The DEFI Project

Archibal MICHIELS

DEFI is a five-year project, employing the equivalent of three full-time researchers. It started in October 1995 and aims at providing a prototype of an online reading comprehension tool. Being a research project, it is not end-user-oriented, and is not limited by the constraints (computational efficiency, uptodateness and elegance of the user interface) which would characterise efforts geared towards the production of a commercial tool within the same time boundaries. It involves two languages, viz. English and French, and two directions, both languages playing the part of source and target language. The French to English direction is likely to lag behind a bit on account of the lesser availability of NLP analysis tools for French on our development platform (Windows). DEFI is meant to act as a filter on a bilingual dictionary (a merge of the Oxford/Hachette (OH) and Robert/Collins (RC) English-French and French-English bilinguals) to provide the user with the most likely translation(s) of the item he has requested help about.

The tasks involved are the following:

- recognition of terminological units, be they single-word or multi-word. The test bed is Mycenaean archeology, but the recognition strategies are designed to be domain-independent;

✉ A. Michiels, English Dpt, University of Liège, 3, pl. Cockerill, B-4000 Liège, Belgium
amichiels@ulg.ac.be
http://engdep1.philo.ulg.ac.be/michiels

- recognition of general language multi-word units (mwu's) stored in the bilingual dictionaries. This task also includes the presentation to the user of relevant dictionary examples, because the concept of mwu is extended here to cover examples as selected and/or edited by lexicographers;
- for both multi-word units and single word lexical items, restriction of the range of translations, such restriction to be based on properties of the source text, i.e. the textual environment of the item the user has asked to get the translation of. In the best of cases, the translation that ranks highest according to the DEFI matcher is the one that is most appropriate to the context.

The first two of these three tasks are essential to the quality of the tool. Although not all mwu's are monosemic (far from it), the first help to provide the user who has requested the translation of a word belonging to an mwu, is that mwu and its translation(s). Restricting the range of translations is again a matter of matching the context with the constraints or preferences that the dictionary associates with the source item under a specific translation. The recognition of terminological units and mwu's should not be dependent in any way on the particular item within the mwu that the user has selected. It would be a great pity for an electronic dictionary tool not to be free of physical storage considerations or to be dependent on a specific indexing scheme which would force the user to guess which word the mwu is most likely to be stored under (assuming the user is able to recognize that there is an mwu in the string he is interested in).

These three tasks are carried out by a dictionary/text matcher implemented as a Prolog program. This program has access to binary trees stored in Prolog idb's (internal data bases). Such binary trees are constructed on the basis of:

- two bilingual dictionaries (OH and RC) with reformatting into Prolog clauses (and later merging of the two dictionaries) accomplished by AWK programs
- a thesaurus data base, namely part of the Prolog *WordNet* package as distributed by Princeton University (see MILLER: 1990), with

slight editing and adaptation of the accompanying application programs (query tools)

- a data base of metalinguistic descriptions derived from RC and OH, to allow the measure of semantic distance in terms of metalinguistic slot sharing (cf. MONTEMAGNI *et al.*: 1996)

The tools used in the project are the following:

- the AWK programming language
- Arity Prolog for Windows
- LingSoft's *engcg* surface parser for English (for French we plan to develop local grammars in the INTEX formalism, in a first stage geared towards the texts used as test beds)

DEFI's strategies are based on the following observations:

- Mwu's (and examples) in the dictionaries are simply strings, and therefore do not provide enough information for the DEFI matcher to work on if it is to spot mwu's in text in other than their canonical forms. But we do not think it is reasonable to expect specific local grammatical descriptions to become available for every such string. We believe that the more reasonable solution is to apply a robust surface parser to BOTH dictionary mwu's and examples on the one hand and the user's text on the other and to let the matcher loose on the enriched dictionary/text pair. Thereby we avoid the need to design equivalence-measuring procedures for diverging grammatical information, as is bound to happen when we confront specific local descriptions in the dictionary with a tagged or fully parsed text.
- We believe that bilingual dictionaries do not provide enough information on their source side, and that we have to harness other lexical resources, such as thesauri (especially helpful in the treatment of collocational information, where collocates are very often to be interpreted as thesauric heads rather than as lemmas). We think that monolingual dictionaries, especially recent learner-oriented dictionaries, can help to enhance information on the source language, although pairing the source side of a bilingual with a monolingual is a major task, and a delicate one, in so far as

the semantic space of a source item is divided in a bilingual dictionary according to the semantic space divisions enforced by the target language

- DEFI is a research program, and therefore we can envisage wider avenues than we would be able to if a commercial product had to be delivered at the end of the project. This is not to say that computational efficiency can be neglected: no serious testing and debugging can take place when the system is too slow. But it means that we can think of developing a prototype tool in such a high level language as Prolog, making use of its automatic backtracking procedures and built-in unification.

Our Web site[1] offers the following papers on the DEFI project:

1. general presentation of the project: this paper;
2. turning bilingual dictionaries as provided by the publishers into something that DEFI can use, viz. Prolog clauses ready to be stored in binary trees (DUFOUR (Nicolas): "From the Oxford Hachette SGML tape to DEFI dictionaries");
3. parsing text and dictionaries and enhancing the results by means of AWK programs (MICHIELS (Archibal) and DUFOUR (Nicolas): "From SGML tape to Dic Clauses");
4. general description of the dictionary/text matcher (MICHIELS (Archibal): "The DEFI matcher: an introduction");
5. the full commented code of the part of the matcher taking care of mwu's (MICHIELS (Archibal): "The Defi Matcher: Prolog Program");
6. a preliminary results file, exemplifying mwu recognition and translation selection ("The DEFI matcher: preliminary results");
7. an early DEFI paper coupling a monolingual to a bilingual dictionary for the purposes of translation selection and word sense discrimination (MICHIELS (Archibal): "Target Selection and Word Sense Discrimination in DEFI").

[1] http://engdep1.philo.ulg.ac.be/michiels/defi.htm.

References

Dictionaries and thesauri

CIDE = PROCTER (Paul), ed.: 1995, *Cambridge International Dictionary of English* (CUP) , first ed.

COBUILD = SINCLAIR (John), ed.: 1987, *Collins Cobuild English Dictionary* (Collins), first ed.

LDOCE = PROCTER (Paul), ed.: 1979, *The Longman Dictionary of Contemporary English*, first ed.

OH = CORRÉARD (M.H.) and Grundy (V.), eds.: 1994, *The Oxford-Hachette French Dictionary* (Oxford: OUP).

RC = ATKINS (Beryl T.) *et al.*, eds: 1995, *Collins-Robert French/English English/French Dictionary* (4th edition, Glasgow: HarperCollins).

WordNet = WordNet Prolog Package, downloadable from the Princeton University WWW site. See also MILLER: 1990.

Tools

The surface parser: ENGCG was developed at the General Linguistics Department of the University of Helsinki. It is marketed by Lingsoft Inc. (http://www.lingsoft.fi).

Awk: MKS and Thompson implementations for Windows and associated documentation; see also AHO *et al.*: 1988.

Prolog: Arity implementation for Windows; Arity Corporation, Damonmill Square, Concord, Mass.

Intex: as distributed by Max Silberztein for NextStep; cf. SILBERZTEIN: 1993.

Other references

AHO (Alfred V.), Kernighan (Brian W.), Weinberger (Peter J.): 1988, *The AWK Programming Language* (Reading, Mass.: Addison-Wesley).

MILLER (G.A.): 1990, eds., *WordNet: An On-Line Lexical Database, International Journal of Lexicography, 3*, 4.

MONTEMAGNI (S.), FEDERICI (S.) and PIRRELLI (V.): 1996, *Example-based word sense disambiguation: a paradigm-driven approach*, paper read at the 1996 Euralex Congress.

SILBERZTEIN (Max): 1993, *Dictionnaires électroniques et analyse automatique de textes* (Paris: Masson).