WORD FORMATION IN NATURAL LANGUAGE PROCESSING SYSTEMS

Roy J. Byrd

IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598

ABSTRACT

Systems which process natural language require a reliable source of information about words. Not only must their lexical subsystems handle a large number of known words; they must also cope with coinages. The morphological principles underlying the notion "possible word" are under active study by linguists, and are articulated in the theory of word formation. This paper presents a technique for building lexical subsystems which embody these principles by emulating the behavior of word formation rules. These subsystems combine totally idiosyncratic lexical information, stored in a dictionary, with systematic information derived from word structure. Applications for lexical subsystems built along the lines described here will be discussed.

1. Introduction.

In many published descriptions of computer systems that process natural language, little or no information is given about the sources of lexical information. Instead, fully analyzed words, bearing all required categorial, syntactic, and semantic features, appear in the data structures presented as illustrations of the systems' operation. The dictionary structures which store these words, and the processes by which feature information is derived, are not usually revealed.

When system descriptions do shed some light on their lexical processing, two general approaches can be discerned. In the first, typified by Sager(1981), the. words in the input text serve as keys for accessing dictionary entries where information about the words is stored. Different inflected forms of a word serve as independent keys into the dictionaries' data structures, although morphological regularities are sometimes captured by allowing different dictionary entries for inflectionally related words to share sublists of common properties.

The second category of lexical subsystems uses "affix stripping" to economize on word storage, or, equivalently, to increase the apparent size of their word lists. In the simpler versions of the procedure, the systems will merely test if a valid ending is attached to a valid word (as in Peterson(1980)). In some cases adjustments are made to the spelling of the base after stripping the affix and before looking up the base in the dictionary (as in Winograd(1972)). These systems usually handle only words with inflectional suffixes, reducing them to their uninflected forms. Words with derivational affixes again have separate dictionary entries.

A more complex form of the affix stripping procedure is presented in Cercone(1974). In that system, prefixes as well as suffixes are (recursively) removed from an incoming word, and the resulting bases are looked up in the. dictionary, as above. Only if the base is of an appropriate category for the affix, will the input be accepted, however. Furthermore, information that is redundantly associated with the affix--such as the fact that words ending in -ness are nouns--is asserted for the input word. Cercone's system handles derivational as well as inflectional affixes.

The system of morphological rules described in this paper is superior to the affix stripping systems in several respects. First, it allows for the features associated with a word to be a combination of totally idiosyncratic information from a dictionary, together with systematic information derived by the recursive application of morphological rules. Second, it provides for the observation*of a complex set of restrictions on morphological rule application. The nature of these restrictions is known from linguistic research, and their observation allows for much tighter control over the attachment of affixes to bases. This enhanced control provides for the third type of improvement over previous systems: this system can exhibit much more reliable behavior in the presence of coinages--the creation of new words by human users.

Section 2 of this paper describes word formation rules and the restrictions that govern their application. Section 3 presents a system of morphological rules based in the linguistic results of section 2. The final section lists applications of the techniques developed in this paper.

2. Word Formation Rules.

Linguistic studies of morphology have settled on the word formation rule as the best means of accounting for the structure of words. Extensive descriptions of two varying views of this model can be found in Aronoff(1976) and Selkirk(1982). Briefly, word formation rules can be represented as in (1). Rule (la) will apply to transitive verbs that select for animate objects, and will produce animate nouns. Using this rule, we can derive draftee from draft, but we can't get *singee (sing doesn t take animate

objects) or *abdicatee (abdicate is intransitive).

Rule (Ib) derives latinate adjectives from transitive verbs. For example, we can get interruptible and singable, but not *abdicatable. Rule (Ic) says that the inflectional affix -ed can be applied to verbs bearing the morphological feature [-ablaut], yielding a past tense verb. It will apply to walk to produce walked, but not to sing which is [+ablaut].

(1)a. +ee: $[\{[+\text{animate object}]_V, [+\text{transitive}]_V\}\ \_\_]_N\ [+\text{animate}]$

b. #abl: $[[+\text{transitive}]_V\ \_\_]_A\ [+\text{latinate}]$

c. -ed: $[[-\text{ablaut}]_V\ \_\_]_V\ [+\text{past}]$

Three types of restrictions apply to word formation rules: boundary restrictions, subcategorization restrictions, and selectional restrictions. Boundary restrictions refer to the fact that some types of affixes may only occur "outside." of other types. Selkirk(1982) distinguishes three types of boundary.

a) The "innermost" are those associated with non-neutral affixes which can alter the phonology of the base to which they apply. Such an affix is +ee in (Ia) which alters the stress pattern of assign when forming assignee. These boundaries are denoted by "+".
b) Boundaries at the next level, denoted by "#", are associated with neutral affixes, such as #abl, which do not alter the phonology of their base. Note that the stress of the base is not shifted in assignable.
c) The 'outermost" boundaries are associated with inflectional affixes, ana are denoted by "-". An example is the past tense suffix -ed. An important fact about inflectional affixes is that they do not pile up"; at most one is ever found on a single word.

Subcategorization restrictions limit the occurrence of affixes to environments containing only certain word categories. Thus, in example (Ib), #abl may only cooccur with verbs. It cannot be affixed to nouns (*elementable) or adjectives (*temporaryable). Cercone's morphological analysis system captures subcategorization restrictions when it tests that the string remaining after affix stripping belongs to a certain category.

Selectional restrictions specify constraints on non-categorial features of an affix's base. Thus, the base to which the affix +ee can apply must not simply be a verb, it must also bear the features [-••transitive] and [+animate object].

Since the word formation process is recursive, and since word formation rules themselves can assert new features for the words that they form, the selectional restriction mechanism can be a powerful tool. Consider the rule in (2).

1. [+latinate] is an abstract morphological feature which roughly indicates that words so marked are of Greek or Latin origin. A particular use of this abstract feature will be given below.

(2) +ity: $[[+\text{latinate})_A\ \_]_N$

Because +ity only attaches to latinate adjectives (we get falsity but not *wrongity), its ability to cooccur with singable in singability must mean that singable is [+latinate]. But since sing is not a latinate word, this must be so because of the suffix +abl in singable. Thus, the derivational history of a word, in addition to its ultimate base, determines what features will be associated with it.[2]

3. The Morphological Rule System.

The linguistic theory of word formation rules, outlined above, can be advantageously exploited for natural language processing systems. A useful way of organizing the lexical subcomponent of such systems is to incorporate a dictionary, where truly idiosyncratic information about words is stored, and an interpretive mechanism for applying morphological rules derived from word formation rules.

The information contained in the dictionary should include not only the usual categorial, syntactic, and semantic features. It should also include morphological, etymological, and phonological information relevant to word formation processes. This information can be efficiently encoded, and the payoff--as this paper attempts to show--is well worth the additional cost.

The morphological rules themselves consist of five parts: a) an affix name, b) a boundary specification, c) a pattern, d) a condition, and e) an assertion. The parts of a rule combine to emulate the behavior of a word formation rule in linguistic theory. The pattern specifies the base for the rule by describing the affix to be removed and further adjustments to be made. The condition embodies the subcategorization and selectional restrictions on the base. The assertion allows statement of the categorial and diacritic features for the output of a rule application. The boundary specification insures the observation of boundary restrictions.

An example of a morphological rule is given in (3). The five parts of the rule are labeled a) through e).

(3) +ion: ation5*e* (verb) (noun +sg +abstr)
b)a) c) d) e)

This rule would operate as follows while analyzing the word realization. The position of the boundary marker indicates that this is a suffix rule. Hence, the right end of the word is checked for the pattern characters ''a-t-i-o-n''. This check succeeds, so 5 characters are removed and the "*" causes the result ("realiz") to be looked up either in the dictionary or via a recursive invocation of the morphological rule processor. This look up fails--.it would have succeeded had the original word been, say, relaxation—so "e" is added, yielding "realize" which is

2. In spite of appearances, the word singability is not a violation of the boundary restrictions mentioned earlier. See Aronoff(1976) for a justification of the existence of two suffixes, #abl and +abl. +abl is the one in singability.

successfully looked up. The categorial and diacritic features of the base are checked against the condition. In this case, the fact that "realize" is a verb suffices. Furthermore, the boundary specification is not violated, since "realize" contains no neutral or inflectional affixes. So the assertion is applied, associating the categorial feature "noun" and the diacritic features "+singular -♦-abstract" with the word "realization".

These rules actually manipulate morphographemes (i.e., the written form of words) rather than morphemes. While this is perfectly acceptable for computer applications, we must be clear about the correspondence between the abstract word formation rules, and their realization as morphological rules. Thus, the three morphological rules in (4) represent different orthographical facets of a single word formation rule which handles the suffix #abl.

(4)a. #abl: able4*e* (v +trans) (a +lat -sg -pi)

b. #abl: sible5d* (v +trans) (a +lat -sg -pi)

c. #abl: able3te* (v +trans) (a +lat -sg -pl)

Rule (4a) will derive reachable from reach and likable from like. (4b) derives defensible from defend. (4c) derives delegable from delegate. All will check that the base is a transitive verb, and assert that the result is a latinate adjective. Similar clusters of rules will exist for cases of allomorphy, as in in+continent, im+practical, irreversible, and il+//ogical. In fact, the patterns of morphological rules can be used to capture the allomorphy and truncation phenomena discussed by Aronoff(1976). In addition, they can encode purely orthographical phenomena, such as the spelling rules for silent e suppression, consonant doubling, c-k alternations, etc.

In a lexical subcomponent constituted as suggested here, the information known about a given word is a combination of the inherent features found in the lexical entry for the ultimate base and the systematic features associated with the word's structure. Thus, if one form of the verb realize is known to be transitive and to require an animate subject (as in "the sculptor realized his masterpiece in bronze"), then we also know that realization is likewise transitive and requires an animate subject (as in "the sculptor's realization of his masterpiece"). Furthermore, we know that the semantic properties of realize and realization are closely related. This means of attaching inherent information to many related words captures the lexical redundancy relations introduced by Chomsky(1970) and elaborated by Jackendoff(1975). It goes far beyond Sager's(1981) scheme of having multiple dictionary entries for inflectionally related words point to shared sublists of common properties.

4. Applications.

A lexical subcomponent based on the ideas presented here has been implemented and is being used in a system that produces syntactic and stylistic critiques of English language texts in a word-processing environment. The system is described in Heidorn, et al.(1982). As part of the development of that system's dictionary, a set of morphological rules for analysing inflectional affixes was used to automatically identify redundant words in a source dictionary. From among more than 80,000 original entries, it was possible to omit approximately 10,000 which were found to have completely predictable analyses.

The morphological rule interpreter itself is being used as part of a linguistic study of restrictions on word formation rules, described in Byrd(1983). This study should yield a highly detailed set of word formation rules which, in turn, will provide the basis for morphological rules to be used in various computer applications.

In automatic text-to-speech synthesis systems, accurate morphological decomposition as well as reliable phonological information, such as can be stored in our dictionary, is essential. A version of the morphological rule interpreter has been implemented which can decompose an input word into a base, for which a pronunciation is known, plus an ordered list of affixes. See Allen(1976) for another approach to this problem. Finally, with a highly articulated set of morphological rules, it seems likely that this technology could be used to generate words as reliably as it analyses them. Such a capability would be invaluable in text generation applications.

References.

Allen, J. (1976) "Synthesis of Speech from Unrestricted Text," Proceedings of the IEEE 64, 433-442.

Aronoff, M. (1976) Word Formation in Generative Grammar, Linguistic Inquiry Monograph 1, MIT Press, Cambridge, Massachusetts.

Byrd, R. J. (1983) "On Restricting Word Formation Rules," unpublished paper, New York University.

Cercone, N. (1974) "Computer Analysis of English Word Formation," Technical Report TR74-6, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada.

Chomsky, N. (1970) Remarks on Nominalizations," in R. Jacobs and P. S. Rosenbaum, eds. Readings in English Transformational Grammar, Ginn, Waltham, Massachusetts.

Heidorn, G. E., K. Jensen, L. A. Miller, R. J. Byrd, and M. S. Chodorow (1982) "The EPISTLE Text-Critiquing System," IBM Systems Journal 21, 305-326.

Jackendoff, R. S. (1975) "Morphological and Semantic Regularities in the Lexicon," Language 51, 639-671.

Peterson, J. L. (1980) "Computer Programs for Detecting and Correcting Spelling Errors," Communications of the ACM, 23, 676-687.

Sager, N. (1981) Natural Language Information Processing: A grammar of English and Its Applications, Addison-Wesley, Reading, Massachusetts .

Selkirk, E. 0. (1982) The Syntax of Words, Linguistic Inquiry Monograph 7, MIT Press, Cambridge, Massachusetts.

Winograd, T. (1972) Understanding Natural Language, Academic Press, New York.