BATMULT Fellowship Report

Pavel Vondřička

January 4 to October 8, 2006

Project description

The research project aims on lexical description of the Norwegian nouns, both monolingually and in a contrastive analysis with other languages. The background of this research is the desirability of reusable lexical resources supporting natural language processing (NLP). While tradional printed dictionaries are large and detailed, a lot of information contained in them is implicit, cannot easily be digested by computers, and does not reflect all the contextual information of the words. The goal of the PhD project is to examine the possibilities of a structured and explicit lexical description and its possible utility as a resource in NLP. The concrete goal of the fellowship was to explore methods for representing highly structured lexical information about Norwegian nouns, such that it is applicable in NLP contexts.

In the context of the project, a system was designed and implemented which is more general than language specific dictionary editors, but less general than formal systems used in NLP. This system has been tested on the Norwegian nouns. The system is based on the ideas of object oriented description with default inheritance, feature structures and unification, but with its own rules for interpretation of the data description, which is in many aspects bound to a specific language or language type. The goal of the system is to provide a description of the morphological structure (and syntactic structure in the case of multi-word entries) of the lemmas together with information about their syntactic behavior, collocability and lexical semantic relations of their senses. In addition, the system was designed to handle wide language variability and detailed usage attributes for different variants – important factors that are often underestimated in human oriented dictionaries and almost completely ignored by most NLP projects.

Different parts of the lexical description (morphemes, words and multi-word expressions, valency frames and their slots, usage labels, semantic relations and translational equivalency, etc.) are represented by objects and language or theory specific functions used to connect them. A separate language module takes care for interpretation of the data and reconstruction of the list with valid word forms and their features and for formation of a human readable output or output in XML format.

The BATMULT training site and AKSIS institute have great experience in both NLP, building of lexical resources and building and processing of multilingual corpora. They also provide all necessary tools needed to prepare Norwegian text for the parallel corpus.

Activities

- Possible resources concerning nouns in Norwegian and their use were considered. The most useful general description and classification was found in Norwegian Reference Grammar *(Norsk referansegrammatikk)*, and different mono- and bilingual dictionaries were used for examples and comparison. From local resources, the SCARRIE database was used to create a stylistic database of word forms. The morphological database used in LOGON project was not found suitable as resource for a humane oriented dictionary, because of use of a custom inflectional classification. More advanced methods of automatic data extraction are not accessible at the moment because of lack of higher-quality resources like word-aligned parallel corpora and parallel treebanks.
- An implementation of the proposed database has been developed using the object-oriented

web-application framework *Ruby on Rails* and *MySQL* database and continually tested with different types of problematic examples from the morphology (inflexion and word-formation), syntactic and collocational patterns and semantic and translational relations, with regard to their variability and stylistic features. Different solutions for description of problematic issues were considered and compared. The possibility to interpret the resulting description to compute valid word forms and to reconstruct graphical output resembling a printed dictionary entry was proven.

- Quantitative statistics about the distribution of alternative word forms (orthographic and inflectional variants) in Norwegian were collected for the stylistic database created from SCARRIE database by parsing three different Norwegian corpora: the local *Aviskorpus* and *Talemålskorpus*, and the *Bokmålskorpus,* created at the University in Oslo.
- Texts for parallel Norwegian-Czech corpus have been collected both from the publishers and by scanning books from the university library.
- Tools for tagging and alignment of texts, developed at Aksis, were tested on the new texts and improved; new complementary tools for processing of large XML texts with the Oslo-Bergen tagger were created.
- A summary of the results and problems with description of nouns at different levels was written.

Results

- Lexical database implementation: a testing implementation of the proposed database system has been developed using object-oriented web-application framework *Ruby on Rails* and *MySQL* database
- Analysis of Norwegian nouns: an analysis of problems in the description of Norwegian nouns was carried out on the basis of the Norwegian Reference Grammar, corpus data and local NLP projects
- Experimental examples: a basic structure of templates and simple example entries was created in the lexical database for both typical and irregular nouns, describing their inflexion, word-formation, syntactic and collocational patterns and semantic and translational relations, with regard to their variability and stylistic features
- Interpretation of the examples: interpretational methods have been implemented in the database to prove the ability to generate basic resources both for NLP and human readable dictionaries from the description
- Statistics about usage of variant noun forms: a database of Norwegian variant noun forms was extracted from the SCARRIE database and statistics about the distribution of the forms were collected from three different corpora
- Improvement of tools and methods: methods for tagging of Norwegian texts and their alignment to Czech texts have been developed, current tools (Oslo-Bergen Taggers SOAP interface, TCA2 aligner) were tested and consequently significantly improved by their developers at Aksis; new complementary tools were developed
- New texts for parallel corpus: new texts for the parallel Norwegian-Czech corpus were acquired both from the publishers and by scanning printed books; these texts will be used for further research on the bilingual equivalency
- Base for the thesis: a summary of the theoretical and practical results was written as a basis for the thesis and a possible scientific paper
- Ongoing work: agreements about possible cooperation on futher building of the parallel Norwegian-Czech corpus were achieved; a discussion has started about further proofs of the database system and its usability