Automatic Noun Sense Disambiguation*

Paolo Rosso[1], Francesco Masulli[2], Davide Buscaldi[3],
Ferran Pla[1], and Antonio Molina[1]

[1] Dpto. de Sist. Informáticos y Computación, U. Politécnica de Valencia, Spain
{prosso,fpla,amolina}@dsic.upv.es
[2] INFM-Genova and Dip. di Informatica, Università di Pisa, Italy
masulli@disi.unige.it
[3] Dip. di Informatica e Scienze dell'Informazione, Università di Genova, Italy
buscaldi@disi.unige.it

Abstract. This paper explores a fully automatic knowledge-based method which performs the noun sense disambiguation relying only on the WordNet ontology. The basis of the method is the idea of conceptual density, that is, the correlation between the sense of a given word and its context. A new formula for calculating the conceptual density was proposed and was evaluated on the SemCor corpus.

1 An Extension of the Conceptual Density

The task of Word Sense Disambiguation (WSD) consists of examining word tokens and specifying exactly which sense of each word is being used. The WordNet (WN) ontology, based on synsets (*sets* of *syn*onyms), is the external lexical resource which is often used to perform the WSD task. In most of the WSD approaches, a word is disambiguated along with a portion of the text in which it is embedded, that is, its context. When the initial input source of information (i.e., the word and its context) is processed only together with the lexical knowledge source (e.g. WN), a fully automatic method which does not require any kind of training process is needed to perform WSD.

Conceptual Density (CD) is a measure of the correlation among the sense of a given word and its context. The foundation of this measure is the *Conceptual Distance*, defined as the length of the shortest path which connects two concepts in a hierarchical semantic net. The starting point for our work was the CD formula of Agirre and Rigau [1], which compares areas of subhierarchies:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^i}{\sum_{i=0}^{h-1} nhyp^i} \tag{1}$$

where c is the synset at the top of subhierarchy, m the number of word senses falling within a subhierarchy, h the height of the subhierarchy, and $nhyp$ the

* This work was supported by the Spanish Research Projects CICYT TIC2000-0664-C02 and TIC2000-1599-C01-01, and the VIDI of the Univ. Politécnica Valencia.

averaged number of hyponyms for each node (synset) in the subhierarchy. The numerator expresses the expected area for a subhierarchy containing m marks (word senses), while the divisor is the actual area.

The synsets of the senses of the word to be disambiguated fall in different places in the hierarchy, and in most cases this means that the hierarchy can be partitioned into subhierarchies (we refer them as *clusters*), each containing exactly one sense of the word to be disambiguated (therefore, a word having six senses in WN should determine six partitions). When two or more senses of the word are one hyponym of each other the partition cannot be done. Therefore, in such conditions the word sense disambiguation cannot be carried out.

Formula 1 considers the averaged number of hyponyms of each node in the subhierarchy. Due to the fact that the averaged number of hyponyms for each node in WN1.6 is greater than in WN1.4 (the version which was used in the original work presented in [1]), we decided to consider only the *relevant* part of the subhierarchy determined by the synset paths (from c to an ending node) of the senses of both the word to be disambiguated and its context. The base formula is based on the M number of relevant synsets (corresponding to the *marks* m in Formula 1) divided by the total number nh of synsets of the subhierarchy.

$$baseCD(M, nh) = M/nh \tag{2}$$

Formulas 1 and 2 do not take into account sense frecuency. It is possible that both formulas select subhierarchies with a low frecuency related sense. In some cases this would be a wrong election. This pushed us to modify the CD formula by including also the information about frequency that comes from WN:

$$CD(M, nh, f) = M^{\alpha}(baseCD)^{\log f} \tag{3}$$

where M is the number of relevant synsets, α is a constant (the best results were obtained with α near to 0.10) , and f is an integer representing the frequency of the subhierarchy-related sense in WN (1 means the most frequent, 2 the second most frequent, etc.). This means that the first sense of the word (i.e., the most frequent) gets at least a density of 1 and one of the less frequent senses will be chosen only if it will exceed the density of the first sense. The M^{α} factor was introduced to give more weigth to the subhierarchies with a greater number of relevant synsets, when the same density is obtained among many subhierarchies.

We included some adjustment factors based on context hyponyms, in order to assign an higher conceptual density to the related cluster in which a context noun is an hyponym of a sense of the noun to be disambiguated (the hyponymy relation reflects a certain correlation between the two lexemes). We refer to this technique as to the *Specific Context Correction (SCC).* The idea is to select as the winning cluster the one where one or more senses of the context nouns fall beneath the synset of the noun to be disambiguated.

An idea connected to the previous one, was to give more weight to the clusters placed in deeper positions. We named this technique as *Cluster Depth Correction (CDC).* When a cluster is below a certain averaged depth (which was determined in an empirical way to be about 4) and, therefore, its sense of the noun to be disambiguated is more specific, the conceptual density of Formula 3 is augmented proportionally to the number of the contained relevant synsets:

$$CD * (depth(cl) - avgdepth + 1)^{\beta} \tag{4}$$

where $depth(cl)$ returns the depth of the current cluster (cl) with respect to the top of the hierarchy; $avgdepth$ is the averaged depth of all clusters in the subhierarchies obtained from Semcor; its value was empirically determined to be equal to 4; and β is a constant (the best results were obtained with $\beta = 0.70$).

Finally, we investigated the possibility of expanding the context with the gloss of the noun to be disambiguated. This led to worse results, since the gloss was examined without considering the syntactic category of its words and a certain "noise" was introduced as consequence of considering all lexemes as possible nouns. A refinement was done by considering only monosemic words of the gloss but, in spite of that, the performance for the noun disambiguation task did not increase. In order to consider only nouns, we first Part-Of-Speech tagged the gloss. We used a POS tagger based on Lexicalized-HMM. This tagger was evaluated achieving a precision of 96.8% on the *Wall Street Journal* corpus [6].

2 Experimental Results and Conclusions

The first goal of our work was to determine an effective window context size. Like many other researchers have done [4], we have carried out WSD experiments using the Semcor corpus[1]. The best results in term of precision were obtained with a context window size of 2 nouns, confirming that closer nouns give a more precise definition of the context than farther ones. The drawback of this approach is the average recall (around 60%). This is mainly due to the fact that many nouns have senses that differ slightly one from each other. This can be viewed in a hierarchy as deep clusters with only one synset inside them (corresponding to the sense of the noun to be disambiguated). In most cases, there are no context nouns falling in these "singular" clusters, and the result is that sense disambiguation cannot be done.

We combined different correction models (SCC and CDC) over the whole SemCor corpus and for different window sizes (two, four and six). All these experiments outperformed the baseline precision (76.04%) and the baseline recall (23.21%)[2]. The best precision measure of 81.48% was obtained without any correction factor and with a very small window of size two (recall 60.17% and coverage 73.81%). Using the SCC technique, although precision was not affected significantly, we obtained only small improvements on recall and coverage measures. With regard to the CDC technique, the results did not differ significantly to those obtained with the previous correction factor. Improvements on recall

[1] The results were obtained over the 19 randomly selected SemCor files: br-a01,b13,c01,d02,e22,r05,g14,h21,j01,k01,k11,l09, m02,n05,p07,r04,r06,r08,r09.

[2] The baseline precision was calculated assigning the most frequent sense to every noun, whereas the baseline recall was calculated for monosemic nouns only.

(61.27%) and coverage (77.87%) measures were obtained increasing the size of the context window. Recall remained approximately around 60% and varied slightly even when considering many context nouns (e.g. six), whereas coverage improved even if at the price of obtaining a lower precision measure.

For each noun to be disambiguated, we investigated the possibility of expanding its context adding the gloss, excluding the example phrases. In order to reduce the "noise" introduced considering all the words of the gloss, only monosemic words were added to the context of the noun to be disambiguated. In a second approximation, we POS-tagged the words of the gloss and extracted only its monosemic nouns which were included in the context.

The tests with this "expanded context" were conducted over the first 10 files from Brown1 of SemCor, and the CDC factor was also employed. The results of averaged P(recision), R(ecall) and C(overage) are the following: *CDC model and gloss* P=78.42%, R=61.86% and C=78.80%; *CDC model and POS-tagged gloss* P=80.77%, R=62.42% and C=77.24%; *CDC model and no gloss* P=80.91%, R=62.19% and C=76.81%. In order to have a certain balance in terms of precision / recall, a window size of 4 (previous to its expansion with the monosemic nouns of the gloss) was used in the experiments. The size of the expanded context was 5.92 on average (i.e., it contained 6 nouns approximately).

Without POS-tagging the gloss, even considering only its monosemic words, the recall decreased slowly and the precision decreased by an average of more than 2% with respect to the precision obtained without the gloss. The POS-tagging preprocess of the gloss permitted to obtain improvements both on recall and coverage without practically losing in precision. These results are promising if we compare them to those obtained using the original CD formula [1] (precision 81.97% vs. 66.4% and recall 69.02% vs. 58.8% for the file br-a01 of SemCor) especially if we consider that the much more fine-grained 1.6 version of WN was used and only a very small context window size of two to four nouns was needed.

At the moment, we are applying the proposed WSD method to sense-tagged XML documents retrieval [3]. Further work needs to be done to perform the all-word disambiguation task, the evaluation of the method against the Senseval corpus and the comparison with other recent approaches [2, 5].

References

1. E. Agirre, G. Rigau, A Proposal for Word Sense Disambiguation using Conceptual Distance. In Proceedings of RANLP, 1996.
2. D. Fernández-Amorós, J. Gonzalo and F. Verdejo. The role of conceptual relations in Word Sense Disambiguation. In Proceedings of NLDB-01, 2001.
3. M. Mesiti, P. Rosso, M. Merlo, A Bayesian Approach to WSD for the Retrieval of XML Documents. In Proceedings of JOTRI, Valencia, Spain, 2002, pp. 11-18.
4. R. Mihalecea, D. Moldovan, Semantic Indexing using WordNet Senses. In Proceedings of the Workshop on Recent Avances in NLP and IR, 2000.
5. A. Montoyo, Desambiguación léxica mediante Marcas de Especificidad. Ph.D. Thesis, Universidad de Alicante, 2002.
6. F. Pla, A. Molina, Part-of-Speech Tagging with Lexicalized HMM. In Proceedings of RANLP, Tzigov Chark, Bulgaria, 2001.