# A Hidden Markov Model Approach to Word Sense Disambiguation

Antonio Molina, Ferran Pla, and Encarna Segarra

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València (Spain)
{amolina,fpla,esegarra}@dsic.upv.es

**Abstract.** Word Sense Disambiguation is still an open problem in Natural Language Processing. It can be formulated as a tagging problem; therefore different POS tagging techniques can be applied to solve it in a direct way. In this work, we propose a supervised approach to Word Sense Disambiguation which is based on Hidden Markov Models and the use of *WordNet*. We evaluated our system on the *Senseval-2* competition, for the *English all-words* task. The performance of our system is in line with the best approaches in this task.

## 1 Introduction

The availability of linguistic resources has made the application of inductive or corpus-based approaches possible in nearly all the Natural Language Processing (NLP) tasks. These methods have been successfully applied to solve different disambiguation problems, such as part-of-speech (POS) tagging, shallow parsing or chunking, prepositional phrase attachment, etc., using different formalisms: Hidden Markov Models (HMM), transformation-based learning, memory-based learning, decision trees, maximum entropy, etc.

A POS tagger attempts to assign the corresponding POS or morpho-syntactical tag to each word in a sentence, taking into account the context in which this word appears. Word Sense Disambiguation (WSD) consists of selecting the semantic sense of a word from all the possible senses given by a dictionary, as well as taking into account the context in which this word appears. Although a WSD problem can be carried out as a POS tagging task, in practice, the former is more difficult and complex than the latter. First, there is no consensus on the concept of sense, and consequently, different semantic tag sets can be defined. In addition, the size of this set is very large compared to the POS tag set and the few available semantic corpora do not have enough annotated data. Second, the modeling of contextual dependencies is more complicated because a large context is generally needed and sometimes the dependencies among different sentences must be known in order to determine the correct sense of a word (or a set of words). Also, the lack of common evaluation criteria makes it very hard to compare different approaches. In this respect, the knowledge base *WordNet* [1]

and the *SemCor* [1] corpus [2] are the most frequently used resources. *Senseval* [2] competition can be viewed as the most important reference point for WSD.

There has been a wide range of approaches to the WSD problem (a detailed study can be found in [3] and [4]). In general, you can categorize them into knowledge-based and corpus-based approaches. Under the knowledge-based approach the disambiguation process is carried out using information from an explicit lexicon or knowledge base. The lexicon may be a machine-readable dictionary, such as the *Longman Dictionary of Contemporary English*, thesaurus, such as *Rodget's Thesaurus*, or large-scale hand-crafted knowledge bases, such as *WordNet* [5–10].

Under the corpus-based approach, the disambiguation process is carried out using information which is estimated from data, rather than taking it directly from an explicit knowledge base. In general, disambiguated corpora are needed to perform the training process, although there are a few approaches which work with raw corpora. Machine learning algorithms have been applied to learn classifiers from corpora in order to perform WSD, that is, algorithms are applied to certain features extracted from the annotated corpus and used to form a representation of each of the senses. This representation can then be applied to new instances in order to disambiguate them [11–13].

In the framework of corpus-based approaches, successful corpus-based approaches to POS tagging which used HMM have been extended in order to be applied to WSD. In [14], they estimated a bigram model of ambiguity classes from the *SemCor* corpus for the task of disambiguating a small set of semantic tags. Bigram models were also used in [15]. The task of sense disambiguating was carried out using the set of synsets of *WordNet* and using the *SemCor* corpus to train and to evaluate the system.

The last edition of the *Senseval* competition has shown that corpus-based approaches achieve better results than knowledge-based ones. Around 20 different systems participated in the *English all-words* task. The three best systems used supervised methods and they achieved a precision which ranked between 61.8% and 69.0%. The system *SMUaw* by Rada Mihalcea achieved the best precision (69.0%). It used a hybrid method which combined different knowledge sources: the *WordNet*, the *SemCor* corpus and a set of heuristics in order to obtain a set of sense-tagged word-word pairs. The second system in the competition (*CNTS-Antwerp*) by Veronique Hoste used a voting strategy in order to combine different learning algorithms, such as memory-based learning and rule induction. It used *SemCor* to train the different classifiers and obtained a precision of 63.6%. The *Sinequa-LIA-HMM* system by E. Crestan, M. El-Beze and C. Loupy achieved a precision of 61.8%. It used a second-order HMM in a two-step strategy. First, it determined the semantic category associated to a word. Then, it assigned the most probable sense acording to the word and the semantic category. In addition,

[1] The *SemCor* corpus and *WordNet* are freely available at http://www.cogsci.princeton.edu/~wn/

[2] Information about the last edition of *Senseval* can be found at http://www.sle.sharp.co.uk/senseval2/

it used an alternative approach based on classification trees for certain words. The next two systems in the ranking by D. Fernández-Amorós (*UNED-AW-U*) used an unsupervised approach obtaining a precision of 55.0% and 56.9%. They constructed a relevance matrix from a large colecction of English books, which was used to filter the context of the words to be disambiguated. The rest of the systems, which are mainly based on unsupervised methods, gave a significantly lower performance than the methods mentioned above. Only a few of them gave a precision which was higher than 50%; however, they had a very low recall.

Some conclusions can be established from all these works: sense disambiguation is a very difficult task and the semantic resources available to perform it are not sufficient. Despite these drawbacks, the good results obtained by learning techniques in other disambiguation tasks encouraged us to present an approach to WSD based on HMM [16]. A similar technique (Specialized HMM), which takes into account certain words to lexicalize the contextual language model, had been previously applied in order to solve POS tagging [17] and chunking [18] problems. In general, lexicalized HMMs performed better than non-lexicalized ones in these tasks. In [16], we presented a preliminary evaluation of our WSD system on the *SemCor* corpus. We present in this paper an evaluation of our system on the *Senseval-2 English all-words* task, in order to test its portability to other tasks.

The paper is organized as follows: in Section 2, we describe the WSD system proposed. In Section 3, we present the experimental work conducted on the *Senseval-2 English all-words* task. Finally, we present some concluding remarks.

## 2 Description of the WSD System

We consider WSD to be a tagging problem which we propose to solve using a HMM formalism. Let $\mathcal{S}$ be the set of sense tags considered, and $\mathcal{W}$, the vocabulary of the application. From this point of view, tagging can be solved as a maximization problem. Given an input sentence, $W = w_1, \ldots, w_T$, where $w_i \in \mathcal{W}$, the tagging process consists of finding the sequence of senses ($S = s_1, \ldots, s_T$, where $s_i \in \mathcal{S}$) of maximum probability on the model, that is:

$$\begin{aligned} \widehat{S} &= \arg\max_S P(S|W) \\ &= \arg\max_S \left( \frac{P(S) \cdot P(W|S)}{P(W)} \right); S \in \mathcal{S}^T \end{aligned} \tag{1}$$

Due to the fact that this maximization process is independent of the input sequence, and taking into account the Markov assumptions for a first-order HMM, the problem is reduced to solving the following equation:

$$\arg\max_S \left( \prod_{i:1\ldots T} P(s_i|s_{i-1}) \cdot P(w_i|s_i) \right) \tag{2}$$

The parameters of equation 2 can be represented as a first-order HMM where each state corresponds to a sense $s_i$, where $P(s_i|s_{i-1})$ represent the transition

probabilities between states and $P(w_i|s_i)$ represent the probability of emission of symbols, $w_i$, in every state, $s_i$. The parameters of this model are estimated by maximum likelihood from semantic annotated corpora using an appropriate smoothing method. Then, the semantic tagging is carried out using the Viterbi algorithm.

Starting from that general tagging scheme, we made certain decisions in order to improve the disambiguation process.

- We used certain resources, such as *WordNet* to know the possible semantic tags associated to the words. In addition, as we will show in the experimental section, we estimated the frequencies of each possible sense for a word from the *SemCor* corpus. This information is also available in *WordNet*.
- We decided which available input information is really relevant to the task. In this respect, we considered a concatenation of the lemma ($l_i$) and the POS[3] ($p_i$) associated to the word ($w_i$) as input vocabulary, if $w_i$ has a sense in *WordNet*. For the words which do not have a sense in *WordNet*, we only consider their lemma ($l_i$) as input. So, in our HMM, $l_i \cdot p_i$ or $l_i$ are the symbols emitted in the states.
  For example, for the input word *interest* which has an entry in *WordNet*, whose lemma is *interest* and whose POS is *NN*, the input considered in our system is *interest·1*. If the word does not have a sense in *WordNet*, such as the article *a*, we consider as input its lemma *a*.
- We defined the output semantic tag set by considering certain statistical information which was extracted from the annotated training set. In the *SemCor* corpus, each annotated word is tagged with a *sense_key* which has the form *lemma%lex_sense*. In general, we considered the *lex_sense* field of the *sense_key* associated to each lemma as the semantic tag in order to reduce the size of the output tag set. This does not lead to any loss of information because we can obtain the *sense_key* by concatenating the lemma to the output tag. For certain frequent lemmas, we considered a more fine-grained semantic tag: the *sense_key* or *synset*. These choices have been made experimentally by taking into account a set of frequent lemmas, $\mathcal{L}_s$, which were extracted from the training set.
  For instance, the input *interest·1* is tagged with the semantic tag *1:09:00::* in the training data set. If we estimate that the lemma *interest* belongs to $\mathcal{L}_s$, then the semantic tag is redefined as *interest·1:09:00::*.
  For the words without semantic information (tagged with the symbol *notag*), we have tested several transformations: to consider their POS in the states, to consider their lemma or to consider only one state for all these words. The approach that achieved the best results consisted of specializing the states with the lemma. For example, for the word *a* the output tag associated is *a·notag*.

The above decisions do not modify either the learning or the decoding process used. To apply them, we performed a transformation on the original training set

[3] We mapped the POS tags to the following tags: 1 for nouns, 2 for verbs, 3 for adjectives and 4 for adverbs.

to produce a new one which included these decisions [18]. As we will show in the experimental results all these decisions improved the performance of our WSD system.

## 3 Experimental Results

A preliminary evaluation of our system was conducted on the *SemCor* corpus [16]. In that case, the precision results (70.39%) were not very satisfactory because our best system (BIGesp) performed much like to the Baseline (we considered a system that assigned the most frequent sense in the *SemCor* corpus given a lemma and its POS as Baseline). A more objective analysis of the performance of our system should be done on other corpora in which the sense associated to a polysemic word does not usually correspond to the most frequent sense in *WordNet*. The contextual language model could the help in the disambiguation process. To do that, we chose the *English all-words* task in the *Senseval-2* competition.

In the experiments, we used *WordNet* 1.6 as a dictionary which supplies all the possible semantic senses for a given word. In addition, our system disambiguated all the polysemic lemmas, that is, the coverage of our system was 100% (therefore, precision and recall were the same).

We compared the specialized models with respect to non-specialized ones. The basic unigram (UNI) and bigram (BIG) models are non-specialized models which take into account an input vocabulary that only consists of lemmas. UNIpos and BIGpos are also non-specialized models whose input vocabulary consists of the lemma and the POS. UNIesp and BIGesp are models which are specialized as we mentioned in Section 2.

To build the specialized models, we previously selected the set of lemmas $\mathcal{L}_s$. The specialization criterion consisted of selecting the lemmas whose frequency in the training data set was higher than a certain threshold (other specialization criteria could have been chosen, but frequency criterion usually worked well in other tasks as we reported in [18]). In order to determine which threshold maximized the performance of the model, we conducted a tuning experiment on a development partition of *SemCor* (10% of the corpus). For the BIGesp model, the best performance was obtained using a threshold of 20 ($|\mathcal{L}_s|$ was about 1,600 lemmas).

The *Senseval-2* competition did not provide any training corpora for this task, so we used the files contained in the Brown1 and Brown2 folders of the *SemCor* corpus as training data. The test data set provided by *Senseval-2* consisted of three Penn TreeBank documents which contained 2,473 sense-tagged words. POS information was extracted directly from the corresponding Penn TreeBank documents. For unknown words (words that did not appear in the training data set), we assigned the first sense in *WordNet*.

The results for the *English all-words* task are shown in Table 1. The UNIpos and BIGpos models improved the performance of the basic models (UNI and BIG), showing that the POS information is important in differentiating among

| Model | Precision |
|---|---|
| UNI | 40.00% |
| UNIpos | 52.30% |
| UNIesp | 58.80% |
| BIG | 50.10% |
| BIGpos | 58.20% |
| BIGesp | 60.20% |

**Table 1.** Precision results for the *English all-words* task in *Senseval-2*. 100% of the words were tagged.

the different senses of a word. In addition, both Specialized models (UNIesp and BIGesp) outperformed the non-specialized ones. The best performance was achieved by the Specialized Bigram model (BIGesp) with a precision of 60.20%. This result is in line with the results provided for the best systems in *Senseval-2*. This comfirms that Specialized HMMs can be applied to WSD, as successfully as they have been applied to other disambiguation tasks. The most similar approach to our system (Sinequa-LIA-HMM) achieved a result which was slightly better (61.8%), but it combined the HMM model with a classification tree method to disambiguate some selected words.

## 4 Conclusions

In this paper, we have proposed a word sense disambiguation system which is based on HMM and the use of *WordNet*. We made several versions of our WSD system. Firstly, we applied classic unigram and bigram models and, as we had expected, the bigram model outperformed the unigram model. This is because the bigram model better captures the context of the word to be disambiguated. Secondly, we incorporated POS information to the input vocabulary which improved the performance and showed the relevance of this information in WSD. Finally, we specialized both the unigram and the bigram models in order to incorporate some relevant knowledge to the system. Again, as we expected, specialized models improved the results of the non-specialized ones.

From the above experimentation, we conclude that the BIGesp model is the best model. This model gave a precision of 60.20% in the *English all-words* task of the *Senseval-2* competition, which was only outperformed by the three best systems for the same task. This is a good result taking into account that our WSD system is mainly an adaptation of our POS tagging and chunking systems. This adaptation consists of an appropriate definition of the relevant input information and the output tag set. In addition, the portability shown by our system shows that it would also give good results in other languages.

Finally, we think that we could improve our WSD system through a more adequate selection of the set of lemmas which specialize the model. To do this, a

development data set would be necessary in order to tune the specialized model to the task.

## 5 Acknowledgments

This work has been supported by the Spanish research projects CICYT TIC2000–0664–C02–01 and TIC2000–1599–C01–01.

## References

1. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.J.: WordNet: An on-line lexical database. International Journal of Lexicography **3** (1990) 235–244
2. Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a Semantic Concordance for Sense Identificaction. In: Proceedings of the ARPA Workshop on Human Language Technology. (1994) 240–243
3. Ide, N., Véronis, J.: Word Sense Disambiguation: The State of the Art. Computational Linguistics **24** (1998) 1–40
4. Resnik, P., Yarowsky, D.: Distinguishing systems and distinguishing senses: new evaluation methods for Word Sense Disambiguation. Natural Language Engineering **6** (2000) 113–133
5. Lesk, M.: Automated Sense Disambiguation using Machine-readable Dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 1986 SIGDOC Conference, Toronto, Canada (1986) 24–26
6. Yarowsky, D.: Word-sense Disambiguations Using Statistical Models of Roget's Categories Trained on Large Corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, COLING, Nantes, France (1992) 454–460
7. Voorhees, E.: Using WordNet to Disambiguate Word Senses for Text Retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh (1993) 171–180
8. Resnik, P.S.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI, Montreal, Canada (1995) 448–453
9. Agirre, E., Rigau, G.: Word Sense Disambiguation Using Conceptual Density. In: Proceedings of the 16th International Conference on Computational Linguistics, COLING, Copenhagen, Denmark (1996)
10. Stevenson, M., Wilks, Y.: The Interaction of Knowledge Sources in Word Sense Disambiguation. Computational Linguistics **27** (2001) 321–349
11. Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM, ACL (1994) 88–95
12. Ng, H.T.: Exemplar-Base Word Sense Disambiguation: Some Recent Improvements. In: Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, EMNLP. (1997)
13. Escudero, G., Márquez, L., Rigau, G.: A comparison between supervised learning algorithms for Word Sense Disambiguation. In: Proceedings of CoNLL-2000 and LLL-2000, Lisbon, Portugal (2000)

14. Segond, F., Schiller, A., Grefenstette, G., Chanod, J.P.: An Experiment in Semantic Tagging using Hidden Markov Model Tagging. In: Proceedings of the Joint ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources, Madrid, Spain (1997) 78–81
15. Loupy, C., El-Beze, M., Marteau, P.F.: Word Sense Disambiguation using HMM Tagger. In: Proceedings of the 1st International Conference on Language Resources and Evaluation, LREC, Granada, Spain (1998) 1255–1258
16. Molina, A., Pla, F., Segarra, E., Moreno, L.: Word Sense Disambiguation using Statistical Models and WordNet. In: Proceedings of 3rd International Conference on Language Resources and Evaluation, LREC2002, Las Palmas de Gran Canaria, Spain (to be published) (2002)
17. Pla, F., Molina, A.: Part-of-Speech Tagging with Lexicalized HMM. In: proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP2001), Tzigov Chark, Bulgaria (2001)
18. Molina, A., Pla, F.: Shallow Parsing using Specialized HMMs. Journal of Machine Learning Research **2** (2002) 595–613