NAME

index.sense, sense.idx − WordNet's sense index

DESCRIPTION

The WordNet sense index provides an alternate method for accessing synsets and word senses in the WordNet database. It is useful to applications that retrieve synsets or other information related to a specific sense in WordNet, rather than all the senses of a word or collocation. It can also be used with tools like **grep** and Perl to find all senses of a word in one or more parts of speech. A specific WordNet sense, encoded as a *sense_key*, can be used as an index into this file to obtain its WordNet sense number, the database byte offset of the synset containing the sense, and the number of times it has been tagged in the semantic concordance texts.

Concatenating the *lemma* and *lex_sense* fields of a semantically tagged word (represented in a <**wf** ... > attribute/value pair) in a semantic concordance file, using **%** as the concatenation character, creates the *sense_key* for that sense, which can in turn be used to search the sense index file.

A *sense_key* is the best way to represent a sense in semantic tagging or other systems that refer to WordNet senses. *sense_key*s are independent of WordNet sense numbers and *synset_offset*s, which vary between versions of the database. Using the sense index and a *sense_key*, the corresponding synset (via the *synset_offset*) and WordNet sense number can easily be obtained. A mapping from noun *sense_key*s in WordNet 1.6 to corresponding 2.0 *sense_key*s is provided with version 2.0, and is described in **sensemap**(5WN).

See **wndb**(5WN) for a thorough discussion of the WordNet database files.

File Format

The sense index file lists all of the senses in the WordNet database with each line representing one sense. The file is in alphabetical order, fields are separated by one space, and each line is terminated with a newline character.

Each line is of the form:

sense_key synset_offset sense_number tag_cnt

sense_key is an encoding of the word sense. Programs can construct a sense key in this format and use it as a binary search key into the sense index file. The format of a *sense_key* is described below.

synset_offset is the byte offset that the synset containing the sense is found at in the database "data" file corresponding to the part of speech encoded in the *sense_key*. *synset_offset* is an 8 digit, zero-filled decimal integer, and can be used with **fseek**(3) to read a synset from the data file. When passed to the WordNet library function **read_synset( )** along with the syntactic category, a data structure containing the parsed synset is returned.

sense_number is a decimal integer indicating the sense number of the word, within the part of speech encoded in *sense_key*, in the WordNet database. See **wndb**(5WN) for information about how sense numbers are assigned.

tag_cnt represents the decimal number of times the sense is tagged in various semantic concordance texts. A *tag_cnt* of **0** indicates that the sense has not been semantically tagged.

Sense Key Encoding

A *sense_key* is represented as:

*lemma***%***lex_sense*

where *lex_sense* is encoded as:

*ss_type***:***lex_filenum***:***lex_id***:***head_word***:***head_id*

lemma is the ASCII text of the word or collocation as found in the WordNet database index file corresponding to *pos*. *lemma* is in lower case, and collocations are formed by joining individual words with an underscore (_) character.

ss_type is a one digit decimal integer representing the synset type for the sense. See **Synset Type** below for a listing of the numbers corresponding to each synset type.

lex_filenum is a two digit decimal integer representing the name of the lexicographer file containing the synset for the sense. See **lexnames**(5WN) for the list of lexicographer file names and their corresponding numbers.

lex_id is a two digit decimal integer that, when appended onto *lemma*, uniquely identifies a sense within a lexicographer file. *lex_id* numbers usually start with **00**, and are incremented as additional senses of the word are added to the same file, although there is no requirement that the numbers be consecutive or begin with **00**. Note that a value of **00** is the default, and therefore is not present in lexicographer files. Only non-default *lex_id* values must be explicitly assigned in lexicographer files. See **wninput**(5WN) for information on the format of lexicographer files.

head_word is only present if the sense is in an adjective satellite synset. It is the lemma of the first word of the satellite's head synset.

head_id is a two digit decimal integer that, when appended onto *head_word*, uniquely identifies the sense of *head_word* within a lexicographer file, as described for *lex_id*. There is a value in this field only if *head_word* is present.

Synset Type

The synset type is encoded as follows:

1	NOUN
2	VERB
3	ADJECTIVE
4	ADVERB
5	ADJECTIVE SATELLITE

NOTES

For non-satellite senses the *head_word* and *head_id* fields have no values, however the field separator character (**:**) is present.

ENVIRONMENT VARIABLES (UNIX)

WNHOME	Base directory for WordNet. Default is **/usr/local/WordNet-2.1**.
WNSEARCHDIR	Directory in which the WordNet database has been installed. Default is **WNHOME/dict**.

REGISTRY (WINDOWS)

HKEY_LOCAL_MACHINE\SOFTWARE\WordNet\2.1\WNHome

Base directory for WordNet. Default is **C:\Program Files\WordNet\2.1**.

FILES

index.sense	sense index

SEE ALSO

binsrch(3WN), **wnsearch**(3WN), **lexnames**(5WN), **wnintro**(5WN), **sensemap**(5WN), **wndb**(5WN), **wninput**(5WN).