Actes de The First International Conference on Language Resources & Evaluation
pp. 1255-1258; Grenade, Spain; 28-30 Mai 1998

Word Sense Disambiguation using HMM Tagger

Claude de Loupy (1) & (2)**, Marc El-Beze** (1) **& Pierre-François Marteau** (2)

(1)
Laboratoire Informatique d'Avignon
B.P. 1228 Agroparc
339 chemin des Meinajaries
84911 Avignon cedex 9

(2)
Bertin & Cie
Z.I. des Gatines - B.P. 3
78373 Plaisir cedex

Abstract

In Natural Language Processing, it is necessary to take into account the polysemy of some words. In all the domains of language management, stochastic methods have shown their efficiency. To what extent is it possible to use probabilities for word sense disambiguation? What are the available resources and do they fit such a task?

In this article, we evaluate the difficulty of a semantic disambiguation task and present different statistic methods for sense assignation, based on WordNet and using the SemCor to train and test the models.

Introduction

Word sense disambiguation is a very important task for Natural Language Processing. In the information retrieval framework, both precision and recall could be improved by means of a word sense disambiguation component (Gilarranz et al., 1996).
Stochastic systems have shown their efficiency when applied on several problems like speech recognition, spell checking, syntactic tagging (El-Bèze & Mérialdo, 1997). Nevertheless, there is still some uncertainty on how suitable a probabilistic model is for sense disambiguation?
However, few resources are available to perform a semantic tagging. Among others, WordNet (Miller, 1990), freely available on the Web provides semantic information associated with lemmas, and a semantically tagged subset of the Brown Corpus, called SemCor (Miller et al., 1993). These resources could be used to evaluate various approaches and models dedicated to semantic tagging.
Over the last few years, a lot of new tools for semantic disambiguation have appeared in the literature. Some systems use WordNet entries for a semantic classification of words (Resnik, 1995), (Rigeau & Agirre, 1995), (Li et al., 1995), etc. But most of them use a conceptual distance and are not based on Markov models.
The work presented in this paper is an initial step tp evaluate the capability of a stochastic tagger to automatically learn, from WordNet and the SemCor, how to tag each lemma with the most likely sense. Preliminary results are reported, along with some criteria proposed to estimate the task difficulty.

Difficulty of the Task

Resources Used

The experiments presented in this paper rely on the WordNet[1] resource (release 1.5) which contains more than 126 500 entries with semantic information.
We have evaluated the coverage of that lexicon on the Gutenberg[2] project corpus (Hart, 95). The static[3] coverage rate on the entire corpus is 96.5 % and the dynamic one is 99.8 %. The dynamic coverage is very high while the static one is lower than usual[4] because a lot of texts are technical or poetic. Most of them contain many proper nouns. On average, the static coverage is 96.7 % by text. It is strange but interesting to note that the text with the best coverage is Descartes' *Reason Discourse* and the worst is *Zen and the Art of Internet* (in which the three most frequent unknown words are *newsgroup*, *archie* and *telnet*). This reflects that Out Of Vocabulary words are strongly integrated into the language as soon as they appear, which leads us to develop adaptive lexicons and models.

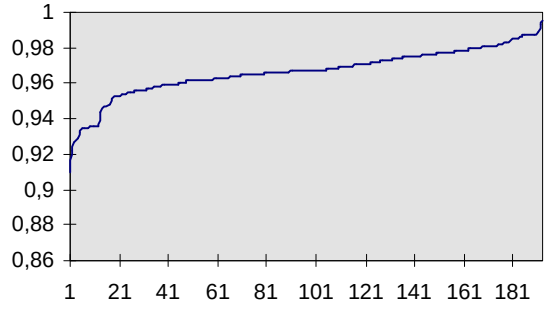

Fig. 1: Static coverage of WordNet on Gutenberg

[1] The release 1.6 of WordNet is available at http://www.cogsci.princeton.edu/~wn.
[2] Project Gutenberg places electronic texts at the community's disposal. By April 1998, more than a thousand texts are available at http://promo.net/pg.
[3] When evaluating a static (resp. dynamic) coverage, each occurrence (resp. only the first occurrence) of an unknown form is taken into account.
[4] With a 20 000 entries lexicon, the static coverage is generally 98 % (Jelinek, 1990)

In Actes de The First International Conference on Language Resources & Evaluation
pp. 1255-1258; Grenade, Spain; 28-30 Mai 1998

To train the parameters of the proposed semantic models, we have used the SemCor. This corpus contains more than 11 000 sentences and about 106 000 semantically tagged words.

How Difficult the Semantic Tagging Task is?

We want to associate a precise sense to each lemma of a text. The difficulty of the task is measured using the ratio between the total number of possible senses and the total number of lemma occurrences. There are 126 520 lemmas and 168 135 different sense units in WordNet 1.5. Consequently, there is, on average, about 1.3 sense per lemma. A sense unit is made of the association between a lemma and a synonym set (or synset). Therefore, we will use "sense" or "synset" indifferently from now on. There are more than 91 000 synsets in WordNet.
But, when estimating on the SemCor, we find 5.1 sense units per lemma. This is because most of the frequent words are polysemous and, even if 48 % of the lemmas used in the SemCor have a single sense, they represent only 22 % of all the occurrences. The difficulty comes from words like *run*, which has 41 different senses and appears 59 times or *have* which appears 595 times and has 21 senses.
Another figure may illustrate how difficult the assignation is. The accuracy for a random tagging is 43 %. One must note how difficult the task is, compared to the 75 % of good assignation for a random syntactic tagging (El-Bèze & Mérialdo, 1997).

Training and Test Corpora

To estimate the parameters of the model, we used 95 % of the SemCor, that is the 10 578 first sentences (about 100 000 words) as a bootstrap. The last 5 % (more than 5 000 words) were used to evaluate the model quality.
65.6 % of the lemmas used in the test corpus are ambiguous. They represent almost 76 % of the lemma occurrences.
Since the Brown Corpus is a collection of different texts, this distribution is consistent with a real task; the test corpus is a different kind of text than those previously observed. This would not have been the same if, for example, we had taken the odd sentences or paragraphs for training and the even ones for test. In that case, there would not be so many discrepancies between test and training corpora in terms of topic.
We didn't use a Part-Of-Speech tagger in this experiment. We based our models on the lemmas and their associated grammatical tags within the SemCor. Of course, this implies an easier task and better results. Nevertheless, the error rate of POS tagging is very low, and the figures shown in this article should be very close to actual results.

Disambiguating Sense Unit

A Very Simple Model

When looking at the possible senses of a word, they have not, in general, the same probability of occurrence. With an already tagged corpus such as SemCor, one may choose to apply the following decision rule: assign to each lemma its most probable sense (without looking at the context).
In that case, a more realistic measure of the task difficulty for semantic tagging is given by the accuracy rate of a system assigning the most probable sense to each lemma.
With such a simple model (so-called unisem model), the accuracy has been increased to 71.5 %.
This figure seems very low, compared with the unisem efficiency of a part-of-speech assignation: 91.7 %.
In fact, this is not a very surprising result. The information on words is very poor in the SemCor; only 83 words occur more than 100 times in this corpus (Kilgarriff, 1997). 13.2 % of the occurrences in the test part of the corpus concern lemmas which do not appear in the bootstrap.

A More Sophisticated Model

In order to improve this result, we have experimented a bisem model within a Hidden Markov Model. The tags used are synsets.
The probability of a synset σ for lemma λ_i in a given context H is calculated by the following approximation:

$$P(\sigma_{i,j}|H) \approx \left(\alpha . P(\sigma_{i,j}|\sigma_{i-1,k}) + (1-\alpha) . P(\sigma_{i,j})\right) . P(\lambda_i|\sigma_{i,j}) \quad (1)$$

Using the Viterbi (Viterbi, 1967) algorithm, we found 28 % discrepancies between the result and the SemCor with a combination of unisem and bisem (α=0.1). This represents thus only a 0.5 % improvement compared to the unisem model. A pure bisem model did not lead to better results than the unisem one.
There are different explanations for this rather disappointing result. The SemCor contains 90 784 different bisems. This seems to be very poor, when referring to the number of possible pairs : $(91\,591)^2 \approx 8.4\ 10^9$. Moreover, for more than 57 % of the occurrences of the bisems in the SemCor, the two corresponding lemmas are ambiguous.
Lastly, less than 7 % of the bisems occurred more than once. Spurious effects are observed when HMMs are undertrained (Gale & Church, 1990). If a synset σ occurred only once in the training corpus, necessarily no more than one synset σ' has been observed after σ. It is thus impossible to predict a sense σ'' since the pair (σ,σ'') never occurred whatever $\sigma'' \neq \sigma'$. That part of the model is clearly not probabilistic anymore.
In order to obtain better reliability during the training phases, it will be important to enlarge the training corpus with other resources. Nevertheless, a combination of bisem and unisem (α = 0.1) slightly improves the results.

Using WordNet Category

The Category Assignation

(Segond et al., 1997) have proposed a model based on 45 semantic categories given by WordNet.
For example, a noun can be referenced as an animal, another as a body part, etc. There are 26 categories for nouns, 15 for verbs, 3 for adjectives and 1 for adverbs.
In the SemCor, there is an average of 2.4 categories per lemma (compared to 5.1 for synsets) and both categories and bicategories are well represented. 20.4 % of the lemmas used in the test corpus are ambiguous for the category, and they represent 50.4 % of the occurrences.
A random tagging gives 66.6 % of good assignation.

In Actes de The First International Conference on Language Resources & Evaluation
pp. 1255-1258; Grenade, Spain; 28-30 Mai 1998

When applying a formulation similar to (1) on categories, we found better results (86.3 %) than those obtained with synsets[5].
But we have to examine to what extent a precise sense can be derived from a given lemma, λ, and a given category c.

Using Categories to Assign Synsets

The problem is that it is not possible to assign a sense unit to a given pair (λ, c) in a deterministic way. 51 % of the lemmas used in the test corpus have, at least, one associated category c such as (λ, c) is associated with more than one synset. These lemmas represent almost 59 % of the occurrences.
In spite of these figures, better results for synset allocation could be expected using a model based on both general category and precise sense.
We have chosen to estimate the upper bound of the score that could be obtained with such a combination by using the category given with the lemma in the SemCor.
In that way, there is no more ambiguity on the category, and the search is limited to the best synset associated to a couple (lemma, category).
The Viterbi algorithm gives 83.2 % of accuracy with the SemCor. This is really interesting compared with the 72 % we found with a synset-based model.
If we try to first assign a category and then assign a synset (both with Viterbi algorithm), the result is worse than when applying only a synset based model (71 % for the best score).

A Model Combining Synsets and Categories

We are now going to check whether a model combining a category and a synset assignation at the same moment could improve the score. For that purpose we use the following model to assign a couple ($\sigma_{i,j}$, $c_{i,k}$) to a given lemma λ_i, in a particular context H.

$$P(\sigma_{i,j}, c_{i,k} | H) \approx P(\lambda_i | \sigma_{i,j}, c_{i,k}).P(\sigma_{i,j} | c_{i,k}). \left(\alpha.P(c_{i,k} | c_{i-1}) + (1-\alpha).P(c_{i,k})\right) \quad (2)$$

This model leads to a slight improvement of the result (71.8 % of accuracy) with $\alpha = 0$ (unisem). This is not really surprising since 83.2 % (correct assignation of a synset, given a lemma and a category) of 86.3 % (correct assignation of a category, given a lemma) gives 71.8 %. This shows the consistency of the model.

Pattern Recognition

Principle

Another way to have better assignation is to look more precisely at the only units that are known with certainty: the lemmas. We made the following assumption: when lemma series are observed in the training part, and appear in the text to tag, it is more probable that the senses of the series are the same. The longer the series, the higher the likelihood that its sense will be identical to the one of the same series observed in another context.

[5] It is difficult to compare our results with the ones published by (Segond et al., 1997) because they did not use exactly the same method, nor the same split between test and training corpus.

It is interesting to note that this leads to a paradox, because the longer the series, the lower its number of occurrences. Therefore, the reliability of this series should be weaker, according to the usual probabilistic evidence.
Besides, humans themselves use such pattern recognition mechanisms. It is easier for them to disambiguate word groups than single words.

Applying to the SemCor

The training and test parts of the SemCor have 682 series in common. The 3 longest ones are composed of 4 lemmas, and only 59 of 3 lemmas. Therefore, this model is mainly a bigram one.
The algorithm is the same as for the synset disambiguation, except that, when known series are observed, the corresponding senses get a bonus.
This leads to a slight improvement: 72.2 % of accuracy. 600 different rules were applied on 1 234 lemmas. Compared to the synset model, 25 new errors were added, and 38 corrected.
By observing the results, we found that most of the errors were made when the series contained the verb "to be". So, we decided to eliminate all the series containing this verb. 436 different rules were applied on 886 lemmas. 14 errors were added and 32 corrected. Therefore, the percentage of accuracy is 72.3 %.
This improvement, though weak, emphasizes the importance of working on a corpus bigger than the SemCor for training.

Managing Several Senses

One Sense is Not Enough

By keeping only the best tag, alternate ones are rejected even if they have almost the same score. Secondly, even if a word is used in a particular sense, its other senses are always in the mind of the person who says this word and the one who hears it. For example, in "Mayor_Hartsfield announced that he would not run for reelection", two or three among the 41 possible senses of "run" are kept in mind.
Therefore, since the goal of this work is to improve Information Retrieval, it would be risky to keep only the most probable sense of a lemma. We have to order the possible senses of a lemma according to its context, and keep as much senses as we need.

Using the Baum-Belch Algorithm

In order to achieve such a result, we used the Baum-Welch algorithm (Baum et al., 1970), (Cutting et al., 1992).
The following table gives the percentage of assignations in which the SemCor semantic tag appears, according to the number of senses kept:

1 sense	2 senses	3 senses
71.7 %	86.4 %	92.3 %

We can see that, when we keep 3 senses (if the considered lemma has at least 3 senses), the sense given in the SemCor is very often in the list of senses chosen by the model.

In Actes de The First International Conference on Language Resources & Evaluation
pp. 1255-1258; Grenade, Spain; 28-30 Mai 1998

Conclusion

Sense disambiguation is a very difficult task and semantic resources to perform it are not sufficient. We implemented different stochastic methods to perform a word sense assignation.

In fact, according to (SemCor, 1995), the error rate for polysemous words in the SemCor is estimated at 13 %. Moreover, (Ng & Lee, 1996) put forward a much worse rate.

The NLP community needs more semantic resources, and particularly, semantically tagged corpus with a high level of confidence for several languages[6]. The SENSEVAL[7] and ROMANSEVAL[8] projects will perhaps boost the production of such resources.

Finally, on one hand, synsets in WordNet are too fine-grained. Are the 41 distinctions for "run" really useful for Information Retrieval? On the other hand, WordNet categories do not provide enough information. It is necessary to have nested senses (Kilgarriff, 1997), so that different levels of analyzes are possible.

References

Baum L.E., Petrie T., Soules G. & Weiss N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. AMS Vol. 41, No. 1 (pp 167-171).

Cutting D., Kupiec J., Pedersen J. & Sibun P. (1992). A Practical Part-of-Speech Tagger. In *Third Conference on Applied Natural Language Processing* (pp 133-140), Trento.

El-Bèze M. & Mérialdo B. (1997). HMM Based Taggers. In *Syntactic Wordclass Tagging*. Edited by H. Van Halteren, Kluwer Academic Publishers.

Gale W. A. & Church K. W. (1990). Poor Estimates of Context are Worse than None. In *Proceedings of DARPA Speech and Natural Language Workshop USA* (pp 283-287). San Mateo, Cla.

Gilarranz J., Gonzalo J. & Verdejo F. (1996). An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database. In *Proceedings of AAAI-96 Spring Symposium Cross-Language Text and Speech Retrieval*.

Hart M. S. (1995). History and Philosophy of Project Gutenberg. http://gutenberg.etext.org/pl/history.html.

Jelinek F. (1990). Self-Organized Language Modeling for Speech Recognition. *In Readings in Speech Recognition*. A. Waibel & K. F. Lee Ed.; Morgan Kaufman Publishers (p 468).

Kilgarriff A. (1997). Evaluating Word Sense Disambiguation Programs: Progress Report. In *Proceedings SALT Workshop on Evaluation in Speech and Language Technology*. Sheffiel, U.K.

Miller G. (1990). Five Papers on WordNet. In *Special Issue of International Journal of Lexicography* 3(4).

Miller G., Leacock C., Randee T., Bunker R. (1993). A Semantic Concordance. In *Procedings of the 3rd DARPA Workshop on Human Language Technology* (pp 303-308). Plainsboro, New Jersey.

Ng H. T. & Lee H. B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense : an Exemplar-Based Approach. In *ACL Proceedings*.

Li X., Szpakowicz S. & Matwin S. (1995). A WordNet-Based Algorithm for Word Sense Disambiguation. In *Proceedings of IJCAI-95*. Montréal, Canada.

Rigau G. & Agirre E. (1995). Disambiguating Bilingual Nominal Entries Against WordNet. *Workshop on the Computational Lexicon, ESSLI'95*.

Resnik P. (1995). Disambiguating Noun Grouping With Respect to WordNet Senses. In *Proceedings of the 3rd Workshop on Very Large Corpora*; M.I.T.

Segond F., Schiller A., Grefenstette G. & Chanod J.-P. (1997). An Experiment in Semantic Tagging Using Hidden Markov Model Tagging. In *Proceedings of EACL'97* (pp 78-81). Madrid.

SemCor man pages (1995). SemCor - Discussion of Semantic Concordance of Semantically Tagged Text. *http://www.cosgi.princeton.edu/~wn/man/semcor.7WN.html*.

Viterbi A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Transactions on Information Theory, IT-13* (pp 260-269).

[6] The LIA and Bertin are involved in the EWN project extension for the creation of a French semantic wordnet :
http:// www.let.uva.nl/~ewn

[7] SENSEVAL is a project aiming to evaluate word sense disambiguation tools :
http://www.itri.brighton.ac.uk/events/senseval

[8] ROMANSEVAL is a part of SENSEVAL project dedicated to French and Italian languages :
http://www.lpl.univ-aix.fr/projects/romanseval