Abstract

There are several electronic dictionaries, thesauri, lexical databases, and so forth today. WordNet is one of the largest and most widely used of these. It has been used for many natural language processing tasks, including word sense disambiguation and question answering. This is an attempt to explore and understand the structure of WordNet, and how it is used for what applications it is used, and also to see where its strength and weakness lies.

1. WordNet as a lexical database

1.1 Background

Before the 1990s, most of the dictionaries for English existed only in paper form. The dictionaries that were available in electronic form were limited to a few groups of researchers. This was something that hindered much work to be done in certain areas of computational linguistics, for example word sense disambiguation (WSD). In 1993, WordNet was introduced. It is a lexical database, organized as a semantic network. The development began in 1985 at Princeton University by a group of psychologists and linguists, and the university still is the maintainer of this lexical database. Even though it was not created with the intention to serve as knowledge source for tasks in computational linguistics, it has been used as such. It has been widely used as a lexical resource for different tasks, and has spawned many different subsets. One task that it has been widely used for is the previous mentioned WSD.

1.2 Structure

WordNet consists of three separate databases, one for nouns, one for verbs and one for adjectives and adverbs. It does not include closed class words. The current version available for download is WordNet 3.0, which was released in December 2006. It contains 117,097 nouns, 22,141 adjectives, 11,488 verbs and 4,601 adverbs. There is a later release, 3.1, which is available for online usage.

The basic structure is *synsets*. These are sets of synonyms, or more correct, near-synonyms, since there exists none to few true synonyms. Synsets contain a set of lemmas, and these sets are tagged with the sense they represent. These senses can be said to be concepts, all of the lemmas (or words), can be said to express the same concept. Word forms which have different meanings appear in different synsets. For example the noun *bank*, has 10 different senses in WordNet, and thus it also connected in some way to other synsets, expressing some kind of relation. Which these relations are depend on the part of speech of the word itself, although the hypernym/hyponym (the what do you mean with this sentence? (some of these relations?)
Semantic analysis in language technology

1.2.1 Nouns

Nouns have the richest set of relations of all parts of speech represented in WordNet, with 12 different relations. As previously stated, the hyponym/hypernym relation is the most frequent used one. For example, if we look at the noun **bass** again (which have 8 different senses), now in the context of sea bass, it is a saltwater fish, which is a kind of seafood, which is a kind of solid food, and so on. These relations are also transitive, which means that sea bass is a type of food, as much as it is a type of saltwater fish.

<table>
<thead>
<tr>
<th>Sense 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>sea bass, bass</td>
</tr>
<tr>
<td>➞ saltwater fish</td>
</tr>
<tr>
<td>➞ seafood</td>
</tr>
<tr>
<td>➞ food, solid food</td>
</tr>
<tr>
<td>➞ solid</td>
</tr>
<tr>
<td>➞ matter</td>
</tr>
<tr>
<td>➞ physical entity</td>
</tr>
<tr>
<td>➞ entity</td>
</tr>
</tbody>
</table>

Table 1: Hypernyms of bass in the sense of sea bass.

WordNet also separates the hyponyms between types and instances. A chair is a type of furniture. Hesse, however is not a type of author, but an instance of author. So an instance is a specific form of hyponyms, and these instances are usually proper nouns, describing a unique entity, such as persons, cities and companies. These instances go both ways, just like the types.

Meronymi, the part-of relationship is divided into three different types, member meronymi, part meronymi and substance meronymi. It also has it's counterpart, just like hyponyms, holonymi. Where meronymi is has-part, holonymi is part-of. And just like homonyms, meronyms are a transitive relationship. If a tree has branches, and a branch has leaves, the tree has leaves.

Part meronymi, which is the relationship most commonly associated with the word, describes parts of an entity. is the blue sentence an explanation/example for meronyms?

Substance meronymi describes substances contained in an entity. For example, using the word water in the sense of the chemical substance H2O, it has substance hydrogen and substance oxygen.

The last subset of meronymi, member meronymi, describes the relationship of belonging to a larger kind of group. Looking at the word tree again, we can see that it is a member of the entity forest.
wood and woods. See table 2 for a description of the different types of meronymis.

![Table](image)

Table 2: Different types of meronymi used in WordNet.

*Antonyms* describe words that are semantically opposed. If you are a parent, you cannot be a *child* in the sense of someone's child. However, they do not have to rule out one another. Even though *poor* and *rich* are antonyms, just saying that one is rich does not automatically mean that they are *poor.*

1.2.2 Verbs

Verbs, just like nouns, have the hypernym relationship. Where the counterpart to hypernyms in the case of nouns is called hyponyms, this relationship among verbs are called *troponyms.* These go from the event to a superordinate event, and from an event to a subordinate event, respectively. Troponyms can also be described as in which manner something is done, therefore explaining the difference of names. Antonym also exists for verbs, and functions the same way, *stop* is an antonym of *start.*

The third relation, *entails* goes from an event to an event it *entails.* Entailment is used in pragmatics to describe a relationship between two sentences, where the truth condition of one sentence depends on the truth of the other. If *sentence A is true, then sentence B also has to be true.* For example: *If A entails B “The criminal was sentenced to death” (A), and “The criminal is dead” (B). If A is true, then B also has to be true.* This is the kind of relationship described by entailments in WordNet. If you snore, you are also sleeping, which is represented as an entailment relation of the two words, and thus you have an entailment mapping from *snore* to *sleep.*
1.2.3 Adjectives and adverbs

Adjectives are mostly organized in the terms of antonymi. As in the case of nouns and verbs, these are words which have meanings that are semantically opposed. As all words in WordNet, they are also part of a synset. The other adjectives in this particular synset also have their antonyms, and thus the antonyms of the other words become indirect antonyms for the synonyms.

*Pertainyms* is a relation which points the adjective to the nouns that they were derived from. This is one of the relations that cross the part of speech, though there are a few rare cases in which it points to another adjective. There's an extra paragraph for cross relations, maybe put it there?

The amount of adverbs are quite small. This depends on the fact that most of the adverbs in English are derived from adjectives. Those that does exist are ordered mostly in the same way that adjectives, with antonyms. They also have a relationship that is like the pertainym relation of adjectives, which also is a cross part of speech pointer, and points to the adjective that they were derived from. there's also the relation "similar to" an "also see" which connects as it sais "similar adjectives, that are not so synonymous to be in obe synset. you could explain that as well.

1.2.4 Relations across part of speech

Most of the relations in WordNet are relations among words of the same part of speech. There are however some pointers across the subfields the part of speeches it consists of. One has already been mentioned, pertainyms, which points from an adjective to the noun that it was derived from. Other than that, there are pointers that points to semantically similar words which share the same stem, called *derivationally related form*. For many of these pairs of nouns and verbs, the thematic role is also described. The verb *kill* has a pointer to the noun *killer*, and *killer* would be the agent of *kill*.

2. Using WordNet for Natural Language Processing

There are several subfields in natural language processing which can benefit from having a large lexical database, especially one as big and extensive as WordNet. Obviously, many semantic applications can draw benefits from using WordNet, including WSD and sentiment analysis. Many papers have been published regarding WordNet and WSD, exploring different approaches and algorithms, which is the main field for using this. In fact, WordNet can be said to be the de facto standard knowledge source for WSD in English.[4] This success depends on several factors. It is not domain specific, it is very extensive and publicly available.

Since WSD has been the subfield which has used WordNet most extensively, this is what will be good :) focused on here. Though, an interesting mention is that there do exist packages to access WordNet in several programming languages, including Perl and Python. For Python, the Natural Language Tool Kit (NLTK), which offers many modules and tools to analyze and process natural language and is widely used, has tools for using WordNet, such as finding synsets and other relations between words.
2.1 WordNet for Word Sense Disambiguation

WSD is a field which has been around since humans have tried to process natural language with computers. It is has been described as an AI-complete problem and is considered to be an intermediate step in many NLP tasks. The two main approaches to solving this problem are knowledge-based methods and supervised methods. Supervised methods suffer from sparseness in data to train on, in contrast to syntactic parsing, where there exist many resources of tagged data to work with. SemCor is a subset of the Brown Corpus, tagged with senses from WordNet. 186 files out of the 500 that constitutes the Brown Corpus have tags for all of the content words (nouns, verbs, adjectives and adverbs) and another 166 files have tags for the verbs. Even if this may be sufficient for evaluation, it is not enough for building a robust system for WSD. The knowledge-based methods use some kind of knowledge source, such as WordNet, to retrieve word senses. It is for these methods that WordNet has been used extensively.

WordNet keeps occurring in papers regarding WSD to this date. Due to the knowledge-based methods using WordNet performs worse than supervised methods, approaches to extend the knowledge contained in WordNet have been proposed. They range from semantically tagging the glosses in WordNet to enrich the semantic relations, to extracting knowledge from Wikipedia. Combining WordNet with ConceptNet, a semantic network which contains semantic relations, to improve performance have also been proposed.

3. Discussion

WordNet is an impressive database, with its large amount of words and the encoding of the relations. Also being freely available makes it very practical to use for natural language processing, just as it has been. However there are quite a few things that may speak against it. The very fine-grained distinctions in the database can be problematic for several tasks. Difficulty, for example, has four different senses in WordNet, all of them very similar, and can be hard to set apart, just not for computers, but also for humans. As such, not all senses may be relevant to disambiguate a word. Other problems may be that it was mainly annotated and tagged by humans, which may produce some inconsistencies, and that it was not produced just to solve NLP tasks.

WordNet is still widely used by people working in semantic natural language processing, as can be understood when reading papers, specifically regarding WSD. This can be seen in recent research, where WordNet have not been abandoned, but instead been used in combination with other resources, or has been tried to be improved in different ways. And since WordNet 3.0, it also contains a corpus of semantically annotated disambiguated glosses, which itself can prove to be useful. WordNet will be used for a time to come for WSD, mostly because the sparseness of data for supervised methods. Improvement of the lexical knowledge and algorithms to use for this may be the best way to go for the time being.

and since WordNet also contains a corpus, itself can prove to be WordNet will be mainly used for WSD? may be the best wary to find a good sollution for the NLP tasks/WSD
Bibliography


I liked your discussion a lot, especially what the problems of WordNet are. It would be nice if you could extend / restructure the "WordNet for WDS" part, so one can understand what these two methods are, what the difference between them is, how WordNet is used in these approaches and what method performs better. I also like your tables with examples for the WordNet relations. Have a merry Chrismas :)