Information Extraction from Short Business News Items

Gábor Prószéky

MorphoLogic
Budapest
proszeky@morphologic.hu

This paper presents a research project aimed at the extraction of structured information from business news items. Content analysis and extraction have traditional methods, mainly for the English language. As the texts in this project are in Hungarian, and the goal is to extract information to the maximum possible extent. Information extraction (IE) does not require deep parsing of the input. However, a content analysis system working with business news items must be able to identify essential business processes and activities, along with the participants, dates/times, and amounts of money involved. This, in our view, requires more than shallow parsing. Moreover, the texts – short business news items – have many extra-linguistic elements such as dates, times and currency amounts that must be properly interpreted within the textual context. The main goal of the project is to develop a system to analyse Hungarian-language business news items of up to 100 words, and provide structured data of business actions and processes. This will serve as a linguistic pre-processor for the purposes of a query system capable of providing information on actions of a particular company or institution, overall market trends etc. Dates, times, currency amounts must be treated numerically, and processed in the proper context. In order to keep the development process closely related to real-life applications, one of the members of the project consortium is a leading polling institute that validates and tests the content analysis scheme within their own applications.

The research project started in September 2001, and focuses on Hungarian-language news items. This type of texts presents three crucial problems that must be addressed using HLT modules:

(1) Besides the large complexity of the Hungarian morphology, the texts incorporate multiple levels of indirection when referring to an actor (a company, a person, a product, an authority etc.). This phenomenon is specific to the style of Hungarian-language business news items. Therefore, proper NP chunking is difficult, because highest level NP's tend to be very long (often over 10 words), and its internal structure must be revealed to some extent to correctly spot the head of the NP – which can be a multi-word named entity itself.

(2) For the purposes of this application, it is not sufficient to spot named entities: they must be properly identified and mapped to a database entity in order to resolve references. People must be identified either by their job titles or their names; multiple names of the same company must also be recognized. The project members agreed that time references and other numbers must be treated like named entities. Moreover, relative time references must be resolved to absolute dates, based on the release date of the news item under analysis. Here the conclusion is that a sophisticated named entity resolver (NER) module must be applied.

(3) Actions, subjects and objects, and their relationship must be properly parsed as each news item must be described by means of one or more strongly typed event description, which is then suitable to be included in a database. This implies that Hungarian verbal frames must be recognized by the system, as well as the relationship between multiple predicates within the news item, each represented as a single verbal frame.

Here the difficulty lies with the lack of a suitable grammar description of Hungarian. Therefore, a significant effort is required to perform basic research on the Hungarian grammar – all as part of this IE project.

Part of the problem is that the text of news items tends to be ill-formed. Thus the grammar model for the IE application will not be a 'pure' description for the Hungarian grammar, as content must still be extracted when the text does not conform to Hungarian spelling rules or grammar/style conventions. Here it is not necessary to spot the errors: the system must only return extracted information, 'pretending' the input was correct.