SENSEVAL-3: Third International Workshop on the Evaluation of Systems
for the Semantic Analysis of Text, Barcelona, Spain, July 2004
Association for Computational Linguistics

Contextual Semantics for WSD

ERIC CRESTAN[1,2]

(1) Sinequa SAS
51-54, rue Ledru-Rollin
92400 Ivry-sur-Seine, *France*
Crestan@sinequa.com

(2) Laboratoire Informatique d'Avignon
B.P. 1228 Agroparc
339 Chemin des Meinajaries
84911 Avignon Cedex 9, *France*

Abstract

For Sinequa's second participation to the Senseval evaluation, two systems using contextual semantic have been proposed. Based on different approaches, they both share the same data preprocessing and enrichment. The first system is a combined approach using semantic classification trees and information retrieval techniques. For the second system, the words from the context are considered as clues. The final sense is determined by summing the weight assigned to each clue for a given example.

1 Introduction

In the framework of the Senseval-3 evaluation campaign on Word Sense Disambiguation (WSD), we presented two systems relying on different strategy. The system *SynLexEn* is an evolution from the system used during the Senseval-2 campaign. It is based on two steps. The first step uses semantic classification trees on a short context size. A decision system based on document similarity is used as second step. The novelty of this system resides in a new vision level on the context. The semantic dictionary of Sinequa is extensively used in this process.

The second system, *SynLexEn2*, is based on weighted clues summation over a short context size. From the training data, a score is computed for each word in a short context size, for each sense.

In Section 2, the combined approach system for WSD is presented. We first give an overview of the data pre-processing that was applied (Section 2.1). Then, a brief description of Semantic Classification Trees is given (Section 2.2) along with a description of additional data used for semantic view of short and long context (Section 2.3 and Section 2.4). Next, a semantic information retrieval system used in order to select the appropriate sense is proposed (Section 2.5).

Finally, the SynLexEn2 system is presented in Section 3. We then conclude with the evaluation results for both systems in Section 4.

2 Combined approach

The *SinLexEn* system is quite similar to the system used during the last Senseval-2 evaluation campaign (Crestan *et al.*, 2001). It is based on two stages: the first stage uses three Semantic Classification Trees in parallel, trained on different size of context. Then, the second stage brings in a decision system based on information retrieval techniques. The novelty of this approach dwells in the use of semantic resource as conceptual view on extended context in both stages.

2.1 Data pre-processing

The first step in order to get the most from the data is to lemmatize and clean sentences. Each paragraph from the training and the test data are first passed though an internal tagger/lemmatizer. Then, some grammatical words are removed such as articles and possessive pronouns. Only one word is not handled in this process, it is the word to be disambiguated. Indeed, previous works (Loupy *et al.*, 1998) have shown that the form of this word could bring interesting clues about its

possible sense. Other pronouns, such as subject pronouns, are replaced by a generic PRP tag.

2.2 Semantic Classification Trees for WSD

The Semantic Classification Trees (SCT) were first introduced by Kuhn and De Mori (1995). It can be defined as simple binary decision trees. Training data are used in order to build one or more trees for each word to be disambiguate. An SCT is made up of questions distributed along the tree nodes. Then, each test sequence is presented to the corresponding trees and follows a path along the SCT according to the way the questions are answers. When no more question is available (arrived at a leaf), the major sense is assigned to the test.

In order to build the trees, the Gini impurity (Breiman *et al.*, 1984) was used. It is defined as:

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

where *P(s/X)* is the likelihood of sense *s* given the population *X*.

At the first step of the tree building process, the Gini impurity is computed for each possible questions. Then, the best question is selected and the population made up of all the examples is divided between the ones which answer the question (*yes* branch) and the others (*no* branch). The same process is recursively applied on each branch until the maximum tree depth is reached.

In the framework of the *SinLexEn* system, three different trees have been built for each word to be disambiguated. They use different context size, varying from one to three words on each side of the target word. Following is an example of three training sequences using respectively 1, 2 and 3 words on each side of the target (*0#sense*):

		-1#make	*0#sense*	1#of		
	-2#make	-1#more	*0#sense*	1#to	2#annex	
-3#ceiling	-2#add	-1#to	*0#sense*	1#of	2#space	3#and

The number preceding the # character gives the position of the word according to the target. The set of possible questions for the SCT building process is composed of all the words present in considered window width. The tree shown in Figure 1 was built for the word '*sense'* on a window width of 3 words. Each node corresponds to a question, while leafs contain the sense to be assigned to the target. The test sequence–*1#make 0#sense 1#of* will be assigned to *sense%1:10:00::* sense ("*the meaning of a word or expression*") from *WordNet* (Miller *et al.*, 1990). For a more detail description of SCT, see (Crestan *et al.,* 2003).

2.3 Semantic in short context

WSD is much more easy to do for a human than for a machine. By simply reading a sentence, we can determine at a glance the meaning of a word in this particular context. However, we are not relying solely on what the words look like. The human brain is able to see the correlation between '*open a box*' and '*open a jar*'. We have the ability to generalize over similar "concepts". In order to follow this scheme, we used the *WordNet*'s semantic classes (SC). It enables a generalization over words

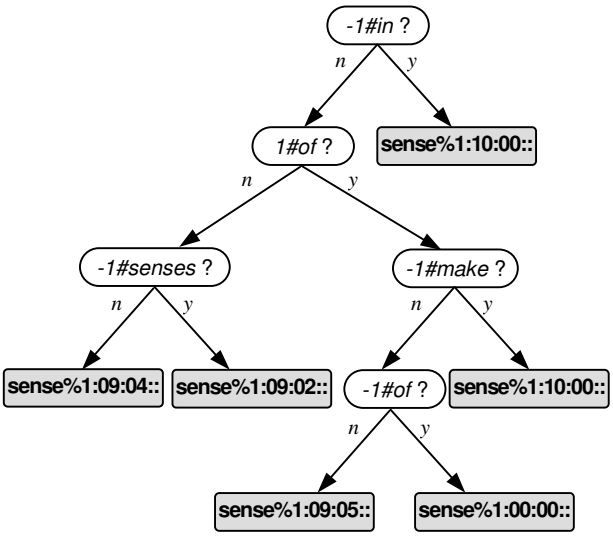

Figure 1. SCT example for the target word *'sense*'

sharing the same high-level hyperonym. Because the correct SC is not known for each word in context, all the possible SC were included in the set of questions for a given word and position. The *WordNet* top ontology is separated in 26 SC for the nouns and 15 for the verbs. An extended description of SC can be found in (Ciaramita and Johnson. 2003). In the following example, the two first sequences share the same sense ('*cause to open or to become open*') whereas the last

sequence corresponds to another sense ('*start to operate or function or cause to start operating or functioning*'):

0#open	1#box	1#06	1#20	1#23	1#25	1#35
0#open	1#jar	1#06	1#11	1#23	1#38	1#42
0#open	1#business	1#04	1#09	1#14		

Two of the five SC are common to both words *box* and *jar*:

- *06*: nouns denoting man-made objects,
- *23*: nouns denoting quantities and units of measure.

However, they have nothing in common with the word *business*.

Although many wrong SC are proposed for each word according to its context, we noticed a 2% improvement on Senseval-2 data while using these "high level information".

2.4 Semantic in full context

The main improvement for this evaluation is the use of semantic clues at a paragraph level. Sinequa has developed along the last 5 years a large scale semantic dictionary of about 100.000 entries. All the word of the language are organized across a semantic space composed of 800 dimensions. For example, a word such as 'diary' is present in the dimensions: *calendar*, *story*, *book* and *newspaper*. It has been wildly used in the information retrieval system Intuition (Manigot and Pelletier, 1997), (Loupy *et al.*, 2003).

For each training sample, we summed the semantic vectors of each word. This step results on a global semantic vector from which only the 3 most representative dimensions (with highest score) where kept. That additional information has been used as possible question in the tree building process. Then, the same semantic analysis has been done on each test sentence. For example, the major dimension represented in the next sentence for the word *material* is '*newspaper*':

'*furthermore , nothing have yet be say about all the research that do not depend on the collection of datum by the sociologist (primary datum) but instead make use of secondary datum - the wealth of* ***material*** *already available from other source , such as government statistics , personal diary , newspaper , and other kind of information .*'

This enables a new vision of the context on a wider scale than the one we used with only short context SCT. Preliminary experiments carried on the Sensenval-2 nouns have shown a 1% improvement. Some nouns such as *dyke*, *sense* and *spade* have been dramatically improved (more than 5%). Although, words such as *authority* and *post* have had about 5% decrease in precision. A first hypothesis can be proposed to explain the gain of some words while others have lost in precision: the use of a wide context semantic is mostly benefic in the case of homonymy, while it is not when dealing with polysemy.

2.5 Semantic similarity for decision system

In order to select the appropriate sense among the three senses proposed by the SCT, a decision system was used. It is based on the *Intuition* search engine used on the *Default* mode: the words and the semantic vectors of documents are used. The final score is a combination between the words score and the semantic score. Moreover, all the sentences linked to a given sense in the training data were concatenated in order to form a unique document (pseudo-document). Then, for a given test instance, the whole paragraph was used to query the engine. The pseudo-document's scores were then used in order to select among the three senses proposed by the SCT. A 2% improvement have been observed during the Senseval-2 evaluation campaign while using this strategy.

3 Maximum clues approach

Starting from the same preprocessing used for the combined approach, we implemented a simple approach based on Gini impurity. Considering a short context, the Gini impurity is computed for all the possible questions in the training data (including the questions about semantic level). For instance, if the question – *1#of* appears 3 times with the sense *S1*, 1 time with *S2* and does not appear in 1 example of sense *S1* and 2 examples for sense *S2*, the final score for this question is:

$$G(-1\#of) = [1-(3/4)^2-(1/4)^2] + [1-(1/3)^2-(2/3)^2] = 0.82$$

Which corresponds to the Gini impurity of the examples where *–1#of* is present, plus the Gini impurity for the examples where it is not. Then, a score is given for each sense according to each question. For the previous example, the score *S1* for the question *–1#of* is:

$$Score(S1, -1\#of) = P(S1/-1\#of) * [G - G(-1\#of)]$$

Where *G* is the initial Gini impurity, minus the Gini impurity of *G(-1#of)* and weighted by the probability of *S1* when *–1#of* was observed.

When disambiguating a test phrase, the score for each sense is computed by summing the individual score for each question. The highest score gives the sense.

This simple approach has shown similar results as those obtained with the combined approach on nouns. Unlike the trees, this system is able to benefit from all the clues in the training corpus. At the opposite, for the SCT, if two questions get rather good scores at first stage, only on question will be selected in order to build the node. This prevents from using clues from the other question because its population is (or might be) divided between the two branches.

4 Results and conclusion

For the third edition of the Senseval campaign, the sense repository for the verbs was different, using *WordsMyth* instead of *WordNet*. The proportion of nouns, verbs and adjectives was also different. Because of these changes, it is difficult to compare this evaluation results with the previous ones.

	Fine	**Coarse**
SinLexEn	67.2%	74.3%
SinLexEn2	66.8%	73.6%

Table 1: Precision for both systems

The precisions of both systems are presented in Table 1. The column named *Fine* corresponds to fine grain scores, while the column named *Coarse* is for the coarse grained senses.

Although we are using different strategies, both systems give approximately the same results for fine grained precision. According to the previous evaluation, we can observe almost 5% increase in precision. However, this increase cannot be taken as significant because of the differences between the evaluations.

A comparative evaluation have to be carried now in order to establish if a combination of both system could improve the final score.

References

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

M. Ciaramita and M. Johnson. 2003. *Supersense Tagging of Unknown Nouns in WordNet*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

E. Crestan, M. El-Bèze and C. de Loupy 2001. *Improving WSD with Multi-Level View of Context Monitored by Similarity Measure*, Proceedings of Senseval-2, 39th ACL.

E. Crestan, M. El-Bèze, C. de Loupy. 2003. *Peut-on trouver la taille de contexte optimale en désambiguïsation sémantique ?*, In Proceedings of TALN 2003.

C. de Loupy, V. Combet and E. Crestan 2003. *Linguistic resources for Information Retrieval*, Proceedings ENABLER/ELSNET, International Roadmap for Language Resources, Paris.

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), p 449-460.

C. de Loupy, M. El-Bèze and P.-F. Marteau. 1998. *WSD based on three short context methods*, SENSEVAL Workshop, Herstmontceux.

L. Manigot, B. Pelletier. 1997. Intuition, une approche mathématique et sémantique du traitement d'informations textuelles. Proceedings of Fractal'1997. pp. 287-291. Besançon.

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), p 235-244.