Visualizing WordNet Structure

Jaap Kamps

Abstract

Representations in WordNet are not on the level of individual words or word forms, but on the level of word meanings (lexemes). A word meaning, in turn, is characterized by simply listing the word forms that can be used to express it in a synonym set (synset). As a result, the meaning a word in WordNet is determined by its sets of synonyms. This is essentially a recursive definition of word meaning. Hence meaning in WordNet is a structural notion: the meaning of a concept is determined by its position relative to the other words in the larger WordNet structure. We have implemented a set of scripts that visualize the WordNet structure from the vantage point of a particular word in the database.

1 Introduction

This paper report on visualization tools for Princeton's WordNet lexical database (Miller, 1990; Fellbaum, 1998). One of WordNet's greatest assets is its wide coverage of the English language. The down-side of this coverage is that it is highly non-trivial to get a good overview of particular parts of the lexical database. We want to visualized the WordNet structure from the vantage point of a particular word in the database. We will focus here on WordNet's main relation, the synonymy or SYNSET relation. The notion of meaning used in WordNet is lexical meaning, and the SYNSET relation denotes coincidence of lexical meaning. So our goal is to visualize parts of the WordNet SYNSET structure.

2 Relatedness and Minimal Path-Length

The first problem we face it that simply plotting all SYNSET relation immediately results in a knotted graph that fails to provide insight in the underlying WordNet structure. That is, we need to find a way that abstracts from the synonymy relation while still preserving the WordNet structure. For this reason, we investigate distance measures.

We will define the notion of n-relatedness based on the SYNSET relation (this is similar to the graph-theoretic notion of connectedness).

Definition 1 *Two words w_0 and w_n are n-related if there exists an $(n+1)$-long sequence of words $\langle w_0, w_1, \ldots, w_n \rangle$ such that for each i from 0 to $n-1$ the two words w_i and w_{i+1} are in the same SYNSET.*

For example, the verbs 'be' and 'endure' are 2-related since there exists a 3-long sequence $\langle \mathsf{be}, \mathsf{live}, \mathsf{endure} \rangle$. Two words may of course be related by many different sequences, or by none at all. We will only be interested in the shortest sequences relating words.

Definition 2 *Let* MPL *be a partial function such that* $\mathsf{MPL}(w_i, w_j) = n$ *if n is the smallest number such that w_i and w_j are n-related.*

If there is no sequence relating the two words, then the minimal path-length is undefined.

The minimal path-length enjoys some of the geometrical properties we might expect from a distance measure.

Observation 1 *The minimal path-length is a* metric, *that is, it gives a non-negative number* $\mathsf{MPL}(w_i, w_j)$ *such that*

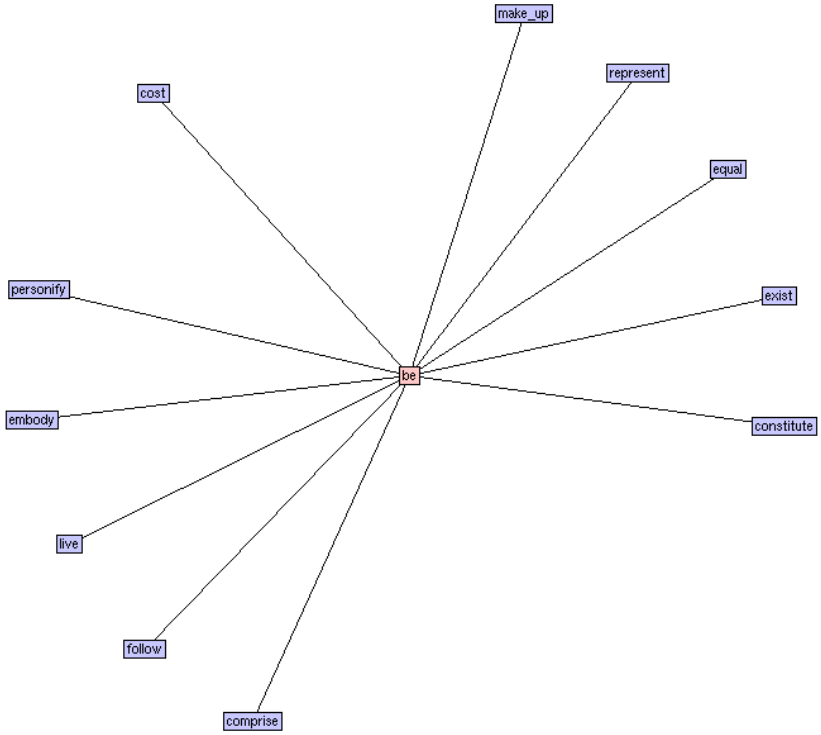

Figure 1: The WordNet database from the vista point of verb ‘be’ and maximal MPL of 1. The edges are SYNSET relations, nodes are only connected by a shortest path.

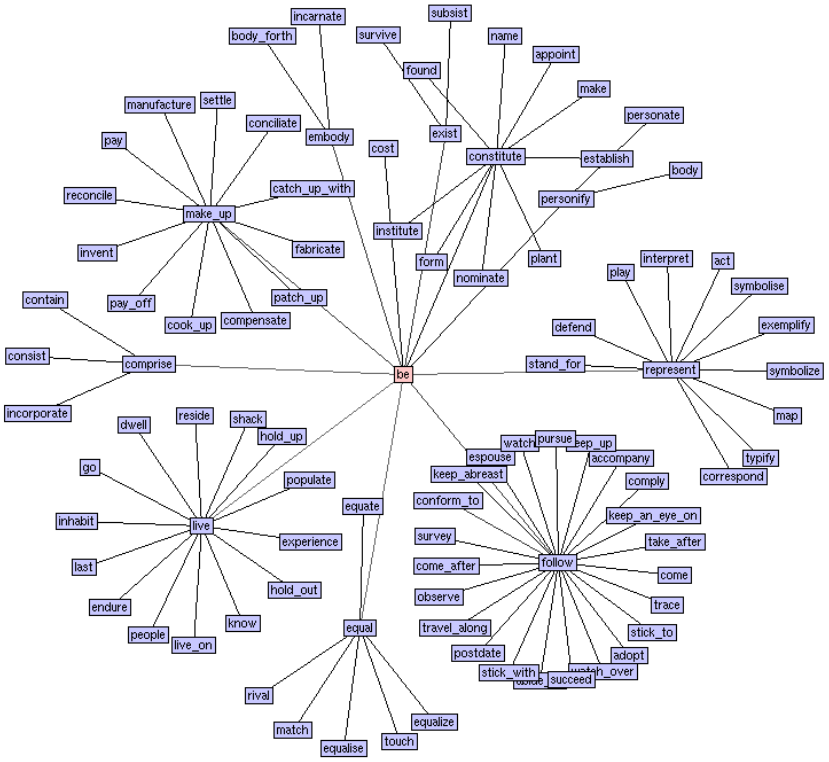

Figure 2: The WordNet database from the vista point of verb ‘be’ and maximal MPL of 2.

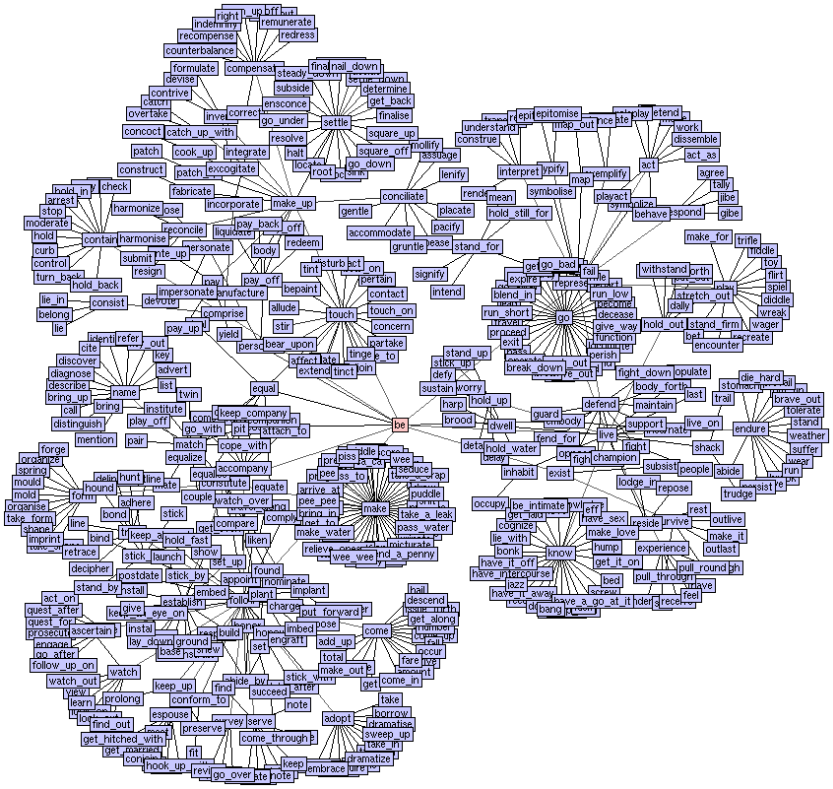

Figure 3: The WordNet database from the vista point of verb ‘be’ and maximal MPL of 3.

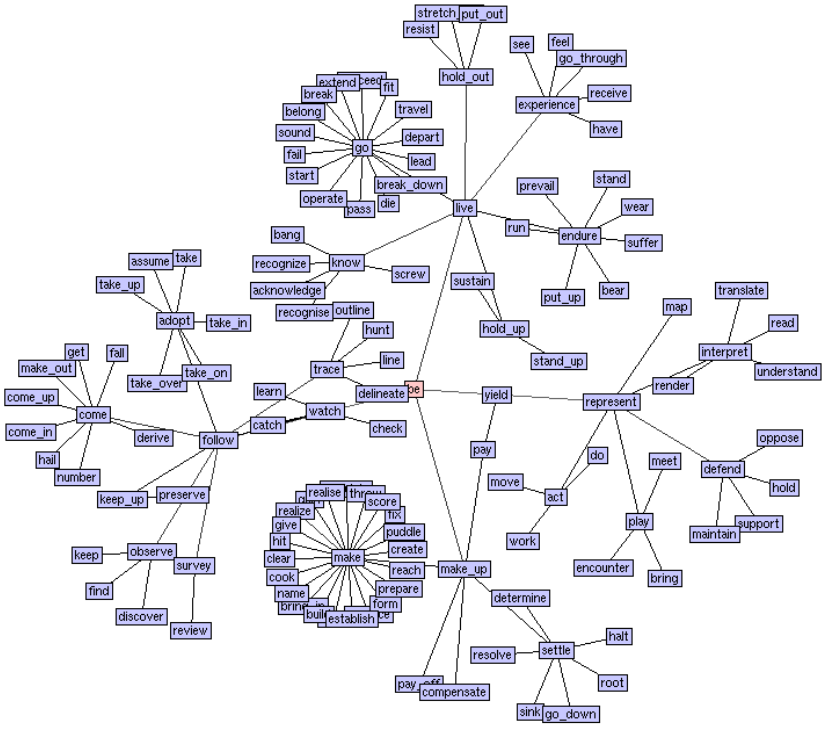

Figure 4: The WordNet database from the vista point of verb ‘be’ and maximal MPL of 3 and polysemy count ≥ 5.

i) $\mathsf{MPL}(w_i, w_j) = 0$ *if and only if* $w_i = w_j$,
ii) $\mathsf{MPL}(w_i, w_j) = \mathsf{MPL}(w_j, w_i)$, *and*
iii) $\mathsf{MPL}(w_i, w_j) + \mathsf{MPL}(w_j, w_k) \geq \mathsf{MPL}(w_i, w_k)$.

The minimal path-length is a straightforward generalization of the synonymy relation. For example, using WordNet we now find that

1. $\mathsf{MPL}(\mathsf{be}, \mathsf{live}) = 1$,
2. $\mathsf{MPL}(\mathsf{be}, \mathsf{endure}) = 2$,
3. $\mathsf{MPL}(\mathsf{be}, \mathsf{suffer}) = 3$, and
4. $\mathsf{MPL}(\mathsf{be}, \mathsf{lose}) = 4$.

Our strategy will be to start with a particular word, and draw the graph of words upto a certain MPL. This makes sense considering the SYNSET relation in WordNet is representing similarity of meaning, and our MPL is a straightforward generalization of the SYNSET relation. So the resulting graph still preserving the crucial WordNet structure.

3 Implementation

Our implementation consists of three major ingredients:

1. For a given word we can derive SYNSET related words from Princeton WordNet 1.7.
2. We use *Perl* and Dan Brian's `Lingua::Wordnet` module for efficiently deriving sets of words upto a given MPL.
3. We generate output in the appropriate form for the standard *Java* `Graph.java` class. This java class for the lay-out of graphs will take care of the visualization proper.

The only new part is a *Perl* script that can efficiently generate related words by their MPL. The script starts with a particular word (such as 'be') and recursively generates all synonyms while filtering away words it has encountered earlier. That is, we start with a particular word w (i.e., having minimal path-length zero to itself), then generate all words w_i with $\mathsf{MPL}(w, w_i) = 1$, then with $\mathsf{MPL}(w, w_i) = 2$, etcetera, until the search exhausts, or until we reach a given maximal value of MPL. For every new word, we also keep track of the word whose synonym it is. When finished, we will simply add a node for each word, and draw an edge to the word whose synonym it is. That is, for each related word, we only add one edge corresponding to (one of) its minimal paths to the initial word. In this way, the graph visualizes a small subset of the SYNSET relation, precisely those that give rise to minimal paths. To make the graph more appealing, we can influence the length of this edge by giving it a weight, which we let decrease as a function of the MPL. This list of nodes and edges is fed to the Graph Layout Java script, which will take care of the actual visualization.

Consider that we want to know the WordNet structure in the neighborhood of the verb 'be.' Figures 1 and 2 show screendumps of the graphs for $\mathsf{MPL} \leq 1$ and ≤ 2, respectively. The Java script is dynamic in the sense that the nodes can be manipulated. Although the graphs based on MPL are much sparser than the full SYNSET relation, the graphs get crowded when we increase the maximal MPL. This is simply due to rapid increase in the number of words, see for example figure 3. For this reason, the script has an additional argument that allows us to ignore words with a low word familiarity or polysemy count. See figure 4 for the same part of the WordNet lexical database, while filtering away words with polysemy count ≤ 4.

4 Conclusions and Discussion

One of the original design principles of WordNet is the use of a differential theory of lexical semantics (Miller, 1990). Representations in WordNet are not on the level of individual words or word forms, but on the level of word meanings (lexemes). A word meaning, in turn, is characterized by simply listing the word forms that can be used to express it in a synonym set (synset). As a result, the meaning a word in WordNet is determined by its sets of synonyms. This is essentially a recursive definition of word meaning. Hence meaning in WordNet is a structural notion: the meaning of a concept is determined by its position relative to the other words in the larger WordNet structure.

In this paper, we discussed the visualization of WordNet structure from the vantage point of a particular word. That is, we want to position ourselves on a particular word, and overview the larger structure of WordNet from there. This approach reminds of the perspective of modal operators in logic (Blackburn et al., 2001). This way of visualizing local parts of the WordNet database has proven its use for testing and evaluating WordNet similarity measures (Kamps and Marx, 2001).

Since WordNet's main SYNSET relation is too rich to allow for direct visualization, we focused on its straightforward generalization, the minimal path-length—a distance metric. Such measures of distance, similarity, or relatedness are well-known in natural language processing. The use of path-length as similarity metric also discussed in (Rada et al., 1989).

The basic notion of meaning used in WordNet is lexical meaning, and WordNet's main SYNSET relation is denoting coincidence of lexical meaning. Interestingly, WordNet is partly inspired by psycholinguistic theories of human lexical memory. That is, the meaning of words is also determined by its place in the larger structure of the database. Also note that this larger structure shows some resemblance with our own lexical memory. This may explain some of the intuitive appeal of the generated graphs.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grants # 400-20-036). Thanks to Patrick Blackburn, Maarten Marx, Michael Masuch, Rob Mokken and Ivar Vermeulen for their comments. All data is derived from Princeton WordNet 1.7, using Perl and Dan Brian's excellent `Lingua::Wordnet` module, and the `Graph.java` class.

On-line examples of the WordNet visualization scripts are available at the following URL: `http://www.illc.uva.nl/~kamps/wordnet/`.

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION
UNIVERSITY OF AMSTERDAM
NIEUWE ACHTERGRACHT 166
1018WV AMSTERDAM
THE NETHERLANDS
kamps@illc.uva.nl

References

Patrick Blackburn, Maarten de Rijke, and Yde Venema. 2001. *Modal Logic*, volume 53 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, Cambridge UK.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication Series. The MIT Press, Cambridge MA.

Jaap Kamps and Maarten Marx. 2001. Words with attitude. Technical Report PP-2001-16, Institute for Logic, Language and Computation, University of Amsterdam.

George A. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312. Special Issue.

Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:17–30.