NAME

Format of semantic concordance file

DESCRIPTION

A semantic concordance consists of texts that have been syntactically and semantically tagged. The semantic tagging was done by hand, using various tools to annotate the English text with WordNet senses. The "raw" data were reformatted and syntactically tagged before semantic tags were assigned. See **semcor**(7WN) for more information about semantic concordances.

For historical reasons, semantic concordance files are referred to as *context files*. Some of the field element within context files reflect that history, as well.

File Format

Regardless of which files comprise a semantic concordance, and which words are tagged, the format of each concordance file is the same. The context file format follows SGML guidelines, using elements and attribute/value pairs to record information about the file, paragraph and sentence boundaries, and syntactic and semantic information. SGML elements give a file its structure; attribute/value pairs provide additional information. All SGML elements require both start and end tags.

SGML attribute/value pairs are of the form:

attribute=*value*

Our SGML format deviates from the standard in that *value*s are only enclosed in quotation marks when there may be embedded spaces. Due to the large number of attribute/value pairs, the presence of quotation marks around each *value* would have substantially increased the size of the concordances.

Nomenclature

The structure of a context file is specified below in pseudo-BNF notation. Each SGML element is on a separate line. Terminals are printed in **boldface** and are represented in the file as written. Items in *italics* are variables. Upper-case strings are non-terminals.

File Structure

CONTEXTFILE ::= **<contextfile concordance=***conc***>**
CONTEXT+
</contextfile>

CONTEXT ::= **<context filename=***filename* **paras=yes>**
PARA+ | SENT+
</context>

PARA ::= **<p pnum=***paragraph_number***>**
SENT+
</p>

SENT ::= **<s snum=***sentence_number***>**
SENT_TOK+
</s>

SENT_TOK ::= (WORD_FORM | PUNC)+

WORD_FORM ::= **<wf cmd=tag** RDF SEP POS **>***word***</wf>**
| **<wf cmd=ignore** DC SEP POS **>***word***</wf>**
| **<wf cmd=done** RDF SEP POS SEM_TAG OT**>***word***</wf>**
| **<wf cmd=(update | retag)** RDF SEP POS TAGNOTE NOTE**>***word***</wf>**

POS ::= **pos=**POS_TAG

POS_TAG ::= **CC** | **CD** | **DT** | **EX** | **FW** | **IN** | **JJ** | **JJR** | **JJS** | **LS** | **MD** | **MD|VB**
| **NN** | **NNP** | **NNPS**| **NNP|NP** | **NNP|VBN** | **NNS** | **NN|SYM** | **NP** | **NPS**
| **PDT** | **POS**| **PP** | **PR** | **PRP** | **PRP$** | **RB** | **RBR** | **RBS** | **RP**
| **TO** | **UH** | **VB** | **VBD** | **VBG**| **VBN** | **VBP** | **VBZ** | **WDT** | **WP**| **WP$** | **WRB**

SEM_TAG ::= LEMMA WNSN LEXSN PN | NULL

LEMMA ::= **lemma=***lemma*

WNSN ::= **wnsn=***sense_number*

LEXSN ::= **lexsn=***lex_sense*

PN ::= **pn=**CATEGORY | NULL

CATEGORY ::= **person** | **location** | **group** | **other**

RDF ::= **rdf=***redefinition* | NULL

DC ::= **dc=***distance* | NULL

SEP ::= **sep="***separator_string***"** | NULL

TAGNOTE ::= **tagnote=**TAGNOTE_TYPE

TAGNOTE_TYPE ::= **sns_miss** | **indist_sns** | **wd_miss** | **insuffctxt** | **sense_lost** | **misc**

NOTE ::= **note="***note***"**

OT ::= **ot=**OTHER_TAG | NULL

OTHER_TAG ::= **notag** | **metaphor** | **idiom** | **complexprep** | **foreignword** | **nonceword**

PUNC ::= **<punc>**PUNC_CHARACTER**</punc>**

PUNC_CHARACTER ::= [**, . ? ! , ; ([ )] ‘ ’ $ " :**]

Interpretation of SGML Elements

<contextfile concordance=*conc***>**

This element indicates the beginning of a context file. *conc* specifies the name of the semantic concordance that the file is in. A semantic concordance file contains one or more context elements from the same semantic concordance. See **semcor**(7WN) for a list of semantic concordance names.

<context filename=*filename* **paras=yes>**

This element indicates the beginning of a context. *filename* is the name of the original corpora file from which this context was extracted. **paras** indicates that this document contains paragraph delimiters.

<p pnum=*paragraph_number***>**

Start of a new paragraph. *paragraph_number* is an integer. The first paragraph in a **context** is numbered **1**, and paragraph numbers are incremented sequentially.

<s snum=*sentence_number***>**

Start of a new sentence. *sentence_number* is an integer. The first sentence in each **context** is numbered **1**, and sentence numbers are incremented sequentially throughout the **context**.

Sentence numbers *do not* restart at **1** in each paragraph.

<wf *attribute/value_pairs***>***word***</wf>**

This element represents a word form. *word* is the orthographic form as it appears in the original document. All of the syntactic and semantic information is stored as attribute/value pairs described below.

cmd=*cmd* Indicates the status of the **wf** element.

cmd	**Meaning**
tag	word is to be tagged
done	word is semantically tagged
ignore	word should not be tagged
update	used during semantic concordance development only
retag	used during semantic concordance development only

pos=*pos* *pos* is the syntactic tag assigned by Eric Brill's stochastic part-of-speech tagger. See Syntactic Tags below for a list of possible values.

lemma=*lemma* The base form of the word or collocation that the other attribute/value pairs in this **wf** pertain to. This is the form of the string used to search the WordNet database. If **rdf** is present, *lemma* is the base form of the redefinition. When **pn** is present, *redefinition*, *lemma* and *category* all have the same value.

wnsn=*sense_number*
sense_number is the integer sense number corresponding to the WordNet output display.

lexsn=*lex_sense* *lex_sense*, when concatenated onto *lemma* using **%** as the concatenation character, creates a *sense_key* that indicates the WordNet sense to which *word* should be linked. This is the semantic tag for *word*. The format of a *sense_key* is described in **senseidx**(5WN).

pn=*category* Indicates that *word* is a proper noun categorized as one of the values of CATEGORY. When **pn** is present, *redefinition*, *lemma* and *category* all have the same value.

rdf=*redefinition* If present, *word* has been "redefined" to something else. This is mainly used to define discontinuous collocations, correct typographical errors in the text, or to enter a string that should be used to search WordNet instead of *word* in order to find an appropriate sense for the semantic tag. When **pn** is present, *redefinition*, *lemma* and *category* all have the same value.

dc=*distance* Indicates that *word* is part of a discontinuous collocation in which the words comprising the collocation are not adjacent. *distance* is an integer that specifies how many **wf** elements away the semantic tag for the collocation is. It may be either negative, indicating **wf** elements prior to this one, or positive, indicating **wf** elements following in the file.

sep="*separator_string***"**
Indicates that the space between this **wf** element and the next should be displayed as *separator_string*. The string may be one or more character. The default word separator is one space.

tagnote=*tagnote_type*
A **tagnote** attribute/value pair is always present if *cmd* is **update** or **retag**. This is used only during semantic concordance development, and indicates the type of problem encountered during semantic tagging. See TAGNOTE_TYPE for a list of possible values.

note="*note***"** A **note** attribute/value pair is always present with **tagnote**. *note* may contain a string that provides additional information about the **tagnote**, or may be empty.

ot=*other_tag* If present, a semantic tag cannot be assigned to *word* for one of the reasons listed in OTHER_TAG.

Syntactic Tags

The following tags are assigned by Brill's stochatstic part-of-speech tagger.

Syntactic Tag	Interpretation
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential "there"
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNP	Proper noun, singular
NNPS	Proper noun, plural
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PP	Personal pronoun
PR	Pronoun
PRP	Pronoun
PRP$	Pronoun, plural
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	"to"
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

SEE ALSO

escort(1WN), **senseidx**(5WN), **wndb**(5WN), **morphy**(7WN), **semcor**(7WN).