Industrial Applications of Unification Morphology

Gábor Prószéky
MorphoLogic
Fő u. 56-58. I/3.
H–1011 Budapest, Hungary
h6109pro@ella.hu

Abstract

Industrial applications of a reversible, string-based, unification approach called Humor (High-speed Unification Morphology) is introduced in the paper. It has been used for creating a variety of proofing tools and dictionaries, like spelling checkers, hyphenators, lemmatizers, inflectional thesauri, intelligent bi-lingual dictionaries and, of course, full morphological analysis and synthesis. The first industrialized versions of all of the above modules work and licensed by well-known software companies for their products' Hungarian versions. Development of the same modules for other agglutinative (e.g. Turkish, Estonian) and other (highly) inflectional languages (e.g. Polish, French, German) have also begun.

1 Supported Morphological Processes

1.1 Morphological Analysis/Synthesis and Lemmatizing

The morphological analyser is the kernel module of the system: almost all of the applications derived from Humor based on it. It provides all the possible segmentations of the word-form in question covering inflections, derivations, prefixations, compounding and creating basic lexical forms of the stems.

Morphological synthesis is based on analysis, that is, all the possible morphemic combinations built by the core synthesis module are filtered by the analyzer.

Lemmatizer is a simplified version of the morphological analysis system. It provides all the possible lexical stems of a word-form, but does not provide inflectional and derivational information.

1.2 Spelling Checking and Correction

Spelling checking of agglutinative languages cannot be based on simple wordlist based method because of the incredibly high number of possible word-forms of these languages. Algorithmic solutions, that is morphology based applications, are the only way to solve the problem (Solak and Oflazer 1992). The spelling checker based on our unification morphology method provides a logical answer whether the word-form in question can be constructed according to the actual morphological descriptions of the system, or not. In case of negative answer a correction strategy starts to work. It is based on orthographic, morpho-phonological, morphological and lexical properties of the words. This strategy also works in real corpus applications where automatic corrections of some typical mis-typings have to be made.

1.3 Hyphenation

There are languages in which 100% hyphenation cannot be made without exact morphological segmentation of the words. Hungarian is a language of this type: boundaries between prefixes and stems, or between the components of compounds override the main hyphenation rules that cover around 85% of the hyphenation points. Our unification based hyphenator guarantees, in principle, perfect hyphenation (including the critical Hungarian hyphenation of long double consonants where new letters have to be inserted while hyphenated).

1.4 Mono- and Bi-lingual Dictionaries

Besides the above described well-known types of applications there are two new tools based on the same strategy, the inflectional thesaurus called Helyette (Prószéky & Tihanyi 1993), and the series of intelligent bi-lingual dictionaries called MoBiDic. Both are dictionaries with morphological knowledge: Helyette is monolingual, while MoBiDic — as its name suggests[1] — bi-lingual. Having analyzed the input word both systems look for the lemma in the main dictionary. The inflectional thesaurus stores the information encoded in the analyzed affixes, and adds to the synonym chosen by the user. The morphological synthesis module starts to work here, and provides the user with the adequate inflected form

[1] MorphoLogic's Bi-lingual Dictionary

of the word in question. This procedure has a great importance in case of highly inflectional languages.

2 Implementation Details

Humor unification morphology systems have been fully implemented for Hungarian. The same package for Polish, Turkish, German, French are under development. The whole software package is written in standard C using C++ like objects. It runs on any platforms where C compiler can be found.[2]

The Hungarian morphological analyzer which is the largest and most precise implementation needs around 100 Kbytes of core memory and 600 Kbytes disk space for spell-checking and hyphenation (plus 300 Kbytes for full analysis and synthesis). The stem dictionary contains more than 90.000 stems which cover all (approx. 70.000) lexemes of the Concise Explanatory Dictionary of the Hungarian Language. Suffix dictionaries contain all the inflectional suffixes and the productive derivational morphemes of present-day Hungarian. With the help of these dictionaries Humor is able to analyze and/or generate around 2.000.000.000 well-formed Hungarian wordforms. Its speed is between 50 and 100 words/s on an average 40 MHz 386 machine. The whole system can be tuned[3] according to the speed requirements: the needed RAM size can be between 50 and 900 Kbytes.

The synonym system of Helyette contains 40.000 headwords. The first version of the inflectional thesaurus Helyette needs 1.6 Mbytes disk space and runs under MS-Windows. The size of the MoBiDic packages vary depending on the applied terminological collection. E.g. the Hungarian–English Business Dictionary needs 1.8 Mbytes space.[4]

3 Industrial applications

There are several commercially available Humor subsystems for different purposes: lemmatizers, hyphenators, spelling checkers and correctors. They (called HelyesLem, Helyesel and Helyes-e?, respectively) have been built into several word-processing and full-text retrieval systems.

Spelling checkers and hyphenators are available either as a part of Microsoft Word for Windows, Works, Excel, Lotus 1–2–3 and AmiPro, Aldus PageMaker, WordPerfect, etc. or in stand-alone form for DOS, Windows and Macintosh. Microsoft and Lotus licensed the above proofing tool packages for all of their localized Hungarian products.

Humor-based lemmatizers support free text search in Verity's Topic and Oracle, and it is used by the lexicographers of the Institute of Linguistics of the Hungarian Academy of Sciences in their every-day work. That is, the corpus used in creation of *Historical Dictionary of Hungarian* has been lemmatized by tools based on our unification morphology.

Numerous versions of other Humor-based applications run under DOS, OS/2, UNIX and on Macintosh systems.[5]

References

Prószéky, G., Tihanyi, L. A Fast Morphological Analyzer for Lemmatizing Corpora of Agglutinative Languages. In: *Kiefer, F., Kiss, G, & Pajzs, J. (eds.) Papers in Computational Lexicography — COMPLEX 92.* Linguistics Institute, Budapest: 265–278. (1992) Prószéky, G., Tihanyi, L. Helyette: Inflectional Thesaurus for Agglutinative Languages. *Proceedings of the 6th Conference of EACL*, Utrecht: 473. (1993)

Solak, A. and K. Oflazer. Parsing Agglutinative Word Structures and Its Application to Spelling Checking for Turkish. *Proceedings of the COLING–92*, Nantes: 39–45. (1992)

[2]Up to now, DOS, Windows, OS/2, UNIX and Macintosh environments have been tested.

[3]Even by the end-users.

[4]Its language specific and not application specific parts cannot be multiplied if other vocabularies also need Hungarian and/or English.

[5]For OEM partners there is a well-defined API to Humor.