The Topology of WordNet: Some Metrics

Ann Devitt and Carl Vogel

Computational Linguistics Group, Trinity College, Dublin, Ireland,
Email: devitta@cs.tcd.ie, vogel@cs.tcd.ie

Abstract. This paper outlines some different metrics intended for measuring node specificity in WordNet. Statistics are used to characterise topological properties of the overall network.

1 Introduction

Much work has been done on the notion of semantic relatedness between nodes in WordNet, (see [1] for a comprehensive survey of relatedness measures). This paper addresses a similar question – how comparable are two synsets in the WordNet network, not in terms of their content but in terms of the level or granularity or specifity they represent.

Although WordNet is a substantial knowledge base, it is not comprehensive. We do not know of work that records comparisons with arguably comparable resources like that supplied by CYC [3], however we expect that variant sparseness of coverage is endemic to all comparable knowledge bases. The level of detail in certain domains is essentially an accident of production dependent on the day, on the lexicographer, on the level of interest, etc. (for a case in point, note the recent addition of numerous concepts related to terrorism in WordNet 2, given the current political climate). Applications that use the data in WordNet to carry out some NLP task may themselves be subject to its vagaries. For example, two towns of comparable size in Ireland, Limerick and Drogheda, Limerick is encoded as both a port city and a type of poem where as Drogheda is encoded as a battle, being the site of a 16th century battle. A topic identifier using WordNet as its knowledge base might identify texts about Drogheda to be historical or military, without the second possibility of the topic relating to modern day Ireland.

The aim of this paper is to record statistics about version 1.1.7. that are relevant to our ongoing work in defining a notion of specificity that is determined by the topology of WordNet, and sensitive to variance in coverage across topic areas in WordNet. The measures are applicable to any knowledge source that has a definable topology. The results here are based on an amalgamation of link types assumed in WordNet but, a clear generalization is to factor in link types among nodes. Topological definitions in networks of heterogeneous links have been proposed before [6]. However, it is not yet clear whether any are fully appropriate to the sort of reasoning one would wish to do with WordNet.

The paper is divided into sections each detailing some basic measures for WordNet that characterize its overall topology: graph and node type §2 taxonomic distribution §3, parentage §4, node degree §5, depth and height §6 and clustering coefficients §7. Section 8 sets out some conclusions regarding what information has been gained on how these measures may be combined in an effort to determine node specificity in WordNet.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 106–111.

2 Some Basic Measures

WordNet [2] version 1.1.7 contains 74488 noun synsets. As this paper deals with the *structure* of WordNet rather than its content, we refer to WordNet and its synsets in terms of a graph, a directed acyclic graph and not a tree as it allows multiple inheritance. Henceforth, we use "node" and "synset" interchangeably. Of these synsets or nodes, 58586 or 78.65% are leaf nodes, leaving 15902 internal nodes. Analysis of particular measures across WordNet, such as height and branching factor, must take account of the fact the almost 60,000 leaf nodes may and often do skew results.

3 Dimensional Distribution

There are nine designated most general root nodes to dimensions of the taxonomy, namely:

1 Entity
2 Abstraction
3 Group
4 Act, human action, human activity
5 Psychological feature
6 State
7 Phenomenon
8 Event
9 Possession

The node distribution in these hierarchies is set out in bar chart 1.[1]. As we can see from the chart, the Entity hierarchy is by far the largest and as such merits some investigation as a separate unit. This is concrete evidence of an aspect of the variance mentioned in §2.

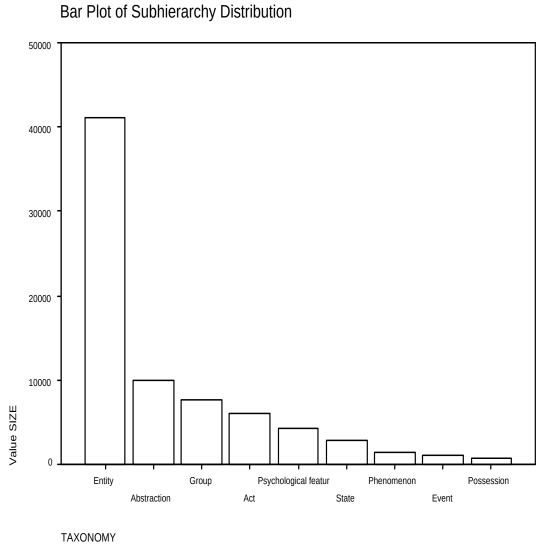

Fig. 1. Bar chart of synset distribution in top hierarchies

[1] The numbers refer to the numbers in the above list

4 Multiple Inheritance Quantified

As noted above, the taxonomy does allow multiple inheritance.

Example 1. The node referring to the multi-talented “Harley Granville-Barker” inherits from the more general nodes: “actor”, “critic”, “theatre producer”, “director” and “playwright”

Example 2. Similarly here, the more general “sphere” and “model” nodes are parents of the synset representing “globe”

In all, these multiple inheritance nodes amount to just 2.28% of the total taxonomy. The histogram in Figure 2 shows the distribution of nodes with more than one parent according to their depth in the hierarchy. The histogram would strongly suggest that these multiple inheritance nodes are normally distributed throughout the depth of WordNet and, thence, their effects propogate down the hierarchy.

However, according to a χ^2 test of independence the distribution of multiple parent nodes in the hierarchy is significantly different within different subhierarchies, χ^2 (8, N=75180)=324.27, p$\leq$0.001. Thus multiple inheritance is significantly more prevalent in certain sub-hierarchies.

One would expect that multiple parentage would imply a more specific concept node, from a content point of view. One might also posit that nodes deeper in the hierarchy are more specific. In this case, synsets in the right tail should be of comparable high specificity. Content inspection reveals the following as a sample of the highly-specific concepts in the right-tail of the distribution.

Example 3. sea bass, cytology, self-condemnation, bombardon

While nodes in the left tail, though with multiple parents, are less specific due to their position in the hierarchy

Example 4. person, artefact

It should be noted that multiple inheritance does not entail an overlap across sub-hierarchies. Only 689 synsets inherit from two distinct subhierarchies and of these only 6 inherit from more than two.

We hope to combine these topological measures to give a dependable measure of content specificity.

5 Branching Factor

The measure of node degree or branching factor here assumes the notion of dominance. Hence,

BranchingFactor= NoOfDescendants + 1 (the node itself).

This is to avoid problems with zero values in subsequent metrics and corresponds to the normal definition of dominance as a reflexive relation [4].

Branching factor (BF) in WordNet ranges from 1 to 573 with an average value of 2.023. Excluding leaf nodes (i.e., BF=1), however, the average branching factor value rises to 5.793.

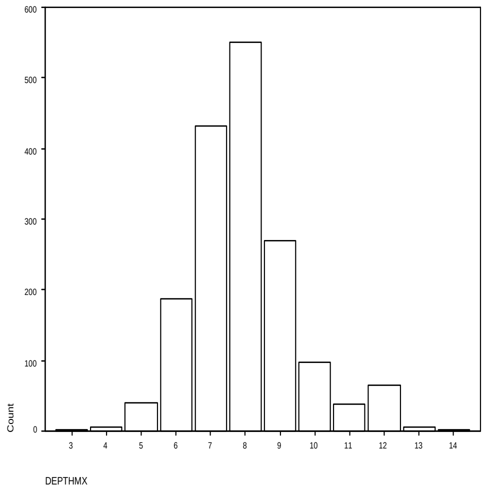

Fig. 2. Histogram of depth for nodes with multiple parentage

Indeed, 75.8% of the remaining 15902 nodes have a branching factor of less than 5 and almost 97% a value of less than 20.

A χ^2 test for BF>4, shows a significant difference in distribution in the phenomenon sub-hierarchy, χ^2(1,N=16406) = 11.23, p$\leq$0.001 alone.

This suggests that overall, in all subhierarchies, the structure is not shallow: small branching with a large number of total nodes suggests greater overall depth in paths. In the following section, we explore the notion of depth further.

6 Depth and Height

As each node may be parent to or descendant of several lineages, nodes may have several possible values for both height and depth. The values discussed here are

- Maximum depth: longest path from node to a top taxonomy node,
- Minimum depth: shortest path from node to a top taxonomy node,
- Maximum height: longest path from node to a leaf node, and
- Minimum height: shortest path from node to a leaf node.

The distribution of depth values in WordNet whether maxima or minima is normal (see figure 3). The data excluding leaf nodes is not substantially different. The means differ by 0.5 (7.1 with leaf nodes, 6.6 without) but the distribution is comparable.

The data for height, however, displays the effects of the preponderance of leaf nodes in the taxonomy.[2] The maximum distance from any node to a leaf node is 5. Two-thirds of all internal nodes are a single node from the bottom of the taxonomy and 93.6% of

[2] Both the data including and the data excluding leaf nodes display the same characteristics. Therefore we confine the discussion to maximum and minimum heights over all of WordNet

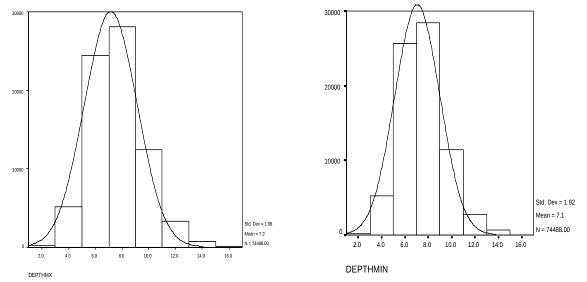

Fig. 3. Histogram of Maximum and Minimum Depth

nodes are a mere 1 or 2 nodes from a leaf node. In fact, for all values of the minimum height variable, the distribution of the depth variable is normal. Figure 4 shows that for both maximum and minimum height values, the distribution is common in natural language: a Zipfian distribution, decreasing at an exponential rate.

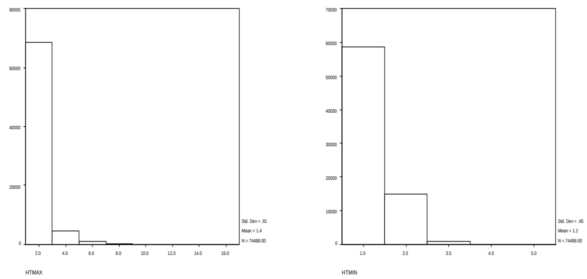

Fig. 4. Maximum and Minimum Height from a leaf node

Given the distribution of measures of height in WordNet, it would seem that depth may be a better measure of specificity. A minimum height value of 2 does little to suggest how precise a concept may be, for within this selection are the following sample nouns: *production, voodoo, group, refracting telescope, citizenry and floor*.

It should be noted that the distribution of these measures is similar within the nine sub-taxonomies of WordNet.

7 Clustering Coefficients

Clustering coefficients as a fine-grained measure of graph topology and connectivity have been posited in [7]. It measures the relative number of connections between neighbouring nodes in a network, hence, how clustered an area of a network may be. The formula to calculate the clustering coefficient C_i of a node i is as follows, where k_i is the number of connections to its neighbouring nodes and E_i is the number of connections between those k_i nodes.

$$\frac{2\Sigma_i}{k_i(k_i - 1)}$$

Higher-order coefficients measure connectivity between a node's immediate and more distant neighbours to a specific distance. The coefficient gives a normalized measure of connectivity across a whole graph.

A first point to note is that the basic cluster coefficient is not useful for a graph such as WordNet. Only 62 synsets have a coefficient higher than zero. This would indicate that the nodes in WordNet do not form strong clusters readily. This is clearly due to the hierarchical rather than network structure of the taxonomy.

The higher order measure, taking immediate neighbours and nodes at one extra remove, is a more useful value, particularly for internal nodes, where the distribution is normal and the mean is 0.337.

This would suggest that although WordNet is not tightly clustered, its nodes may form clusters of wider diameter.

8 Some Conclusions on Node Specificity Measures

The measures set out in the previous sections go some way to outlining the topology of WordNet. We have looked at the contrasting distributions of depth and height, the related concepts of branching factor and cluster coefficients, the notion of multiple inheritance and its significance within the taxonomy.

A model of the topology of WordNet would be useful in guiding interpretation of its content, particularly for non-humans, somewhat in the same way as Sperber and Wilson's relevance theory [5] requires a specific logic to guide inference steps. The more information we have about the shape of the structure in abstract, the more we way be able to extract from the knowledge base in particular.

We are currently working on a qualitative evaluation of various composite measures, combinations of the metrics discussed here using Principal Components Analysis and heuristics, in order to determine specificity of nodes in WordNet.

References

1. Alexander Budanitsky: Lexical Semantic Relatedness and its Application in Natural Language Processing Tech. Rep. CSRG-390 Department of Computer Science, University of Toronto (1999).
2. Christiane Fellbaum: WordNet,an electronic lexical database The MIT Press (1990).
3. D. B. Lenat, R. V. Guha: Building Large Knowledge Based Systems, Reading, Massachusetts: Addison Wesley (1990).
4. Barbara Partee, Alice ter Meulen and Robert Wall: Mathematical Methods in Linguistics, Kluwer Academic Publishers (1993).
5. Dan Sperber and Deirdre Wilson: Relevance: Communication and cognition (2nd ed.) Oxford: Blackwell, (1995).
6. David Touretzky: The Mathematics of Inheritance Systems, Los Altos, CA: Morgan Kaufman (1986).
7. D. J. Watts and S. H. Strogatz: Collective dynamics of small world networks, Nature **401**, 130 (1999).