NAME

semcor – discussion of semantically tagged texts

DESCRIPTION

A semantic concordance is a textual corpus and a lexicon so combined that every substantive word in the text is linked to its appropriate sense in the lexicon. Thus it can be viewed either as a corpus in which words have been tagged syntactically and semantically, or as a lexicon in which example sentences can be found for many definitions. Texts that were used to create the semantic concordances were extracted from the Brown Corpus and then linked to senses in the WordNet lexicon.

The semantic tagging was done by hand, using various tools to annotate the text with WordNet senses. The "raw" data were reformatted and syntactically tagged (using Eric Brill's stochastic part-of-speech tagger) before semantic tags were assigned. After semantic tagging, the files conform to the SGML-like file format described in **cxtfile**(5WN). The tools and programs used to create the semantic concordances are not distributed.

escort(1WN) is a window-based browser used to search the semantic concordances for instances of semantically tagged word forms. It can be used to find semantic tags to one or more senses of a word and optional co-occurring senses.

Semantic Concordance Organization

The semantically tagged Brown Corpus files are divided into three semantic concordances based on what was tagged and when. Each is stored in a separate directory by the concordance's name (*conc*). The concordances are:

conc	**Contents**	**What's Tagged**
brown1	103 Brown Corpus files	All open class words
brown2	83 Brown Corpus files	All open class words
brownv	166 Brown Corpus files	Verbs

Each directory contains the files **cntlist**, **taglist** and **statistics**, and a **tagfiles** directory. See **cntlist**(5WN) and **taglist**(5WN) for information about these files. See **STATISTICS** for information about the contents of the **statistics** file.

The **tagfiles** directory contains the semantically tagged files. Each file is named using the following convention:

br-*article_code*

where *article_code* is a letter followed by a two digit number that denotes the section and article number that the text was derived from. No file is in more than one semantic concordance.

Semantically Tagged Files

An SGML-like markup language, described in **cxtfile**(5WN), was developed for codifying semantically tagged data. The original content of a Brown Corpus file, with the exception of headlines when they were present, is contained in its corresponding semantically tagged file. The "raw" data are segmented into paragraphs and sentences, then sequentially numbered within the file. Each sentence is separated into word forms and punctuation.

A *semantic tag* associated with a word form indicates one or more senses in the WordNet database that are appropriate for that word form in the textual context. Semantic tags are represented as SGML attribute/value pairs, and are described in detail in **cxtfile**(5WN).

Only nouns, verbs, adjectives, and adverbs (open class words) can be semantically tagged, as these are the

only classes of words represented in WordNet. Proper nouns are generally not in WordNet, but are labeled in the semantically tagged files with one of four categories and assigned semantic tags to predetermined WordNet senses for these categories.

Attribute/Value Pair	WordNet Sense	Sense Key
pn=person	1	person%1:03:00::
pn=location	1	location%1:03:00::
pn=group	1	group%1:03:00::
pn=other	*n/a*	*n/a*

Strings of several words that form a collocation or phrase found in WordNet are joined into one word form in a semantically tagged file and tagged as a single unit. In the case of discontinuous constituents (a collocation in which the words are not adjacent, such as *look up* in the phrase *look the number up*) the first word of the collocation is "redefined" as the entire collocation and is tagged to the appropriate WordNet sense. The remaining words are marked with a special attribute/value pair that indicates which word form contains the semantic tag.

STATISTICS

Category	Semantic Concordance			Total
	brown1	**brown2**	**brownv**	
Miscellaneous				
total word forms (**<wf>**)	198796	160936	316814	676546
word forms with **cmd=done** including **ot=**	122724	98235	53421	274380
word forms with **cmd=done** excluding **ot=notag**	107118	86255	41607	234980
word forms with semantic pointers (**wnsn=**)	106639	86000	41497	234136
word forms tagged to multiple senses	115	551	37	703
total semantic pointers (including multiple senses)	106725	86414	41525	234664
untagged word forms (**cmd=ignore** + **ot=**)	92154	74936	135684	302774
Number of Semantic Pointers				
semantic pointers to nouns	48835	39477	0	88312
semantic pointers to verbs	26686	21804	41525	90015
semantic pointers to adjectives	9886	7539	0	17425
semantic pointers to adverbs	11347	9245	0	20592
semantic pointers to adjective satellites	9970	8347	0	18317
Total semantic pointers	106724	86412	41525	234661
Pointers to Proper Nouns				
pointers to **pn=person**	3815	2783	0	6598
pointers to **pn=location**	600	363	0	963
pointers to **pn=group**	740	440	0	1180
pointers to **pn=other**	447	489	7	943
Total pointers to proper nouns	5602	4075	7	9684
Unique WordNet Senses				
senses pointed to by nouns	11399	9546	0	20945
senses pointed to by verbs	5334	4790	6520	16644
senses pointed to by adjectives	1754	1463	0	3217
senses pointed to by adverbs	1455	1377	0	2832
senses pointed to by adjective satellites	3451	3051	0	6502
Total senses	23393	20227	6520	50140

The previous table was compiled from the data in the **statistics** file in each concordance directory.

Note that there are 7 attribute/value pairs that assign proper nouns to the category "other" in the concordance **brownv**. These proper nouns were incorrectly identified as verbs by the syntactic tagger. See **cxtfile**(5WN) for a detailed discussion of the attribute/value pairs.

ENVIRONMENT VARIABLES

WNHOME Base directory for WordNet. Unix default is **/usr/local/wordnet1.6**, PC default is **C:\wn16**, Macintosh default is **:**.

WNSEARCHDIR Directory in which the WordNet database has been installed. Unix default is **WNHOME/dict**, PC default is **WNHOME\dict**, Macintosh default is **:Database**.

SEMCORDIR Directory in which the semantic concordance package has been installed. Unix default is **WNHOME/semcor**, PC default is **WNHOME\semcor** and Macintosh default is **:Semcor**.

FILES

All files are in **SEMCORDIR/***conc* on Unix platforms, **SEMCORDIR***conc* on PC platforms, and **SEMCORDIR:***conc* on Macintosh platforms.

cntlist file listing number of times each tagged sense occurs in semantic concordance *conc*

taglist file listing location of all tagged senses in semantic concordance *conc*

statistics statistics for tagged files in semantic concordance *conc*

tagfiles/br-* semantically tagged Brown Corpus files in semantic concordance *conc* (Unix)

tagfiles\br-* semantically tagged Brown Corpus files in semantic concordance *conc* (PC)

tagfiles:br-* semantically tagged Brown Corpus files in semantic concordance *conc* (Macintosh)

SEE ALSO

escort(1WN), **cntlist**(5WN), **cxtfile**(5WN), **senseidx**(5WN), **sensemap**(5WN), **taglist**(5WN), **wnpkgs**(7WN).

For a description of the Brown Corpus:

Francis, W. N., and Kucera, H. (1982). *"Frequency Analysis of English Usage: Lexicon and Grammar"*, Houghton Mifflin Company, Boston.

For more information on semantic concordances:

Miller, G.A., Leacock, C., Tengi, R., and Bunker R. T. (1993). A Semantic Concordance, *"Proceedings of the ARPA WorkShop on Human Language Technology"*. San Francisco, Morgan Kaufman.

Landes, S., Leacock, C., Tengi, R. (1998). Building Semantic Concordances. *"WordNet: An Electronic Lexical Database"*, MIT Press, Cambridge, MA.