Classification Method for Learning Morpheme Analysis

László Kovács, Department of Information Technology, University of Miskolc, Miskolc City, Hungary

ABSTRACT

The morpheme analysis module is an important component in natural language processing engines. The parser modules are usually based on rule systems created by human experts. In the paper, a novel approach is tested for implementation of the morpheme analyzer module. The proposed structure is based on the theory of formal concept analysis. The word inflection can be considered as a classification problem, where the class label denotes the corresponding transformation rule. The main benefit of the proposed method is the efficient generalization feature. The proposed morpheme analyzer module was implemented in a prototype question generation application.

Keywords: Formal Concept Analysis, Human Machine Interface, Morpheme Analysis, Natural Language Processing, Parser Modules

INTRODUCTION

Natural Language Processing (NLP) is an active area within human-machine interface development. The processing of input sentences given in human language or generating sentences of human language is still challenging task in IT world. There are many problem areas in NLP where no standard solutions are available for every related task. The input sentences are processed in many different phases, where the usual process includes tokenization, cleaning, morpheme analysis, sentence analysis, semantic graph construction and sentence interpretation. The goal of the morpheme analysis module is to determine the stem of the word and to determine the grammatical role of the word within the sentence. The stem can be used to determine the concept related to the given word. Using some external ontology, the domain specific and universal knowledge elements can be extracted from the related external knowledge base. The ontology databases usually contain information on the specific relationships of the concepts like specialization, generalization, synonyms and specific application. The grammatical role of the words can be encoded on many ways. In some languages, the position of the word conveys the grammatical role. In some other languages, there is no dominant word order, thus other formal elements, like suffixes or prefixes are used to describe the role of the word. As a word may have several grammatical and semantic roles at the same time, several suffix or prefix parts can be attached to the stem word. The main goal of the morpheme analyzer module is to
determine both the different suffix and prefix layers and the stem word.

In the literature there are some standard methods for morpheme analysis which use some rule based systems. These rules are usually created by human experts, thus the generation of the rule set is always a very costly operation. The main goal of our investigation was to investigate the possibility of a learning system which can inference the morpheme structure of the target words. This task has a high complexity as it has a lot of unknown parameters like the set of suffix and prefix elements and the agglutination rules of the morpheme elements.

In this paper, the first phase of the research is summarized which aims at the generation and testing a concept lattice based morpheme analyzer. The proposed system uses a supervised learning mechanism. The training data should contain valid inflection examples: a transaction unit includes the base word, the inflected word and the corresponding morpheme structure. Thus the set of suffixes and prefixes are given as an input parameter. The goal of the concept lattice based classifier is to learn the relationship between the stem form and the corresponding transformation rule. In the proposed system for every possible grammatical roles (for example accusative), a separate concept lattice classifier is generated. Thus the resulting structure is the cluster of classifiers. The possible ordering of the different morpheme units is encoded with a probabilistic finite state automaton. The edges, the transition edges of the automaton are set during the training process. The classification is executed with the application of a concept lattice. The concept lattice is a very flexible structure to determine the most important clusters of the attributes and determine the generalization relationship among them. Using a special, class label attribute in the intent part, the lattice can be used as a classification tool. The main benefit of the concept lattice based classification is that it uses a human-like generalization mechanism. The performed test focused on this property of the classification. The tests were executed with smaller training sets in order to investigate the generalization accuracy of the different morpheme classifiers.

The paper first gives a survey on the internal structure of the NLP engines and it presents the key modules of the engine. The next section presents an overview of the different important stemmer and morpheme methods. Then the formal definition of the concept lattice structure is given and the proposed architecture for concept lattice based classification is introduced. The last section presents a prototype system for automated question generation task. The question generation application uses the proposed morpheme analysis module to determine the stems in the source sentences.

Structure of NLP Engines

Intelligent human-machine interface (HMI) is a key characteristic of future information systems. The umbrella term of HMI covers a wide area of tools and methods. The traditional methods include among others the graphical interface, the tactile (touch) technology and the speech to command interfaces. Among the new technologies, the bionic technology (Cannan et al., 2010) is an emerging new area, where the proposed Skinput system combines the natural conduction properties of the human body to use the skin as user input device. The survey in (Fakhreddine et al., 2008) presents a detailed list of input technologies, like the pointing devices or switch-based devices. The keyboard belongs for example in the family of switch devices. The mice and touch-screen panels are members of the pointing inputs while joystick belongs to both categories. Another important characteristic of current HMI solutions is application of new approaches. The term ambient intelligence (Fakhreddine et al., 2008) refers to the ultimate methods of human-computer interaction that is the deletion of a desktop and embedding of the computer in the environment so that it becomes invisible to humans while surrounding them everywhere hence the term ambient. An important component of ambient technology is the application of spoken dialog system (SDS) which rely on speech recognition processes. This kind of method requires the highest complexity as it is based on the semantic analysis of the input data stream.