Two Web-based Approaches for Noun Sense Disambiguation

Paolo Rosso[1], Manuel Montes-y-Gómez[2,1], Davide Buscaldi[3], Aarón Pancardo-Rodrǵuez[2], and Luis Villaseñor Pineda[2]

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

Abstract. The problem of the resolution of the lexical ambiguity seems to be stuck because of the knowledge acquisition bottleneck. Therefore, it is worthwhile to investigate the possibility of using the Web as a lexical resource. This paper explores two attempts of using Web counts collected through a search engine. The first approach calculates the hits of each possible synonym of the noun to disambiguate together with the nouns of the context. In the second approach the disambiguation of a noun uses a modifier adjective as supporting evidence. A better precision than the baseline was obtained using adjective-noun pairs, even if with a low recall. A comprehensive set of weighting formulae for combining Web counts was investigated in order to give a complete picture of what are the various possibilities, and what are the formulae that work best. The comparison across different search engines was also useful: Web counts, and consequently disambiguation results, were almost identical. Moreover, the Web seems to be more effective than the WordNet Domains lexical resource if integrated rather than stand-alone.

1 Introduction

The problem of the resolution of the lexical ambiguity that appears when a given word in a context has several different meanings is commonly referred as Word Sense Disambiguation (WSD). The state of the art of WSD [15] shows that the supervised paradigm is the most efficient. However, due to the lack of big sense tagged corpora (and the difficulty of manually creating them), the unsupervised paradigm tries to avoid, or at least to reduce, the knowledge acquisition problem the supervised methods have to deal with. In fact, unsupervised methods do not need any learning process and they use only a lexical resouce (e.g. WordNet) to carry out the word sense disambiguation task [1] [16] [17] [19].