TITLE: A FRAMEWORK FOR THE ACQUISITION OF LEXICAL KNOWLEDGE; DESCRIPTION AND APPLICATIONS

AUTHOR: DIMITRIOS KOKKINAKIS .

ABSTRACT

Lexical knowledge acquisition is the task that (a human) or machine performs when lexical information for a textual unit, such as a word or a phrase, is extracted from an information source, such as machine-readable corpora or machine-readable dictionaries (MRDs), and is intended for use in a language processing system, or for the enhancement of dictionaries for human use. Lexical acquisition is a challenging and complex topic which has attracted a lot of attention in the natural language processing (NLP) community. The field is broad and includes a variety of lexical acquisition problems. A few representative examples (though some overlap) include: acquisition of subcategorisation information; thesaurus/ontology construction and enhancement; acquisition of collocations and lexical patterns for eliminating ambiguities in NLP; determining whether (new) lexical entries with associated lexical information can be added in NLP applications; means of refining and extending lexicographic descriptions; accessibility and re-use of MRDs as a basis for formal ones; computer production of lexicons for human use and terminology extraction.

This thesis addresses a few aspects of lexical acquisition. Emphasis is placed on five areas: (i) validation and extraction of subcategorisation information; (ii) ontology enhancement; (iii) acquisition of lexical knowledge for disambiguation tasks; (iv) means of alleviating the task of hand-coding and extending lexical resources for the benefit of computational lexicons intended for NLP, particularly information extraction; and (v) enhancement of lexicographic productivity in terms of devising means of semi-automatically identifying novel usage of words. This primordial need for lexical information cannot be satisfied without the development of techniques and methods for analysing language data (texts, dictionaries) and abstracting away surface differences in texts: tokenisation, lemmatisation, assigning annotations, pattern recognition and parsing. Accordingly, part of the thesis will describe the development, sometimes adaptation, of existing techniques and software that process natural language into Swedish data for the benefit of lexical acquisition, using a linguistic framework entitled LEXAQ.

KEYWORDS: (empirical) natural language processing, lexical acquisition, information extraction, corpus-based lexicography, lexical engineering, extraction patterns, parsing, lexical annotations, corpus processing, subcategorization, written Swedish, disambiguation.

Göteborg university
SE-405 30 Göteborg, Sweden