Multiwords and Word Sense Disambiguation

Victoria Arranz, Jordi Atserias and Mauro Castillo

TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1-3, E-08034 Barcelona, Catalonia
{varranz,batalla,castillo}@lsi.upc.es

Abstract. This paper[1] studies the impact of multiword expressions on Word Sense Disambiguation (WSD). Several identification strategies of the multiwords in WordNet2.0 are tested in a real Senseval-3 task: the disambiguation of WordNet glosses. Although we have focused on Word Sense Disambiguation, the same techniques could be applied in more complex tasks, such as Information Retrieval or Question Answering.

1 Introduction

In the past years there has been a growing awareness of Multiword Expressions (MWEs). Due to their complexity and flexible nature, many NLP applications have chosen to ignore them. However, given their frequency in real language data, and their importance in areas such as terminology [1], they represent a problem that needs to be addressed. Whilst there has been considerable research on the extraction of MWEs [2], little work has been carried out on their identification. This is particularly so in the framework of Word Sense Disambiguation (WSD). However, in order to face WSD in free running text, we should handle MWEs.

The traditional approach to deal with MWEs has been searching for the longest word-sequence match. An exception can be found in the work of Kenneth C. Litkowski. His research on both Question-Answering [3] and Word-Sense Disambiguation [4] explores the idea of inflection in MWEs, even if just by reducing inflected forms to their root forms. Other works have aimed, for instance, at automatically generating MWEs based on some knowledge source. This is the case of Aline Villavicencio's work [5], where she uses regular patterns to productively generate Verb-Particle Constructions.

In the current work, we have gone further in our MWE detection and selection by lemmatizing our MWEs and allowing some inflection of their subparts (cf. section 3). Bearing in mind that our final goal is the WSD of free running text, where segmentation of word units will not be provided, we have applied our treatment of MWEs to a real NLP task: the Senseval-3 Word-Sense Disambiguation of WordNet glosses. The system here described has participated in the Senseval-3 task achieving the third best results. Further, the same WSD system but using the gold tokenization of the solution (including MWEs) has obtained the best scores in the competition.

[1] This work is suported by the European Comision (MEANING IST-2001-34460)