LINGUISTIC PROBLEMS IN MULTILINGUAL MORPHOLOGICAL DECOMPOSITION

G.Thurmair
Siemens AG
ZT ZTI
Otto-Hahn-Ring 6
Munich 83
West-Germany

ABSTRACT

An algorithm for the morphological decomposition of words into morphemes is presented. The application area is information retrieval, and the purpose is to find morphologically related terms to a given search term. First, the parsing framework is presented, then several linguistic decisions are discussed: morpheme selection and segmentation, morpheme classes, morpheme grammar, allomorph handling, etc. Since the system works in several languages, language-specific phenomena are mentioned.

I BACKGROUND

1. Application domain

In Information Retrieval (document retrieval), the usual way of searching documents is to use key words (descriptors). In most of the existent systems, these descriptors are extracted form the texts automatically and by no means standardised; this means that the searcher must know the exact shape of the descriptor (plural form, compound word etc.), which he doesn't; therefore the search results are often poor and meager.

To improve them, we have developed several analysis systems, based on linguistic techniques. One of them is the Morphological Analysis for Retrieval Support (MARS). It expands the terms of search questions morphologically and returns an expanded set of tokens containing the same root as the search term. This is done through analysis of a set of documents: Each word in a document is decomposed into its morphemes, the roots are extracted, allomorphes are brought to the same morpheme representation, and the morphemes are inverted in a way that they are connected to all the words they occur in (see fig.1). Retrieval is done by evaluating these inverted files. As a result, the searcher is independent of the morphological shape of the term he wants to search with. From a pure linguistic point of view, the aim is to find the morphological structure of each word as well as the information about which morpheme occurs in which word.

The system has been developed for several languages: We took 36000 english tokens (from the Food Science Technology Abstracts document files), 53000 German tokens (from the German Patent Office document files) and 35000 Spanish tokens (from several kinds of texts: short stories, telephone maintenance, newspapers etc.). In 95-97% of the tokens, the correct roots were extracted; the retrieval results could be improved by overall 70% (for the English version; the German version is currently being tested).

Since the kernel part of the system consists of a morphological decomposition algorithm, it can also be used for the handling of other phenomena below the lexical level, e.g. handling of lexical gaps.

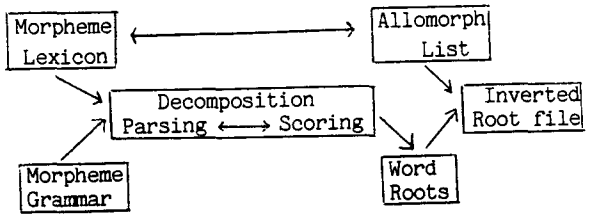
2. The decomposition algorithm

The parser works essentially language independent (see below for language-specific points), using a morpheme list an a morphological grammar of the language in question.

First of all, a preprocessing transforms the input string in order to eliminate some kinds of allomorphes (see below); its operations are just grapheme insertion, deletion and changing; therefore it can be developed language-independent. The transformation conditions and contents, of course, differ for the languages.

Then the transformed string is transferred to the parser. The decomposition works in a left-to-right bredth-first manner. It builds up a network graph consisting of possible morphemes. At a certain state in the graph, the algorithm searches for possible successors: It

Fig.1: System structure



identifies possible morphemes by looking them up in the morpheme list, and checks if they can hold their current position in the word; this is done by means of a morpheme grammar. The morpheme grammar contains the possible sequences of morpheme classes and is represented as a state-transition automaton. If the new morpheme is accepted it is stored in a morpheme chart and connected with the network graph. If the whole input string is processed, the graph is evaluated and put out.

Since the morpheme list and the morpheme grammar are language-specific, they are separated from the parser and stored in files; so the decomposition itself to a large extent is language independent.

In a number of cases (owing both to the morpheme grammar and to true ambiguities in natural language), the parser produces more than one possible result; this means that additional strategies have to be applied to select the most plausible decomposition. The system scores the best result highest and puts it on the top of the decomposition list; nevertheless it keeps the others, because different decompositions may be correct, e.g. for different parts of speech; but this goes beyond pure morphology.

The scored decomposition results are used to extract the root(s) and to disambiguate some morphs. The roots are inverted in such a way that they point to the set of tokens they belong to. Allomorphes of the same morphemes are inverted to the same set of tokens, which means that in searching with any allomorph of a word the system nevertheless will come up with the correct set of tokens.

II LINGUISTIC ISSUES IN DECOMPOSITION

Dealing with large amounts of data, some linguistic problems arise which not only influence the claim that the system should be language independent, but also concern pure morphology. Some of them are presented in the following sections.

1. Morpheme definition

The first problem is to set up rules to define possible morphemes. One crucial point is the morpheme selection: What about proper names (BAGDAD, TOKYO)? What about special terminology (e.g. chemical terms which need special suffixation rules)? What about foreign words and morphemes, which are used quite frequently and have to be considered if the language is seen as a synchronic system? As a result, a pure single-language morphology is highly artificial from the language system point of view, and there is some arbitrariness in selecting morphemes for the morpheme lexicon.

We decided not to deal with proper names and to pick up the morphemes which are quite frequent (with respect to the number of tokens they occur in) and which have many different derivations. So, the morphology of one language (e.g. German) has to be mixed up with the morphology of other languages (Latin, Greek, English) in order to cover a broad variety of the synchronic language system. In addition to this, it has been found that the resulting morpheme lists differ depending on what topic the documents we analyse deal with: Some special vocabulary has to be added to the morpheme list, e.g. with respect to food science. The vocabulary which is considered as basic to all topics consists of approx. 4-5000 morphemes, the rest is special vocabulary. With this morpheme selection, we got error rates (words which could not be decomposed) of 4-8%; most of them were proper names or typing errors.

Another crucial point is morpheme segmentation. As the analysis should be synchronic, no diachronic segmentation is done. Diachronic issues sometimes occur but are not taken into account. But there are two criteria that lead to different segmentations: Purely distributional based segmentation reduces semantically quite different morphemes to the same root (e.g. English ABROAD vs. BROAD, German VERZUG vs. ZUG) and sometimes creates artificial overlappings (e.g. Spanish SOL-O vs. SOL vs. SOL-AR); on the other hand some clear derivations are not recognised because of gaps in the distribution of the lexical material (e.g. -LICH in German OEFFENTLICH). On the other hand, semantically oriented segmentation sometimes leads to a loss of morphological information, e.g. in German prefixation: If the prefixes (VER-LUST, VER-ZUG) are taken as a part of the root, which is correct from a semantic point of view, some information about derivational behaviour gets lost.

We decided to base morpheme segmentation on the semantic criterion to distinguish the meanings of the roots as far as possible, and to segment the morphemes according to their distribution as far as possible: We take the longest possible string which is common to all its derivations even if it contains affixes from a diachronical (and partly derivational) point of view. Since there is some intuition used with respect to which derivations basically carry the same meaning and which affixes should belong to the root (i.e. should be part of a morpheme), the morpheme list is not very elegant from a theoretical point of view; but it must be stated that the language data often don't fit the criteria of morphologists.

These problems are common to the three languages and sometimes lead to irregularities in the morpheme list. The resulting lists consist of 8000 to 10000 morphemes.

2. Morpheme categorisation

Every morpheme has some categorial information attached to it, namely morpheme class, morpheme part of speech, allomorph information, morphological idiosyncrasies and inflectional behaviour.

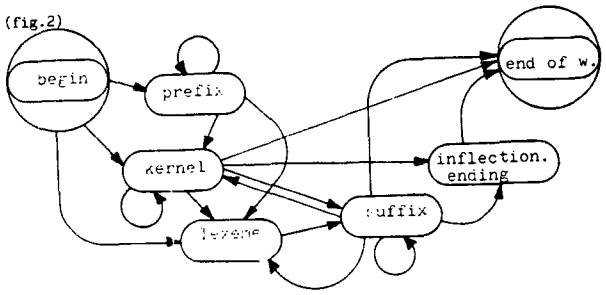
All this information is language dependent: In English, some morphemes duplicate their last consonant in derivation (INFER - INFERRING), and there seems to be phonological but no morphological regularity in that behaviour, so this information has to be explicitly stored. German and Spanish need quite more inflectional information than English does, etc. All this information can be stored within the same data structure for the different languages, but the interpretation of these data has to be programmed separately.

The morpheme classes also depend on the language. Affix classes don't differ very much: Prefixes, suffixes an inflections are common to all the languages considered; but there are fillers in German compound words that don't exist in Spanish, and there are doubleing consonants in English. More differences are found within the lexical morphemes: The three languages have in common the basic distinction between bound and unbound morphemes, but there are special subcategories e.g. within the bound ones: Some need suffixes (THEOR-), some need inflections (the bound version of Spanish SOL-, German SPRACH-), some need doubleing consonants. With respect to this, information about possible succeeding morphemes is stored; but to be able to analyse new or unknown derivations, no additional information should be used: an unbound morpheme can or cannot take a suffix, it can or cannot form a compound word, etc.

3. The morpheme grammar

In this automaton, the sequences of possible morphemes are fixed. For each language it specifies which morpheme class may be followed by which other one: E.g. a prefix may be followed by a bound or unbound lexical morpheme or by another prefix, but not by a suffix or an inflection. The grammar automaton is stored in a file and interpreted by the parser; so the parser can handle different languages. The automaton restricts the number of morphemes that can occur in a given input word (see fig. 2).

(fig.2)



Nevertheless, the most effective constraints work on the subclass level: A prefix can be followed by another one, but not every combination is allowed. An unbound morpheme can be followed by an inflection, but the inflection must fit to the inflectional properties of the morpheme (e.g. verb endings to a noun). All these constraints are written in procedures attached to the transitions between possible morpheme grammar states; these procedures are highly language-specific. In fact, this is the biggest problem when talking about language-independency.

4. Allomorph handling

There are several kinds of allomorphes: Some are quite regular and can be eliminated in a preprocessing step: English RELY vs. RELIES, Spanish CUENTA vs. CONTAR are transformed before the decomposition goes on; this is pure string transformation, which can be performed by a quite similar procedure in each language.

Other allomorphes can not be handled automatically; so we attach the allomorph information to the lexical entry of the morphem in question. This is done with strong verbs, with German derivational and inflectional vowel mutations, with some kinds of Greek and Latin morphemes (eg. ABSORB- ING vs. ABSORP-TION, which in fact is regular but ineffective to deal with automatically), etc. Different spellings of morphemes (CENTRE vs. CENTER) also have to be handled as allomorphes.

In our system, the allomorph stems point to the same set of words they occur in, so that the user searching with FOOD will also find words with FEED or FED.

On the other hand, artificial overlappings (Spanish SOL vs. SOL-O, English PIN vs. PINE) should point to different sets of words in order to disambiguate these morphemes; this can be done by looking at the morphological context of the morph in question; but this is not always sufficient for disambiguation. These kinds of overlappings are very common in Spanish, less frequent in English and rather seldom in German.

5. Selection strategies

In 55% of all cases, the decomposition comes up with only one possible result. This, in over 99% of the cases, is a correct result. In over 40%, however, the result is ambiguous: From a morphological point of view, several decompositions of a word are acceptable. Since the system has no syntactical or semantic knowledge, it cannot find out the correct one (e.g. German DIEN-ST is correct for a verb, DIENST for a noun; similar English BUILD-ING (verb) vs. BUILDING (noun)). We decided not to integrate a scoring algorithm into the decomposition itself but to compare ambiguous results and try to find the most plausible decomposition.

To do this, we apply several strategies: First, we compare the sequences of morpheme classes: Suffixation is more frequent than compounding: The compound LINGUISTIC-ALLY therefore is less plausible than the suffixation LINGUISTIC-AL-LY. The morpheme class sequence information can partly be collected statistically

(by evaluating the decompositions with one correct result); nevertheless it has do be optimalised manually by evaluating several thousands of decomposition results. (The statistics partly depend on the type of the text considered).

This strategy works with different results for the different languages. If the affixes of a language are very ambiguous (as it is in German), this strategy is too poor and has to be supported by several others we are just developing. In English and Spanish, however, the results are quite satisfactory: The first 10 of the morpheme class sequences in English cover 60%, the first 50 over 80% of the tokens.

If the morpheme class sequence strategy falls below a threshold (which mostly happens with long compounds), the strategy is switched to longest matching: The decomposition with the fewest morphemes is scored best.

As a result, the disambiguation returns correct roots in 90-94% of the cases; in German, the ambiguous affixes don't influence the root extraction, although the decompositions as a whole are correct only in 85% of the tokens. Together with the decompositions with only one correct result, the whole system works correctly in about 96% of the input words.

III LIMITATIONS

Although the morphological decomposition works quite well and is useful with respect to information retrieval problems, there are some problems concerning the integration of such an algorithm into a whole natural language system. The reason is, that some information needed therefore is not easily available; this is the information which goes beyond morphology and is based on the interpretation of decomposition results. Two examples should be mentioned here.

1. Parts of speech

It is not easy to derive the part of speech of a word out of its decomposition. In German, the prefix VER- forms verb-derivations, but the derivation VER-TRAUEN from the verb TRAUEN is also a noun, whereas the same derivation VER-TRETEN form the verb TRETEN does not, and the derivation VER-LEGEN (from the verb LEGEN) is also an adjectiv. The past participle GE-FALLEN (from the verb FALLEN) is also a noun, the same derivation from LAUFEN (GE-LAUFEN) is not. This fact is due to the diachronic development of the language which led to a structure of the vocabulary that followed rather the needs of usage than morphological consistency.

2. Semantics

There is some evidence that the meaning of a word can not be predicted out of its morphemes. Words which are morphologically related, like German ZUG vs. BEZUG vs. VERZUG, LUST vs. VERLUST, DAMM vs. VERDAMMEN, are completely different from a semantical point of view. This could mean that the semantic formation rules do not correspond to the morphological ones. But considering large amounts of data, up to now no certain rules can be given how word meaning could be derived from the "basic units of meaning" (what the morphemes claim to be). Semantically and even syntactically regular behaviour can be observed at the level of words rather than morphemes. The result of our research on morphemes tends to support those who stress the status of the word as the basic unit of linguistic theory.

ACKNOWLEDGEMENTS

The work which was described here was done by A.Baumer, M.Streit, G.Thurmair (German), I.Buettel, G.Th.Niedermair, Ph.Hoole (English) and M.Meya (Spanish).

REFERENCES

Meya,M.: Morpheme Grammar. Proc. of 11th Int. Conference of ALLC, April 1984, Louvain-La-Neuve

Niedermair,G.Th., Thurmair,G., Buettel,I.: MARS: A retrieval tool on the basis of Morphological Analysis. Proc. of the ACM Conference on Research and Development in Information Retrieval, July 1984, Cambridge

Hunnicut,S.: A new Morph Lexicon for English. Proc. of the COLING Conference, 1976

Karttunen, L.: KIMMO - a general morphological Processor. Texas Linguistic Forum 22, 1983.

Koskenniemi,K.: Two-level model for morphological analysis. Proc. of the 8th Intern. Joint Conference on Artificial Intelligence, 1983.