Abstract

Universal Semantic Language (USL) is a semi-formalized approach for the description of knowledge (a knowledge representation tool). The idea of USL was introduced by Vladimir Smetacek in the system called SEMAN which was used for keyword extraction tasks in the former Information centre of the Czechoslovak Republic. However due to the dissolution of the centre in early 90's, the system has been lost.

This thesis reintroduces the idea of USL in a new context of quantitative content analysis. First we introduce the historical background and the problems of semantics and knowledge representation, semes, semantic fields, semantic primes and universals. The basic methodology of content analysis studies is illustrated on the example of three content analysis tools and we describe the architecture of a new system. The application was built specifically for USL discovery but it can work also in the context of classical content analysis. It contains Natural Language Processing (NLP) components and employs the algorithm for collocation discovery adapted for the case of cooccurences search between semantic annotations.

The software is evaluated by comparing its pattern matching mechanism against another existing and established extractor. The semantic translation mechanism is evaluated in the task of automated document classification with special attention to the problem of semantic ambiguity and correct translation. Finally we evaluate the ability of the system to discover statistically significant semantic relationships from textual corpora.