Abstract

The paper describes models for representation and methods to handle lexicographic structures supplied by the MORPHO-2 system. It was built to manage monolingual lexicons and to incorporate lexical processing.

1 Introduction

Most advanced systems for natural language processing use powerful lexicons and morpho-lexical processing environments. The system described below enables monolingual lexicon handling and incorporates morpho-lexical processes (i.e. word-form analysis and synthesis) at lexicon level. At present, it works as a component of the JURES environment for building natural language applications (Tufis and Cristea, 1985).

Since in our approach the morphological processes obey a paradigmatic morphology (Tufis, 1989), word-forms analysis and synthesis take into account only grammatical endings (which includes both desinences and suffixes) and the lexicons handled by MORPHO-2 system are root- or lemma-oriented. By lexicon we don't mean only a collection of roots and associated features. At lexicon level we also encounter the control structures needed for morphological processing: morpho-lexical acquisition menus, root modification rules, word-forms synthesis rules, etc.

MORPHO-2 system is root- or lemma-oriented. By lexicon we mean description attached to a leaf node is represented by means of a morpho-lexical acquisition scenario. A scenario entry (further on referred to as a slot) corresponds to a point of the paradigmatic description space.

2 Morphological Model Design

In order to build the morphological model, an integrated environment which allows editing, viewing and compiling the morphological model description, is available to the linguist.

Defining the morphological model takes place in several steps, during which the linguist has to specify the following:

a) the categories, subcategories, features and their values, in a hierarchical manner
b) the paradigmatic descriptions
c) the default feature specifications associated to each paradigmatic description
d) the lemma - entry correspondence, for each paradigmatic description
e) the inflectional paradigms and root detection rules.

The hierarchical description of features is achieved by correlating several feature specifications. A feature specification is given in the form of a (feature: value) pair. We call a paradigmatic description a hierarchical description build of several simple (feature: value) pairs.

Figure 1 partly presents, in the form of an incomplete tree, the morphological model for the Romanian language. By tree traversal, all paradigmatic descriptions of the model may be generated.

Each non-terminal node contains a single feature specification. The leaf nodes may contain one or more feature specifications. According to this we create several selection criteria, which is applied when visiting a non-terminal node, we can distinguish

1. if a word ends in < inflexion > then
2. < inflexion > : = ( < inflectional-paradigm > < slot-number > )

3. with the following meanings:

a) if a word ends in < inflexion > then
3 Lexical Stock Building

MORPHO-2 lets the lexicographer define new entries in the lexicon by means of a user-friendly window-oriented interface.

A lexicon entry has the following formal structure:

\(<entry> ::= \langle lemma \rangle \ |
\langle paradigmatic-description-selector \rangle \ |
\langle inflectional-paradigm \rangle \ |
\langle root \rangle \ |
\langle morphological-description \rangle \ |
\langle semantic-description \rangle \ |
\langle syntactic-description \rangle \ |
\langle unknown-word \rangle\)

The lexicographer's interface is strictly dependent on the specifications from the linguist's interface since a large part of the former is built automatically from the specifications of the latter.

4 Morpho-Lexical Processing

The target natural language processing system is the beneficiary of the morpho-lexical processes executed by MORPHO-2. Wordforms analysis and synthesis are mediated by a process interface.

In the case of lexical analysis, if the interface is given a sequence of words, it will return a sequence of morpho-lexical atoms. The structure of these atoms is presented below:

\(<\langle root \rangle \rangle \ |
\langle\langle paradigmatic-description-selector \rangle \ |
\langle morphological-description \rangle \ |
\langle\langle syntactic-description \rangle \ |
\langle semantic-description \rangle\)

A morphological description contains both contextual and context-free information. The former is obtained from ending analysis and the latter from the lexicon entry corresponding to the root. The information for the other fields from the atom structure is also taken from the lexicon entry corresponding to the root.

With respect to the result of morphological congruence and root retrieval within the lexicon, we may classify the morpho-lexical atoms as unambiguous, ambiguous and undetermined.

The unambiguous morpho-lexical atoms associate the analyzed word with a single lemma. In the case of a root which corresponds to one lemma and has more possible morphological descriptions, for the same paradigmatic description selector, the system will attempt to compact them.

The ambiguous morpho-lexical atoms come from words to which several lemmas may be attached. The association of a root with several lemmas is possible either due to ambiguity of category (e.g. noun vs. verb) or to apparent homography, generated by the absence of prosodic markers in the Romanian language (mődele, mődele, acěłe, acěłe, modul, mǒdul, etc.).

The possible interpretations are ordered in such a way that those which correspond to shorter roots (that means longer ending) have priority.

The undetermined morpho-lexical atoms correspond to words which have no entry in the lexicon. The atoms generated in this situation have the following structure:

\(\langle\langle unknown-word \rangle \rangle \ |
\langle\langle possible-root \rangle \ |
\langle morphological-description \rangle \ |
\langle syntactic-description \rangle\)

where \(\langle entry-identifier \rangle\) may be a lemma, a root or a semantic description.

We have to point out that previous to morpho-lexical analysis and synthesis, the target processor may configure the structure of morpho-lexical atoms according to the designed application, by means of a communication protocol.

5 Implementation

The MORPHO project, started in 1986, has achieved as a first result, a prototype version now available on a PDP-11 compatible computer. The second version of the system, the one presented in this paper, is implemented in C on an IBM-PC compatible.

The network representation of data and techniques used for implementation, like lexicon indexing using prefixed virtual B³ trees (based on which, for 20000 pseudo-random generated words, several variable length retrieval requires 2 external accesses only), have led to an average response time of lexical processes, quite independent of the lexicons size (for more details on performance analysis see (Tufis and Dumitrescu, 1990)).

References

Dumitrescu, C. MORPHO user manual. I.C.I., Bucharest, 1991
Tufis, D. Cristea D. IURES: A human engineering approach to natural language question answering. Artificial Intelligence: Methodology, Systems, Applications, North Holland, Elsevier, 1985
Tufis, D. It would be much easier if WENT were GOED. Proceedings of the 4th Conference of ECACL, Manchester, 1989
Tufis, D., Dumitrescu, C.; MORPHO - A dictionary management system, Proceedings of the 13th International Seminar on DBMS, Mamaia, 1990