Context Expansion with Global Keywords for a Conceptual Density-Based WSD

Davide Buscaldi[1], Paolo Rosso[2], and Manuel Montes y Gómez[3,2]

[1] Dipartimento di Informatica e Scienze dell'Informazione (DISI),
Università di Genova, Italy
buscaldi@disi.unige.it
[2] Dpto. de Sistemas Informáticos y Computación (DSIC),
Universidad Politecnica de Valencia, Spain
{prosso, mmontes}@dsic.upv.es
[3] Lab. de Tecnologías del Lenguaje,
Instituto Nacional de Astrofísica, Optica y Electrónica, México
mmontes@inaoep.mx

Abstract. The resolution of the lexical ambiguity, which is commonly referred to as Word Sense Disambiguation, is still an open problem in the field of Natural Language Processing. An approach to Word Sense Disambiguation based on Conceptual Density, a measure of the correlation between concepts, obtained good results with small context windows. This paper presents a method to integrate global knowledge, expressed as global keywords, in this approach. Global keywords are extracted from documents using a model based on term frequency and distribution. Preliminary results show that a slight improvement in recall can be obtained over the base system.

1 Introduction

The resolution of lexical ambiguity that appears when a given word in a context has several different meanings is commonly referred as Word Sense Disambiguation (WSD). Supervised approaches to WSD usually perform better than unsupervised ones [4]. However, such approaches are afflicted by the lack of large, semantically annotated corpora. The unsupervised approach to WSD based on *Conceptual Density* and the frequency of WordNet senses [5] is an unsupervised approach which obtained good results, in terms of precision, for the disambiguation of nouns over SemCor (81.55% with a context window of only two nouns, compared with the MFU-baseline of 75.55%), and in the Senseval-3 all-words task (73.40%, compared with the MFU-baseline of 69.08%) as the CIAOSENSO-2 system [2].

Our approach obtained the above results with a context window of only two nouns, one before and one after the noun to disambiguate, exploiting the relationship existing between adjacent words. The obtained results [5] show that a larger context deteriorates the performance of the approach. We suppose that

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 263–266, 2005.

such decrease is due to the fact that distant words have little or no meaning for the disambiguation of a given word. The only relationship existings between two distant words in the same document is that they are related to the content of the document itself.

In order to introduce this information into our approach we needed to select the most representative words in a document, and adding them to the context of the word to disambiguate. The selected model for extracting document keywords was based on term frequency and distribution as presented in [3].

2 The CD-Based Approach

Conceptual Density *(CD)* is a measure of the correlation among the sense of a given word and its context. Our approach carries out the noun sense disambiguation by means of a formula [5], derived from the original Conceptual Density described in [1].

Due to the granularity of the version 2.0 of WordNet, we consider only the *relevant* part of the subhierarchy determined by the synset paths (from the synset at the top of subhierarchies to an ending node) of the senses of both the noun to be disambiguated and its context, and not the portion of subhierarchy constituted by the synsets that do not belong to the synset paths. In order to take into account also the information about frequency contained in WordNet the following fomula was introduced [5].

3 Extraction of Global Keywords

Document keywords appear usually in very different locations in the document. The Information Retrieval (IR) model proposed by [3] allows to use distribution characteristics of words to determine keywords, by computing their standard deviation. The standard deviation for the i-th word in document is computed as:

$$s_i^2 = \frac{1}{(f_i - 1)} \sum_j (l_{ij} - m_j)^2 \tag{1}$$

where f_i is the frequency of the i-th word, l_{ij} is the j-th position of the word in document, and m_j is the mean of relative location j. Thereafter, we can extract document keywords, having great frequency and standard deviation, that is, wide distribution over the text.

We applied this IR model to the three documents which are part of the Senseval-3 all-words corpus, obtaining the global keywords as shown in Table 1.

Document 1 is a part of a novel, document 2 is a newspaper article about presidential elections, while document 3 is a collection of excerpts from a bulletin board. It is noteworthy how representative are the global keywords extracted from document 2.

Table 1. Keywords extracted for each document in the Senseval-3 all-words corpus, sorted by standard deviation. Frequency is the total number of occurrences in the document, positions are the numbers identifying words' positions in the document, deviation is the standard deviation calculated over the document

Document	keywords	frequency	positions	deviation
doc1	guy	5	65,229,648,1658,1875	330.8
	course	4	124,990,1207,1994	332.9
	something	4	202,1011,1127,1907	302.1
	accident	4	776,1193,1969,1999	260.6
doc2	level	4	33,1271,1278,1344	274.2
	ticket	4	35,490,789,1258	222.6
	gop	5	6,126,431,951,1232	211.3
	pattern	4	155,498,891,1266	208.4
	election	5	51,113,510,666,1200	186.4
doc3	berkeley	3	278,356,1405	296.7
	bay	3	11,96,1105	286.9
	line	3	59,454,1214	276.7
	phone	3	58,723,1213	273.3
	book	3	301,663,1283	234.1
	room	3	306,662,1289	234.6
	night	3	5,128,887	225.2

4 Experimental Results

The Global Keywords (GK) extracted were added to the context of each word, taking them into account for the computation of Conceptual Density. Table 2 shows the obtained results, compared with those obtained with the CIAOSENSO-2 system at Senseval-3 [2] and the Most Frequent Sense (MFS) heuristic.

We obtained a slight improvement in Recall (1.7%) and Coverage ($\sim$ 3%), but there was a $\sim$ 1% loss in precision. In order to obtain better results, we decided to add to the context only two words for each document. The two words were selected on the basis of the following criteria:

1. Polysemy (i.e., those having fewer senses);
2. Depth in the WordNet hierarchy (i.e., the words whose synsets' average depth is the greatest);
3. Specificity (i.e., the words whose synsets' averaged number of hyponyms is smaller).

in Table 2 we show the characteristics of polisemy, depth and specificity of all the extracted global keywords.

5 Conclusions and Further Work

The number of the experiments and the size of the used corpus are too small to fully understand the impact of representative global information on WSD. How-

Table 2. Results obtained over the nouns in the Senseval-3 all-words corpus using context expanded with GK (*CD+GK*), the CD approach (*CIAOSENSO-2*), the MFS heuristic, and filtering GK by their characteristics - polisemy (CD+Less Polysemic), averaged depth of synsets (CD+Deepest), and averaged number of hyponyms (CD+Most Specific)

	Precision	Recall	Coverage
CIAOSENSO-2	0.743	0.497	66.9%
MFS	0.691	0.691	100%
CD+GK	0.734	0.508	69.2%
CD+Less Polysemic	0.729	0.506	69.3%
CD+Deepest	0.731	0507	69.3%
CD+Most Specific	0.730	0.507	69.4%

ever, it seems that a slight improvement in recall and coverage can be obtained without losing too much in precision. This has to be proofed over a larger corpus, such as SemCor. Filtering global keywords depending on their polisemy, depth and hyponyms features extracted from WordNet did not prove to be helpful in WSD, even if we suppose that this can be exploited in other applications, such as IR, to improve the model based on frequency and distribution of words.

Acknowledgments

We would like to thank R2D2 CICYT (TIC2003-07158-C04-03), CONACyT-Mexico (43990A-1, U39957-Y) and ICT EU-India (ALA/95/23/2003/077-054) projects, as well as the Secretaría de Estado de Educacin y Universidades de España, for partially supporting this work.

References

1. Agirre, E., Rigau, G.: A proposal for Word Sense Disambiguation using Conceptual Distance. Proc. of the Int. Conf. on Recent Advances in NLP (RANLP'95). 1995.
2. Buscaldi, D., Rosso, P., Masulli, F.: The upv-unige-CIAOSENSO WSD System. Senseval-3 Workshop, Association for Computational Linguistics (ACL-04). Barcelona, Spain (2004).
3. Lee, J., Baik, D..: A Model for Extracting Keywords of Document Using Term Frequency and Distribution Lecture Notes in Computer Science, Vol. 2588. Springer-Verlag (2004)
4. Mihalcea, R., Moldovan, D.I.: A Method for Word Sense Disambiguation of Unrestricted Text. In: Proc. of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99). Maryland, NY, U.S.A. (1999) NOTA: sostituire con Senseval-3
5. Rosso, P., Masulli, F., Buscaldi, D., Pla, F., Molina, A.: Automatic Noun Disambiguation. Lecture Notes in Computer Science, Vol. 2588. Springer-Verlag (2003) 273–276.