Combining EWN and sense-untagged corpus for WSD

Iulia Nica,[1,2] Mª. Antònia Martí,[1] Andrés Montoyo,[3] Sonia Vázquez [3]

[1] CLiC-Department of General Linguistics
University of Barcelona, Spain
amarti@ub.edu

[2] University of Iasi, Romania
iulia@clic.fil.ub.es

[3] Department of Information Systems and Languages
University of Alicante, Spain
montoyo@dlsi.ua.es, svazquez@dlsi.ua.es

Abstract. In this paper we propose a mixed method for Word Sense Disambiguation, which combines lexical knowledge from EuroWordNet with corpora. The method tries to give a partial solution to the problem of the gap between lexicon and corpus by means of the approximation of the corpus to the lexicon. On the basis of the interaction that holds in natural language between the syntagmatic and the paradigmatic axes, we extract from corpus implicit information of paradigmatic type. On the information thus obtained we work with the information, also paradigmatic, contained in EWN. We evaluate the method and interpret the results

Key words: Word Sense Disambiguation, semantic annotation

1 Introduction

Word Sense Disambiguation (WSD) is an open problem for Natural Language Processing. The focus of interest in the area of WSD has been centred principally on the heuristics used and less on the linguistic aspects of the task. However, some recent experiments ([22], [32]) have revealed that the process is in a higher degree dependent on the information used than on the algorithms that exploit it.

On the basis of these results, the present paper investigates the intensive use of linguistic knowledge in the process of Word Sense Disambiguation. We analyse some essential questions for the task of WSD: the distance between the information in the lexicon and the one in the corpus, and the identification and the treatment of local context for an ambiguous occurrence.

Depending on the sense characterisation which is taken as reference in the WSD process, there took shape two principal approaches to the task: knowledge-driven methods, which use structured lexical sources (machine readable dictionaries, semantic nets, etc.) and corpus-based methods, which use sense-tagged examples. Between