NAME

morphy − discussion of WordNet's morphological processing

DESCRIPTION

Although only base forms of words are usually stored in WordNet, searches may be done on inflected forms. A set of morphology functions, Morphy, is applied to the search string to generate a form that is present in WordNet.

Morphology in WordNet uses two types of processes to try to convert the string passed into one that can be found in the WordNet database. There are lists of inflectional endings, based on syntactic category, that can be detached from individual words in an attempt to find a form of the word that is in WordNet. There are also exception list files, one for each syntactic category, in which a search for an inflected form is done. Morphy tries to use these two processes in an intelligent manner to translate the string passed to the base form found in WordNet. Morphy first checks for exceptions, then uses the rules of detachment. The Morphy functions are not independent from WordNet. After each transformation, WordNet is searched for the resulting string in the syntactic category specified.

The Morphy functions are passed a string and a syntactic category. A string is either a single word or a collocation. Since some words, such as **axes** can have more than one base form (**axe** and **axis**), Morphy works in the following manner. The first time that Morphy is called with a specific string, it returns a base form. For each subsequent call to Morphy made with a **NULL** string argument, Morphy returns another base form. Whenever Morphy cannot perform a transformation, whether on the first call for a word or subsequent calls, **NULL** is returned. A transformation to a valid English string will return **NULL** if the base form of the string is not in WordNet.

The morphological functions are found in the WordNet library. See **morph**(3WN) for information on using these functions.

Rules of Detachment

The following table shows the rules of detachment used by Morphy. If a word ends with one of the suffixes, it is stripped from the word and the corresponding ending is added. Then WordNet is searched for the resulting string. No rules are applicable to adverbs.

POS	Suffix	Ending
NOUN	"s"	""
NOUN	"ses"	"s"
NOUN	"xes"	"x"
NOUN	"zes"	"z"
NOUN	"ches"	"ch"
NOUN	"shes"	"sh"
NOUN	"men"	"man"
NOUN	"ies"	"y"
VERB	"s"	""
VERB	"ies"	"y"
VERB	"es"	"e"
VERB	"es"	""
VERB	"ed"	"e"
VERB	"ed"	""
VERB	"ing"	"e"
VERB	"ing"	""
ADJ	"er"	""
ADJ	"est"	""
ADJ	"er"	"e"
ADJ	"est"	"e"

Exception Lists

There is one exception list file for each syntactic category. The exception lists contain the morphological transformations for strings that are not regular and therefore cannot be processed in an algorithmic manner. Each line of an exception list contains an inflected form of a word or collocation, followed by one or more base forms. The list is kept in alphabetical order and a binary search is used to find words in these lists. See **wndb**(5WN) for information on the format of the exception list files.

Single Words

In general, single words are relatively easy to process. Morphy first looks for the word in the exception list. If it is found the first base form is returned. Subsequent calls with a **NULL** argument return additional base forms, if present. A **NULL** is returned when there are no more base forms of the word.

If the word is not found in the exception list corresponding to the syntactic category, an algorithmic process using the rules of detachment looks for a matching suffix. If a matching suffix is found, a corresponding ending is applied (sometimes this ending is a **NULL** string, so in effect the suffix is removed from the word), and WordNet is consulted to see if the resulting word is found in the desired part of speech.

Collocations

As opposed to single words, collocations can be quite difficult to transform into a base form that is present in WordNet. In general, only base forms of words, even those comprising collocations, are stored in WordNet, such as **attorney general**. Transforming the collocation **attorneys general** is then simply a matter of finding the base forms of the individual words comprising the collocation. This usually works for nouns, therefore non-conforming nouns, such as **customs duty** are presently entered in the noun exception list.

Verb collocations that contain prepositions, such as **ask for it**, are more difficult. As with single words, the exception list is searched first. If the collocation is not found, special code in Morphy determines whether a verb collocation includes a preposition. If it does, a function is called to try to find the base form in the following manner. It is assumed that the first word in the collocation is a verb and that the last word is a noun. The algorithm then builds a search string with the base forms of the verb and noun, leaving the remainder of the collocation (usually just the preposition, but more words may be involved) in the middle. For example, passed **asking for it**, the database search would be performed with **ask for it**, which is found in WordNet, and therefore returned from Morphy. If a verb collocation does not contain a preposition, then the base form of each word in the collocation is found and WordNet is searched for the resulting string.

Hyphenation

Hyphenation also presents special difficulties when searching WordNet. It is often a subjective decision as to whether a word is hyphenated, joined as one word, or is a collocation of several words, and which of the various forms are entered into WordNet. When Morphy breaks a string into "words", it looks for both spaces and hyphens as delimiters. It also looks for periods in strings and removes them if an exact match is not found. A search for an abbreviation like **oct.** return the synset for { **October, Oct** }. Not every pattern of hyphenated and collocated string is searched for properly, so it may be advantageous to specify several search strings if the results of a search attempt seem incomplete.

BUGS

Since many noun collocations contains prepositions, such as **line of products**, an algorithm similar to that used for verbs should be written for nouns. In the present scheme, if Morphy is passed **lines of products**, the search string becomes **line of product**, which is not in WordNet

ENVIRONMENT VARIABLES

WNHOME Base directory for WordNet. Unix default is **/usr/local/WordNet-2.0**, Windows default is **C:\Program Files\WordNet\2.0**.

WNSEARCHDIR Directory in which the WordNet database has been installed. Unix default is

WNHOME/dict, Windows default is **WNHOME\dict**.

FILES

In directory **WNSEARCHDIR**:

*pos***.exc** morphology exception lists

SEE ALSO

wn(1WN), **wnb**(1WN), **binsrch**(3WN), **morph**(3WN), **wndb**(5WN), **wninput**(7WN).