Project update

DiET – Diagnostic and Evaluation Tools for NLP Applications

Klaus Netter & Tillmann Wegst, DFKI, Saarbrücken

The main goal of the DiET project, which drew to an end after two years in March 1999, was to provide tools to support the development and application of linguistic data for the evaluation and diagnostics of human language applications, as well as to produce such data in English, French, and German. The project was partly based on the experiences of the TSNLP project, which had led to the insight that the maintenance of test data requires the support of sophisticated tools, and that the value of artificially constructed test data can be considerably increased if such data can be customised to specific applications and domains.

DiET brought together partners from industry and from independent research organisations with different backgrounds and interests. On the user side, these were **Aerospatiale** France, **IBM** Germany, and the **Localisation Resources Centre** in Dublin. The development was carried out by the (coordinating) **Language Technology Lab** of **DFKI** in Saarbrücken, **ISSCO** in Geneva, **SRI** Cambridge, and **IBM** in Heidelberg.

The major results of the project are a core system for the construction, annotation and maintenance of test data, customisation tools for text profiling and lexical replacement, and test data at different levels of linguistic abstraction.

The *annotation system* consists of Java clients equipped with a graphical user interface for the annotation, configuration and retrieval of linguistic data, and a central SQL database server. The system emphasises openness and flexibility by offering a wide range of options for configuring test data and annotations, and the relations between them. Thus, users can freely design annotation types for their particular purposes. They can draw upon a set of annotation modes, each providing display, editing, storage and control functions for one of six data types (from boolean, strings, integral and float numbers, to trees and text markings). Some of the annotation modes are quite sophisticated. So, for example, there is a fully configurable tree annotation tool, which allows to annotate both nodes and edges, thus providing facilities for both phrase structural and relational markup. Similarly, a marking tool allows the user to freely markup substrings in a text, e.g., for specifying anaphoric or other discourse relations. Annotation values can be entered freely and/or chosen from lists, in the case of numbers constrained to ranges, and preset with defaults. Annotations may be entered not only by the user, but also by external services: a protocol was designed which was proven to allow remote servers (such as POS taggers and morphological analysers) to be plugged in easily, which allows for a dynamic extension of the DiET system and its power. Annotation types are organised in hierarchical schemata, and types can be shared across these schemata. Test items are grouped into test suites, which can also be reorganised through the sharing and copying of test items. All of the data may be exported and imported to or from files to allow data interchange with other systems, and to expose the data to external processing.

Test items, which may have been constructed for a different or no specific domain, are often not suitable by themselves for the evaluation of applications which are tuned and adapted for specific domains. To address this issue, *customisation tools* were developed in DiET, which allow the user to adapt her data to a specific domain, either by means of text profiling or through lexical replacement. The *text profiler* is a complex system which can be run on a stand-alone basis or as a service to DiET. It allows the user to derive a general text profile of a corpus, including a broad range of statistical parameters, and it can establish links between phenomena and their occurrence in a corpus. For the latter purpose the user designs a search pattern in a GUI and submits the query, upon which hits are returned and automatically imported into the DiET database. Users can thus combine the manual construction of controlled, systematic and non-redundant test items with real-life representative data from domain-specific corpora. *Lexical replacement* allows for the derivation of new items from existing ones on the basis of lexical equivalence classes, defined at different levels of linguistic abstraction.

In terms of *linguistic test data***,** the languages English, French, and German were covered by items representing phenomena at three different levels of linguistic abstraction (morphology, syntax, and discourse). These data, altogether some 18,000 items, are all well-structured and decorated with detailed annotation schemata comprising some 250,000 values. For *morphology*, DiET provides a common reference standard in the form of equivalence or ambiguity classes on the basis of which different morphological components can easily be compared to each other. For *syntax*, the test suites consist of systematic grammatical and ungrammatical variations over most of the relevant syntactic phenomena. Most of these data were based on TSNLP data, but they were thoroughly revised and associated with new annotation schemata making use of the new facilities, such as the tree display and editing tool. The development of *discourse* data opened up an entirely new field of test data illustrating mainly elliptical and anaphoric relations.

Altogether, DiET provides a highly versatile, sophisticated and extensible environment for the annotation and maintenance of linguistic data. The range of functions provided by DiET includes many of the commonly expected options, such as complex tree annotations, but also goes well beyond such individual tools by means of seamless integration into a user-friendly and transparent system. The developers of DiET have a great interest in supporting and extending this software package further, and are open to making it available to interested institutions.

(Information box)

Klaus Netter (netter@dfki.de, www.dfki.de/~netter/) is the associate head of the Language Technology lab at DFKI GmbH Saarbrücken. His current main research interests are in Multilingual Information Management, HLT Evaluation, and development of linguistic resources.

Tillmann Wegst (wegst@dfki.de, www.dfki.de/~wegst/) is a senior software engineer at DFKI GmbH Saarbücken. His research interests are in the areas of fuzzy technology and neural grammars.

For more on DiET: diet.dfki.de