Improving WSD with Multi-Level View of Context Monitored by Similarity Measure

E. Crestan[1,2], M. El-Bèze[1] and C. de Loupy[2]

(1) Laboratoire d'Informatique d'Avignon
339 ch. Des Meinajaries, BP 1228
F-84911 Avignon Cedex 9
{eric.crestan, marc.elbeze}@lia.univ-avignon.fr

(2) Sinequa
51-59 rue Ledru Rollin
F-94200 Ivry-sur-Seine
{crestan, loupy}@sinequa.com

Abstract

The approach presented in this paper for Word Sense Disambiguation (WSD) is based on a combination of different views of the context. Semantic Classification Trees (SCT) are employed over a short and a multi-level view of context, including rough semantic features, while a similarity measure is used in some particular cases to rely on a larger view of the context. We also describe our two-step approach based on HMM for the *all-word* task.

Introduction

In the tracks of SENSEVAL-1 (Kilgarriff and Rosenzweig, 2000), the second edition of the word sense disambiguation evaluation campaign offers a new set of words to test improvements in the domain of WSD. It also includes a new task, aimed at disambiguating each word of a text.

Our approach for the lexical sample task is based on three different views of the context, which allows us to consider more information for sense tagging. In order to deal with short-range view of the context, we have chosen to use Semantic Classification Trees (SCT) (Kuhn and De Mori, 1995), which are binary decision trees. Moreover, based on our experience, we will show, that using rough semantic features as a higher-level view of the context yields substantial increases in performance. Finally, a similarity distance is employed in order to capture longer-range context information.

The paper is organized as follows: in the first part (Section 1), the work we have done on the *lexical sample* task is presented. This part includes a brief overview of the SCT approach (Section 1.1) and we show how the coverage it yields could be increased while using more or less rough semantic features thanks to a multi-level view of the context (Section 1.2). In Section 1.3, we propose to use a similarity measure like those used in document retrieval in order to select a sense among those proposed by the SCT systems. The second part (Section 2) is dedicated to the *all-words* task. A two-step approach based on a trisem-bisem model is presented (Section 2.1). Then, we propose to apply a special process on the most frequent words in the task (Section 2.2). In conclusion, the results for both tasks are presented.

1 Lexical Sample Task

The *lexical sample* task of SENSEVAL-2 is composed of 29 nouns, 29 verbs and 15 adjectives in context. We decided to handle the totality of the words, and always assign one and only one sense to each test word (*recall = precision*). For training purpose, we used the corpus supplied for each word to be disambiguated. However, the number of training sentences supplied was greatly reduced compared to that of the first SENSEVAL exercise. By comparison, the average number of training sentences for the nouns in SENSEVAL-1 data was about 410 sentences/word. Here, the average number of training sentences is only 121 sentences/word. This difference leads us to believe that the present evaluation may be much harder than the previous one. The senses used for this evaluation come from the Wordnet 1.7 pre-release (Miller *et al.*, 1990).

1.1 Applying SCT to WSD

Yarowsky (1993) states that most clues for the purpose of disambiguation are present in a micro-context of *3* or *4* words. SCT seems to be an adequate approach to handle short contexts. Moreover, SCT, which are binary decision trees, permit a simple interpretation of the results, by recovering the successive questions asked along each path from the root to a leaf. Kuhn and

De Mori (1995) have shown that these extracted rules correspond to regular expressions. However, this approach requires a certain amount of data in order for the trees to be grown with reliable questions in its nodes.

Relying on previous work in this field (Loupy *et al.*, 2000), the training corpus was used to build one tree for each word to be disambiguated. While growing the trees, the list of possible questions is built at each node, taking into consideration the position of an element of the context (lemma in this case). The Gini impurity $G(X)$ (Breiman *et al.*, 1984) is then computed (*formula 1*) for each question in the list, in order to extract the one which generates the highest decrease in impurity ΔG_q (*formula 2*).

$$G(X) = 1 - \sum_{s \in S} P(s/X)^2 \qquad (1)$$

Where $P(s/X)$ is the probability of sense s given population X,

$$\Delta G_q = G(T) - p_{Yes_q} G(Yes_q) + p_{No_q} G(No_q) \quad (2)$$

Here Yes_q and No_q correspond respectively to the population answering *yes* or *no* to the question q; p_{Yes_q} (respectively p_{No_q})is the proportion of population T answering *yes* (respectively *no*) to question q.

A more detailed description of our approach to SCT can be found in Crestan and El-Bèze (2001).

The data had to be pre-processed before they could be used. Motivated by conclusions drawn from recent work (see for example Loupy and El-Bèze (2000)), the context was lemmatized, except for the word to be disambiguated. The determiners, possessive pronouns, adverbs and adjectives were removed, because they bring more noise to the tree growing process than they help capture relevant clues. However, some adjectives were preserved, when they were part of a compound noun, as in "*short circuit*". For the part-of-speech (POS) tagging process and lemmatization process, the English Tree-Tagger (Schmid, 1994) was used.

1.2 Rough semantic features as a multi-level view of context

Regarding previous work using SCT, the novelty of our approach consists in the introduction of rough semantic features into the context in order to increase the coverage of the trees. The process of tree growing can quickly suffer from lack of data. The ability of our system to view the context, not only as a succession of lemmas, but also as a multi-level view makes it more robust and reliable.

We used the Semantic Classes (SC) proposed in Wordnet in order to improve the coverage of the trees. There are *26* SC associated with nouns (e.g. <noun.body> for body related nouns) and *15* SC associated with verbs (e.g. <verb.motion> for motion related verbs). Because most of the adjectives and adverbs were removed during the pre-processing phase and because they have only one or two possible SC, their respective SC are not employed.

During the SCT building process, there is now not just one question to ask at a given position in a training sentence, but *n+1* (where *n* is the number of possible SC associated with a lemma). For example, the sentence sample in *figure* 1 leads to 16 possible questions if considering SC, and only *7* questions if considering only lemmas.

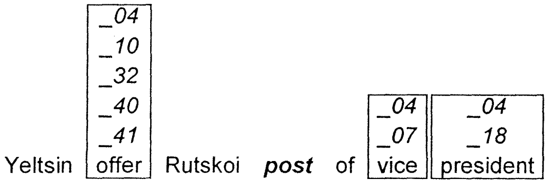

Figure 1: Example of SC usage

SC are added regardless of the POS. In the example above, the term *offer* can only be a verb, but we still associate with it the classes *_04* (noun.act) and *_10* (noun.communication), which are associated with the noun-senses of *offer*. There are two reasons for this choice: First, in the case of erroneous POS tagging, we would not be able to characterize a sense using the adequate SC. Second, tests have shown that results obtained using POS related SC or all the SC are comparable. This last point could be explained by the aptitude of SCT to select the best questions. Therefore, SCT are able to partially disambiguate the local context at a coarse-grained sense level when enough data are available. Consequently, it seems useless to make assumptions about POS.

Experiments carried on the SENSEVAL-1 data, has shown an improvement of about 2.5% on nouns and about 3% on verbs when using the Semantic Classes.

1.3 Similarity measure for a long-range view of the context

Experience has shown us that a window size of *WS*=3 is enough for disambiguation in many cases, but there are still numerous cases for which a larger window size is required. However, if a larger window size can provide more information for sense detection, it may also add more noise. In order to cope with this drawback, a similarity measure is employed (a technique usually applied in the field of document retrieval), as a ruler to decide which sense seems the more likely, considering the whole sentence (Figure 2). Firstly, three different Window Sizes (*WS*) are considered and run through the appropriate SCT process (trained on the same WS). Secondly, for each sense proposed by the SCT systems (E_1, E_2, and E_3), a pseudo-document is built with the corresponding sentences from the training corpus. Then, a similarity measure as those used in document retrieval is computed between the test sentence ($WS=|S|$) and each of the pseudo-documents (*i.e.* senses). Finally, only the sense having the best score is kept. The similarity measure used here is the Cosine measure (Salton and McGill, 1983).

The analysis of the results has shown that monitoring several SCT based views of the context by using the here above described technique leads to an average precision improvement of about 2%.

2 All-Words Task

The second task proposed in SENSEVAL consists in tagging almost all the words of a text. This is a more difficult task because in the first one, only some words have to be studied, whereas the behavior of all words must be known in order to correctly tag an entire text. Hidden Markov Models (HMM) have shown their efficiency in many NLP domains: part-of-speech tagging (El-Bèze and Merialdo, 1999), speech recognition (Jelinek, 1998), etc. Moreover, they have been used in semantic disambiguation with some success (Loupy *et al.*, 1998). Therefore, we decided to use this method for the all words task.

The test corpus supplied is composed of *2473* words to be disambiguated out of *5836* words. All POS are represented: *1140* nouns, *544* verbs, *453* adjectives and *299* adverbs (according to the supplied TreeBank-tagged file).

2.1 A coarse to fine-grained sense strategy

In a previous experiment (Loupy *et al.*, 1998), HMM were applied directly to disambiguate senses at fine-grained level using a unisem-bisem model, after training on the SemCor (Miller *et al.*, 1993). However, even if this method achieves correct results (72 % of correct assignation), it does not really improve over the unisem model. Therefore, it is recognized that there are not enough data to correctly learn the transitions between senses.

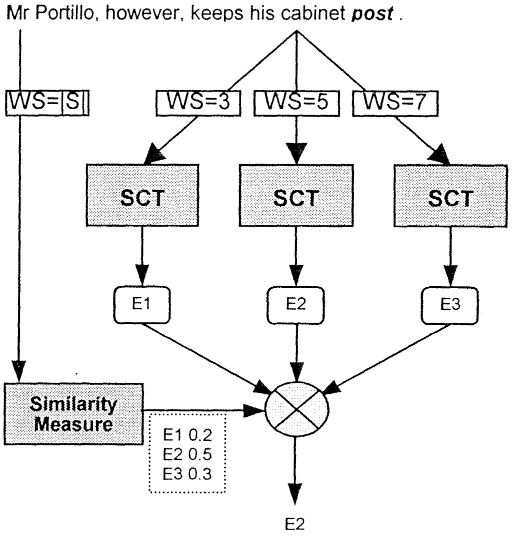

Figure 2: Sense selection using a similarity measure

On the other hand, an HMM unisem-bisem model brings a slight improvement as compared to unisem alone when applied to a coarser semantic level, that is SC (Loupy *et al.*, 1998). We adopt the following two-step strategy:

- Firstly, determine the SC associated with each word in the text (*formula 4)*

$$\begin{aligned}\tilde{G} &= Arg\,\underset{G}{Max}\left[P(G/L)\right] \\ &= Arg\,\underset{G}{Max}\left[P(L/G)P(G)\right]\end{aligned} \tag{4}$$

  where G is the set of possible coarse-grained semantic classes associated to the lemma L.

- Secondly, assign the most probable fine-grained sense according to the word and the previously retrieved SC (*formula 5*).

$$\tilde{S} = Arg\,\underset{S}{Max}\left[P(S/G,L)\right] \tag{5}$$

  where S is the set of possible senses associated with the lemma L and its possible semantic classes G.

To cope with the well-known sparse data problem, some assumptions allow us to use a HMM (trisem-bisem model), in order to estimate *P(G)* (*formula 6*) and *P(L/G)* (*formula 7*).

$$P(G) \approx \prod_i \lambda \times P(g_i | g_{i-2}, g_{i-1}) + (1-\lambda) \times P(g_i | g_{i-1}) \quad (6)$$

and $$P(L|G) \approx \prod_i P(l_i | g_i) \quad (7)$$

In the same way, assumptions were made in order to estimate the probability *P(S/G,L)* (formula 8).

$$P(S|G,L) \approx \prod_i P(s_i | g_i, l_i) \quad (8)$$

2.2 Using Lexical Sample Task Experience

In view of our experience with the lexical sample task, we decided to take advantage of it. The most frequent words among those to be disambiguated in the all-words task and which were also present in the SENSEVAL-2 lexical sample task were extracted. For those words, the technique presented in Section 1 was applied. In this way, 4 verbs (*call*, *develop*, *find* and *use*) and 2 nouns (*child* and *church*) were disambiguated by the SCT-Cosine method, as described in Section 1.3.

Results and Conclusion

As mentioned in section 1, the scores for the second edition of the *lexical sample* task are much lower than for the first edition (about 20%). However, our system achieved satisfactory results comparing to other participants (see table 1) and even accessed the top-5 systems. The use of SC as a multi-level view of the context has generated significant improvements in the results. As well as, the combination of different window sizes using similarity measure on a larger context as a judge has shown noticeable improvements.

	Lexical Sample		All-Words	
	Precision	Recall	Precision	Recall
Fine	61.3%	61.3%	61.8%	61.8%
Coarse	68.2%	68.2%	62.6%	62.6%

Table 1: Results for fine and coarse-grained senses

For the *all-words* task, our system has proven to be one of the bests, achieving an average precision/recall of 61.8%, and this, despite the absence of mapping between Wordnet 1.6 senses used for training purpose (SemCor) and Wordnet 1.7 senses used as test references.

References

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

E. Crestan and M. El-Bèze (2001): "*Improving Supervised WSD by Including Rough Semantic Features in a Multi-Level View of the Context*", SEMPRO-2001 Workshop, Edinburgh. http://www.lia.univ-avignon.fr/publications/fich_art/LIA-SEMPRO-2001.pdf

M. El-Bèze and B. Mérialdo (1999): "*HMM Based Taggers*", in Syntactic Wordclass Tagging, ed. Hans Van Halteren, Kluwer Academic Publishers, Text and Language Technology, pp 263-284.

F. Jelinek (1998): "*Statistical Methods for Speech Recognition*", MIT Press, Cambridge.

A. Kilgarriff and J. Rosenzweig (2000): "*English SENSEVAL: Report and Results*", In Proc. LREC, Athens, Greece, Vol 3, pp 1239-1244.

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

C. de Loupy and M. El-Bèze (2000): "*Using Few Clues can compensate the small amount of resources available for Word Sense Disambiguation*", LREC, Athens, Vol 1, pp 219-223.

C. de Loupy, M. El-Bèze and P.-F. Marteau (1998): "*Word Sense Disambiguation using HMM Tagger*", LREC, Grenade, Vol 2, pp. 1255-1258.

C. de Loupy, M. El-Bèze and P.-F. Marteau (2000): "*Using Semantic Classification Trees for WSD*", Computer and the Humanities, N° 34, Kluwer Academic Publishers, pp 187-192.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller (1990): "*Introduction to WordNet: An on-line lexical database*," International Journal of Lexicography, vol. 3(4), pp 235-244.

G. A. Miller, C. Leacock, R. Tengi, and T. Bunker (1993): "*A Semantic Concordance*", Proceedings of ARPA Workshop on Human Language Technology, Plainsboro, New Jersey, pp 303-308.

G. Salton and M.J. McGill (1983): "*Introduction to Modern Information Retrieval*", McGraw-Hill, New York.

H. Schmid (1994): "*Probabilistic Part-of-Speech Tagging Using Decision Trees*". In Proceedings of the Conference on New Methods in Language Processing. Manchester, UK, pp 44-49.

D. Yarowsky (1993): "*One sense per collocation*", In Proceedings of the ARPA Workshop on Human Language Technology, pp 266-271.