Crossing Parallel Corpora and Multilingual Lexical Databases for WSD

Alfio Massimiliano Gliozzo, Marcello Ranieri, and Carlo Strapparava

ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica, I-38050 Trento, ITALY
{gliozzo,ranieri,strappa}@itc.it

1 Introduction

Word Sense Disambiguation (WSD) is the task of selecting the correct sense of a word in a context from a sense repository. Typically, WSD is approached as a supervised classification task to get state-of-the-art performance (e.g. [1]), and thus a large amount of sense-tagged examples for each sense of the word is needed, according to the word-expert approach. This requirement makes the supervised approach unfeasible for "all-words" tasks, consisting on disambiguating all the words in texts. This problem has been called the Knowledge Acquisition Bottleneck and many solutions have been proposed for it (see for example [2]) .

In this paper we propose the use of aligned corpora and multilingual lexical databases to automatically acquire sense tagged data, exploiting the polisemic differential between two (or more) languages.

Even though the underlying idea of the approach proposed in this paper is not totally original in the WSD literature (see for example [3, 4]) our basic contribution is to show how far we can go in using parallel corpora to collect sense tagged data, by reporting both a quantitative and a qualitative evaluation. It will be shown that having an "ideal" aligned wordnet (i.e. a lexical resource such that all the sense distinctions in one language are reflected in the other), our simple strategy allows to disambiguate 51% of the English/Italian aligned pairs of words with 100% precision, while with the available resources this figures decreases to 67% precision for a subset of 40% words. In the rest of the paper we will evaluate this technique by exploiting two resources recently developed at ITC-irst: MultiWordNet and MultiSemCor.

2 MultiWordNet and MultiSemCor

MultiWordNet (http://multiwordnet.itc.it) is a multilingual computational lexicon, conceived to be strictly aligned with the Princeton WordNet. In our experiment we used the English and the Italian components. The last version of the Italian WordNet contains around 58,000 Italian word senses and 41,500 lemmas organized into 32,700 synsets aligned whenever possible with WordNet English synsets.

The MultiSemCor [5] (http://multisemcor.itc.it) corpus originates from the Princeton SemCor corpus. SemCor texts were taken from the Brown Corpus, and