Software Package for Supporting Information Extraction Research

Zoltán Alexin[1], Tibor Gyimóthy[1], János Csirik[1]

University of Szeged, Department of Informatics,
Árpád tér 2., Szeged, Hungary
e-mail:{alexin,gyimothy,csirik}@inf.u-szeged.hu

Keywords: Information Extraction, Natural Language Processing, Shallow Syntactic Parsing

The research of the technology for IE (Information Extraction) is a dynamically emerging field of NLP (Natural Language Processing). Collecting the relevant information by computers from the vast amount of texts appearing on the Internet and providing it in brief form is a daily need in politics, economy, science, and even in intelligence services. While IR (Information Retrieval), which is one of the characteristic features of web browsers, aims to present the needed documents in original form to the users, the task of IE includes marking and then collecting the relevant information from the texts as well. Hence text compression and IE are closely related.

IE systems do not intend to fully understand or analyze the documents in detail. The main requirements set against them are big capacity, speed, and an acceptable accuracy. They are usually satisfied with identification of major actors in the sentences without complete syntactic parsing. To accomplish this they do shallow parsing. In the identification of actors, *named entities* play important role. Recognizing person names, companies, geographic locations, cities frequently cited in newspapers is done in a separate processing step based on a large lexicon independently from morphological parsing.

In this paper a software package is presented, which is developed at the Department of Informatics at the University of Szeged for supporting IE research. The most important design concept of it was modularity, so that individual components can be developed independently. Modules can be run separately or in a batch. The output of each module can be followed up and be checked. These features are important at the beginning of the research phase, when different experimental approaches are tried. The standardized communication between two subsequent modules eases changing one module to another and so selecting the best one for a specific task.

The presented system was applied for processing short business news. The MTI-Eco, Business+ service [1] has been used to create a database of 6453 articles for training and testing purposes. Most development efforts are directed to modules being in key positions of the processing, such as POS-taggers, shallow syntactic parsers and semantic pattern (semantic-frame) recognizers.

[1] Hungarian News Agency (Magyar Távírati Iroda), `http://www.mti.hu`