projecto Processamento Computacional do Português: Balanço e perspectivas", in Volpe Nunes (ed.), Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000) (Atibaia, São Paulo, Brasil, 19 a 22 de Novembro de 2000), pp. 105-113.

SPECIA, L. (2007) Uma abordagem híbrida relacional para a desambiguação lexical de sentido na tradução automática; PhD Thesis.

STEVENSON, M (2003) Word Sense Disambiguation - The case for combinations of Knowledge Sources, CSLI, Stanford California.

STEVENSON, M.; WILKS, Y. (2003) Word Sense Disambiguation in Mitkov (Editor) Oxford Handbook of Computational Linguistics, Oxford University Press, pages 249-265.

YAROWSKY, D. (1996) Homograph Disambiguation in Speech Synthesis. In J. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), Progress in Speech Synthesis. Springer-Verlag, pp. 159-175.

YAROWSKY, D. (1997). Homograph disambiguation in text-to-speech synthesis. In Jan T. H. van Santen, Richard Sproat, Joseph P. Olive, and Julia Hirschberg (Editors) Progress in Speech Synthesis. Springer-Verlag, New York, pp.157-172.