Computers and the Humanities **34:** 187–192, 2000.

Using Semantic Classification Trees for WSD

C. de LOUPY[1,2], M. EL-BÈZE[1] and P.-F. MARTEAU[2]
[1]*Laboratoire d'Informatique d'Avignon (LIA), BP 1228, F-84911 Avignon Cedex 9 France (E-mail: {claude.loupy,marc.elbeze}@lia.univ-avignon.fr);* [2]*Bertin Technologies, Z.I des Gatines – B.P. 3, F-78373 Plaisir cedex (E-mail: {deloupy,marteau}@bertin.fr)*

Abstract. This paper describes the evaluation of a WSD method within SENSEVAL. This method is based on Semantic Classification Trees (SCTs) and short context dependencies between nouns and verbs. The training procedure creates a binary tree for each word to be disambiguated. SCTs are easy to implement and yield some promising results. The integration of linguistic knowledge could lead to substantial improvement.

Key words: semantic classification trees, SENSEVAL, word sense disambiguation, WSD evaluation

1. Introduction

While developing a set of Information Retrieval components (de Loupy et al., 1998a), the Laboratoire Informatique d'Avignon (LIA) and Bertin Technologies are investigating semantic disambiguation. In a Document Retrieval framework, identifying the senses associated with the words of a query is expected to lead to some noise reduction for short queries (Krovetz and Croft, 1992). As a second benefit, this knowledge should also result in an increase in recall through query expansion relying on synonymy and other semantic links. In de Loupy et al. (1998d), we experimented with this type of enrichment using WordNet (Miller et al., 1993). Performance was improved when words having a single sense (two if they are not frequent words) were associated with their synonyms.

In de Loupy et al. (1998b), we evaluated a first approach based on WordNet, the SemCor (Miller et al., 1993b) and Bisem Hidden Markov Models. These models are not so well adapted to this task for 2 reasons: (i) the context window is too small (2 words), (ii) a very large amount of training corpus (so far unavailable) is required. Semantic Classification Trees (SCT) (Kuhn and de Mori, 1995) offer an alternative to model right and left contexts jointly. Smaller learning resources are required. Short context dependencies are taken into account. We could have used a pure knowledge-based WSD system. But extending such a system to a large scale requires writing rules for each word. The SCT approach can be seen as an attractive trade-off because it allows building an automatic WSD system without excluding the possibility of introducing knowledge.

2. Preparation of the Data

The SCT method, which requires a training corpus, is well-suited to bring out relevant dependencies between the word to be disambiguated and the words (or types of words) surrounding it. As a first step, we have only attempted to tag nouns and verbs (adjectives have not been tested). More precisely, the evaluation of the proposed approach has been performed on 25 different words (see section 4 for the list). In order to train the models, we have used the examples given by DIC[1] (24 examples per word on average) and TRAIN (315 examples per word on average).

"Yarowsky [...] suggests that local ambiguities need only a window of size *k = 3* or *4*, while semantic or topic-based ambiguities require a larger window of 20–50 words" (Ide and Véronis, 1998). Therefore, we have limited the context window size to 3 lemmas before the ambiguous one (call it Λ) and 3 lemmas after. If two possible semantic tags are given for Λ in the learning sample, the information is duplicated to produce one example for each tag. The examples found in DIC and those extracted from TRAIN have been processed exactly in the same way and have the same weight for training.

In order to achieve better WSD, it is important (Dowding et al., 1994; Segond et al., 1993b) to take the grammatical tags of the words into account. For such a task, we have used our ECSTA tagger[2] (Spriet and El-Bèze, 1997). Yarowsky (1993) highlights various behaviors based on syntactic categories: directly adjacent adjectives or nouns best disambiguate nouns. Our assumption is quite different; we would like to check to what extent verbs and nouns could disambiguate nouns and verbs.[3] The words belonging to the three following grammatical classes are therefore not kept for the disambiguation process: determiners, adverbs and adjectives. The other words are replaced by their lemmas and unknown words are left unchanged.

Some words are so strongly related that, in almost all the cases, it is possible to replace one of them by another without any consequence for the sense of the sentence. For instance, it is not necessary to keep precise information on months. Hence, January, February, etc. are replaced by MONTH. In the same way, pseudo-lemma DAY stands for Monday, etc., CD for a number, PRP for a pronoun, NNPL for a location (Paris, etc.), NNPG for a group (a company, etc.), NNP for the name of a person and UNK for an Out of Vocabulary Word if its initial letter is an uppercase.

These substitutions are intended to decrease the variability of the context in which a given word sense can be found. For example, in the definition of *sack*, sense ***504767*** ("the pillaging of a city") is given with the example: *the horrors that accompany the sack of cities?*. This sentence is used to produce the following context example of: / *horror* / *that* / *accompany* / **sack** (***504767***) / *of* / *city* / *?* /.

3. Semantic Classification Trees

A very short description of the SCT method is provided hereafter. For more information, one can refer to Kuhn and de Mori (1995). An SCT is a specialized classification tree that learns semantic rules from training data. Each node T of the

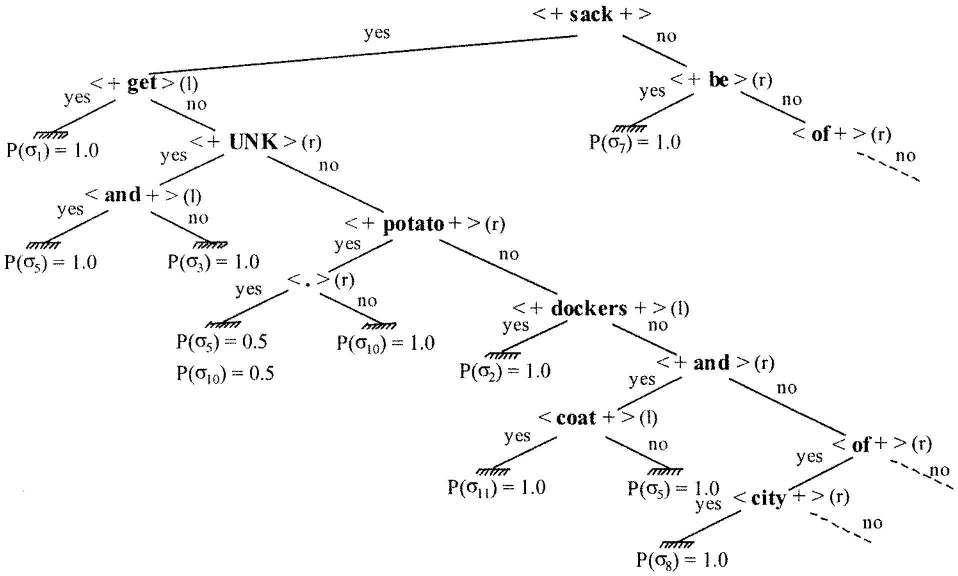

Figure 1. An extract of the SCT for the noun *sack*.

binary tree contains a question that admits a “*Yes/No*” answer corresponding to the two branches attached to the node. The preprocessing procedure described in the previous section produces a set of learning samples. A set of questions[4] corresponding to each sample is then constructed from the words found in the context of the word to be disambiguated. A quantity called Gini impurity (Breiman et al., 1984) is computed in order to choose the most appropriate question for a given *T* node. Let *S* be the set of semantic tags associated with the word to be disambiguated. The Gini impurity is given by $i(T) = \sum_{j \in S} \sum_{k \in S, k \neq j} p(j|T) \times p(k|T)$ where $p(j|T)$ is the probability of sense *j* given node *T*. For each node, the chosen question is the one which leads to a maximal decrease in impurity from the current node to its children, i.e., the one maximizing the *change in impurity*: $\Delta i = I(T) - p_y \times i(Yes) - p_n \times I(No)$ where p_y and pn are the proportions of items respectively sent to *Yes* and *No* by the question.

For instance, let us consider the SCT represented in Figure 1 which has been created for the noun *sack*.[5] Twelve senses are possible for sack-n. Symbols ‘<’ and ‘>’ mark the boundaries of a pattern. ‘+’ indicates a gap of at least one word. For example, < + *sack* + *potato* . > models all the *sack* contexts for which *sack* is not the first word of the context, one or several words follow, then *potato* and a period occur. The sense assigned to *sack* for this context is σ_5”, that is 504756 (“a plain strong bag”), or σ_{10}, that is 505121 (“sack of potatoes”, “something big, inert, clumsy, etc.”). The example given in section 2 gives the rule < + sack of city + > and corresponds to sense σ_8, that is 504767 (“the pillaging of a city”). A linguist would not have used the same questions as the ones found automatically by the system. However, the score obtained for *sack-n* is good: 90.2% of correct

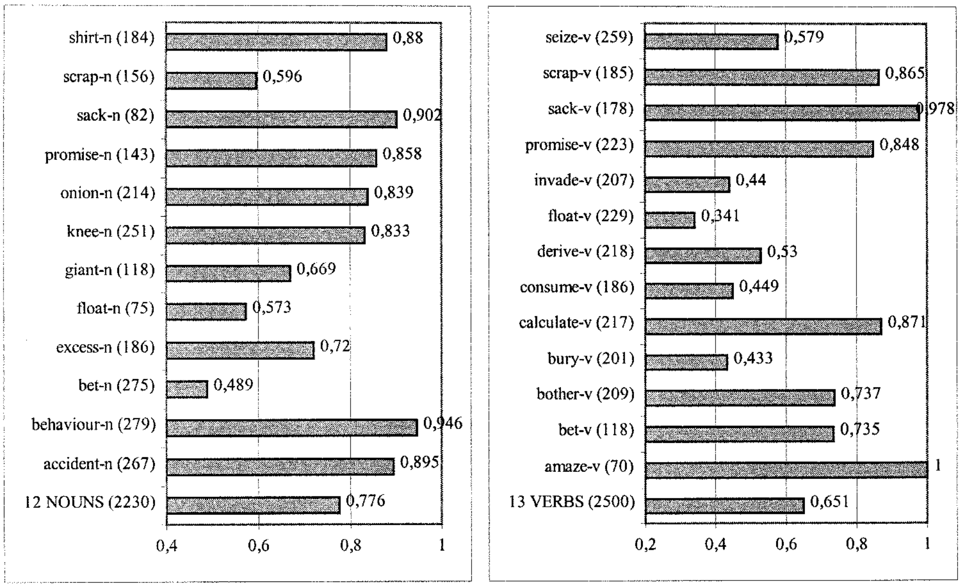

Figure 2. Score of the SCTs for the 25 words (the number of tests per word is given in parentheses).

assignment (the score of a systematic assignation of the most frequent tag being 50.4%).

4. Evaluation of SCTs in SENSEVAL

Within SENSEVAL, the SCT method has been used for semantic tag assignment. The results obtained for the 25 words with *fine-grained* semantic tagging (high precision) are reported[6] in Figure 2. One could argue the most important thing for training is the number of examples for each sense-word pair. Indeed, the best scores are obtained for *behaviour-n* and *amaze-v* for which there is a large number of samples (335 and 139 per sense, respectively). This is not the only explanation: *scrap-v* (13 samples per sense) has better results than *promise-v* (200 samples per sense) and *derive-v* (47 samples per sense). Since *scrap-v* has only 3 semantic tags, the task is obviously easier than for *float-v* (16 tags, 15 samples per sense). Lastly, the task for *amaze-v* is the easiest since there is only one sense! Like other systems tested in SENSEVAL, performance is, on average, better for nouns than for verbs.

It is difficult to compare the experiments carried out with the SCT method and with the HMM model described in de Loupy et al. (1998b) since training and test corpora are different. Moreover, the task described in de Loupy et al. (1998b) requires assignment of semantic tags to each word of the Brown corpus.

5. Conclusion

The approach described in this article has yielded some interesting results. Had we used more sophisticated questions when building the SCTs, results could have been better. Moreover, since little data is given for each semantic tag, we have used low thresholds in order to build wider trees.[7] Therefore, some rules are too specific and do not reach the generalization objective.

Other methods have been tested, leading to the conclusion that SCTs perform better than alternative approaches presented in de Loupy et al. (1998c) (0.51 precision for the other two methods on nouns). Further experiments are necessary in order to assess this result with more reliability. This method is a numerical one and requires no expertise. Nevertheless, linguistic knowledge could be integrated into the whole process, particularly when drawing up the list of questions. For example, the following word is often a good way to determine the sense of a verb (ex: *look around, look for, look after, . . .*).

Moreover, the LIA is developing a French semantic lexicon within the framework of the EuroWordNet project (Ide et al., 1998) and intends, with the support of its industrial partner Bertin Technologies, to use it in a cross-language Document Retrieval frame. Future research will be focused on this topic.

Acknowledgements

We are indebted to Frédéric Béchet, Renato De Mori and Roland Kuhn for their help on the implementation of the SCT method.

Notes

[1] DIC and TRAIN are used here as in SENSEVAL to abbreviate dictionary and training corpus.

[2] ECSTA was evaluated for French in Spriet and El-Bèze (1997), but we do not have a real estimate of its performance for English.

[3] Within the SENSEVAL evaluation, we found that using nouns and verbs to disambiguate nouns improved the effectiveness from 6 to 34% compared to the use of adjectives and nouns, except for 3 nouns for which scores are similar (11.5% improvement on average). For the verbs it is not so clear since the average improvement is less than 2%.

[4] Questions are formulated as regular expressions. An example is given in the following paragraph.

[5] The noun *sack* is a better illustration of the SCT method than *onion*. The SCT for *onion* can be found in de Loupy et al. (1998c).

[6] The SCTs always make a decision. Therefore, precision and recall are the same.

[7] The use of high thresholds would lead to building very poor trees and even, with a very high threshold, reduce to a single node (the root) so that the most frequent tag would be systematically assigned.

References

Breiman, L., J. Friedman, R. Olshen and C. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

Dowding, J., R. Moore, F. Andry and D. Moran. *Interleaving Syntax and Semantics in an Efficient Bottom-up Parser*. ACL-94, Las Cruces, New Mexico, 1994, pp. 110–116.

Ide, N., D. Greenstein and P. Vossen (eds). *Special Issue on EuroWordNet, Computers and the Humanities*, 32(2–3) (1998).

Ide, N. and J. Véronis. "Introduction to the Special Issue on WSD: The State of the Art; Special Issue on Word Disambiguation". *Computational Linguistics*, 24(1) (March 1998), 1–40.

Krovetz, R. and W.B. Croft. "Lexical Ambiguity and Information Retrieval". *ACM Transaction on Information Systems*, 10(1).

Kuhn, R. and R. De Mori. "The Application of Semantic Classification Trees to Natural Language Understanding". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5) (May 1995), 449–460.

de Loupy, C., P.-F. Marteau and M. El-Bèze. *Navigating in Unstructured Textual Knowledge Bases*. La Lettre de l'IA, No. 134-135-136, May 1998, pp. 82–85.

de Loupy, C., M. El-Bèze and P.-F. Marteau. *Word Sense Disambiguation using HMM Tagger*. LREC, Granada, Spain; May 28–30 1998, pp. 1255–1258.

de Loupy, C., M. El-Bèze and P.F. Marteau. *WSD Based on Three Short Context Methods*. SENSEVAL Workshop, Herstmonceux Castle, England, 2–4 September, 1998, http://www.itri.brighton.ac.uk/events/senseval/.

de Loupy, C., P. Bellot, M. El-Bèze and P.F. Marteau. *Query Expansion and Classification of Retrieved Documents*. TREC-7, Gaithersburg, Maryland, USA, 9–11 November 1998.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross and K. Miller. *Introduction to WordNet: An On-Line Lexical Database*. http://www.cosgi.princeton.edu/∼wn, August 1993.

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

Segond, F., A. Schiller, G. Grefenstette and J.-P. Chanod. *An Experiment in Semantic Tagging Using Hidden Markov Model Tagging*. ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications; Madrid, July 1997.

Spriet, T. and M. El-Bèze. "Introduction of Rules into a Stochastic Approach for Language Modelling". In *Computational Models for Speech Pattern Processing*, NATO ASI Series F, editor K.M. Ponting, 1997.

Yarowsky, D. *One Sense per Collection ARP*. A Human Technology Workshop, Princeton, NJ, 1993.