Natural Language
Abstract
The study of word co-occurrences is common in linguistics but only recently the availability of on-line corpora and dictionaries made it possible to extensively collect word associations. Several papers published on this topic claim that their results are useful not only for lexicographers and linguists, but also for NLP, particularly for semantic and syntactic disambiguation in sentence analysis, as well as for lexical choice in generation. However such claims have not been convincingly proved so far. In this paper we argue that word associations derived through pure statistical analysis can hardly cope with the problem of syntactic disambiguation. It is shown that better performances are obtained by integrating wide-coverage techniques such as statistics with traditional NLP methods.

1. Introduction

The challenge for next generation NLP systems is to be able to process texts which are not restricted to a few hundred words. A wider diffusion of AI-based text processing techniques depends upon the ability of current systems to improve their robustness, in particular for what concerns knowledge-intensive, low-coverage techniques such as semantic analysis. In this paper we address the problem of automatically extracting from corpora data which can be useful for syntactic disambiguation.

In traditional NLP systems the disambiguation of compounds, prepositional phrases (PPs) and conjunctions is handled by specifying the selectional restrictions to which words must obey. Hence for example in:

\textit{John goes to Boston by bus}

the PP "by bus" is attached to the verb "go" rather than to "to Boston", because somewhere in the semantic knowledge base it is specified that:

\texttt{[GO]->(INSTRUMENT)->[VEHICLE]}

The task of hand-encoding selectional restrictions is a bottleneck in NLP systems, because it is very time-consuming and it is hard to keep consistency among the data when the lexicon has several hundred or thousand words. The limitations of AI techniques for extensive text processing has been recently stressed in many important linguistic forums [COL 1990],[ACL 1990]. A new tendency is to automatically acquire lexical knowledge from on-line dictionaries and large corpora using statistical methods. However, pure statistics is at the other extreme with respect to pure knowledge-based