Ollscoil Chathair Bhaile Átha Cliath
Dublin City University, Glasnevin, Dublin 9, IRELAND.

Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words

by

R.Richardson
A.F. Smeaton
J. Murphy

School of Computer Applications
Working Paper: CA-1294

Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words

R. Richardson, A. F. Smeaton and J. Murphy

School of Computer Applications
Dublin City University
Glasnevin, Dublin 9, IRELAND.
e-mail: { RRICHARD, ASMEATON, JMURPHY }@COMPAPP.DCU.IE

Abstract

In this paper we propose the use of WordNet as a knowledge base in an information retrieval task. The application areas range from information filtering and document retrieval to multimedia retrieval and data sharing in large scale distributed database systems.

The WordNet derived knowledge base makes semantic knowledge available which can be used in overcoming many problems associated with the richness of natural language. A semantic similarity measure is also proposed which can be used as an alternative to pattern matching in the comparison process.

1 Introduction

This paper describes a proposal to use WordNet in an attempt to introduce semantic knowledge to an information retrieval task. The initial information retrieval task focused on in this research was the locating and relating of information in large scale Federated database systems, in an application we call FEDDICT, [Rich94], however, our system is currently being evaluated in a document retrieval application. It is not difficult to see its applicability to many other aspects of information retrieval and filtering.

Traditional information retrieval systems are particularly susceptible to all the problems posed by the richness of natural language, in particular the polysemous nature of many natural language words and the multitude of ways in which the same

concepts can be described. Many widely used information retrieval systems are little more than clever pattern matching systems. There is no attempt to place index or query terms within an overall context, they are simply viewed as patterns to be matched. As Bates points out in [Bate86], "the probability of two persons using the same term in describing the same thing is less than 20%". As such, attempting to directly match user query terms against data set terms is likely to give bad results. This brings us to the need for a knowledge base that can distinguish different senses of words and can relate concepts that are semantically similar. A similarity function could then be derived which could give degrees of relatedness between index and query terms as opposed to a simple direct match or not.

The remainder of the paper is organised as follows. Section 2 describes how WordNet was adapted for use as a knowledge base. In section 3 there is a description of the proposed similarity estimator. Section 4 briefly comments on how the knowledge base has to be extended due to the requirements of the similarity estimator. The final section presents conclusions and recommendations for future work in this area.

2 WordNet

WordNet is the product of a research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English, [Mill90a, Mill90b, and Beck92]. The system has the power of both an on-line thesaurus and an on-line dictionary, and much more, (refer to Figure 1). Information in WordNet is organised around logical groupings called synsets. Each synset consists of a list of synonymous word forms and semantic pointers that describe relationships between the current synset and other synsets. A word form can be a single word or two or more words connected by underscores, (referred to as collocations). The semantic pointers can be of a number of different types including :

- Hyponym/Hypernym (IS-A/ HAS A)
- Meronym/Holonym (Part-of / Has-Part)
- Meronym/Holonym (Member-of / Has-Member),
- Meronym/Holonym (Substance-of / Has-Substance)

In this work we only use the nouns from WordNet as a knowledge base, ignoring the verbs, adjectives and adverbs. The initial knowledge base consisted of a number hierarchical concept graphs, (HCGs), constructed from WordNet data files. The root concepts of the HCGs were chosen as result of a set of experiments to determine what root concepts would, as a group, provide maximum coverage of the nouns in WordNet whilst minimising the degree of overlap between HCGs. The set of HCG roots we have used which achieves this is as follows :

- { Entity }
- { Location }
- { Abstraction }
- { Event }
- { Group }
- { Phenomenon }.
- { Psychological_feature }
- { Shape }
- { State }
- { Act }
- { Possession }

The resulting HCGs ranged in size from 43950 unique concepts (Entity) to 688 concepts (Shape). The HCGs are organised in the same manner as the WordNet data files, being accessible via index files which index concepts by their byte offsets in the HCG file. One shortcoming of this simple and efficient organisation is that extending the files "is almost impossible", [Beck93].

Constructing the KB in this manner has its advantages and disadvantages. A significant advantage is the fact that the resulting HCGs will serve as comprehensive starting points in obtaining HCGs that contain all relevant concepts in an information domain. Also, WordNet based HCGs will contain a comparatively rich set of semantic link types.[1] However, foremost in the disadvantages is the fact that links in the resulting HCGs are not weighted. Section 4 addresses this problem more fully.

---

[1] The concept graphs reported in [Rada89],[Kim90] and [Lee93] contain only IS-A links whilst the concepts in Ginsberg's WorldViews system, [Gins93], are only related by broader term and narrower term links.

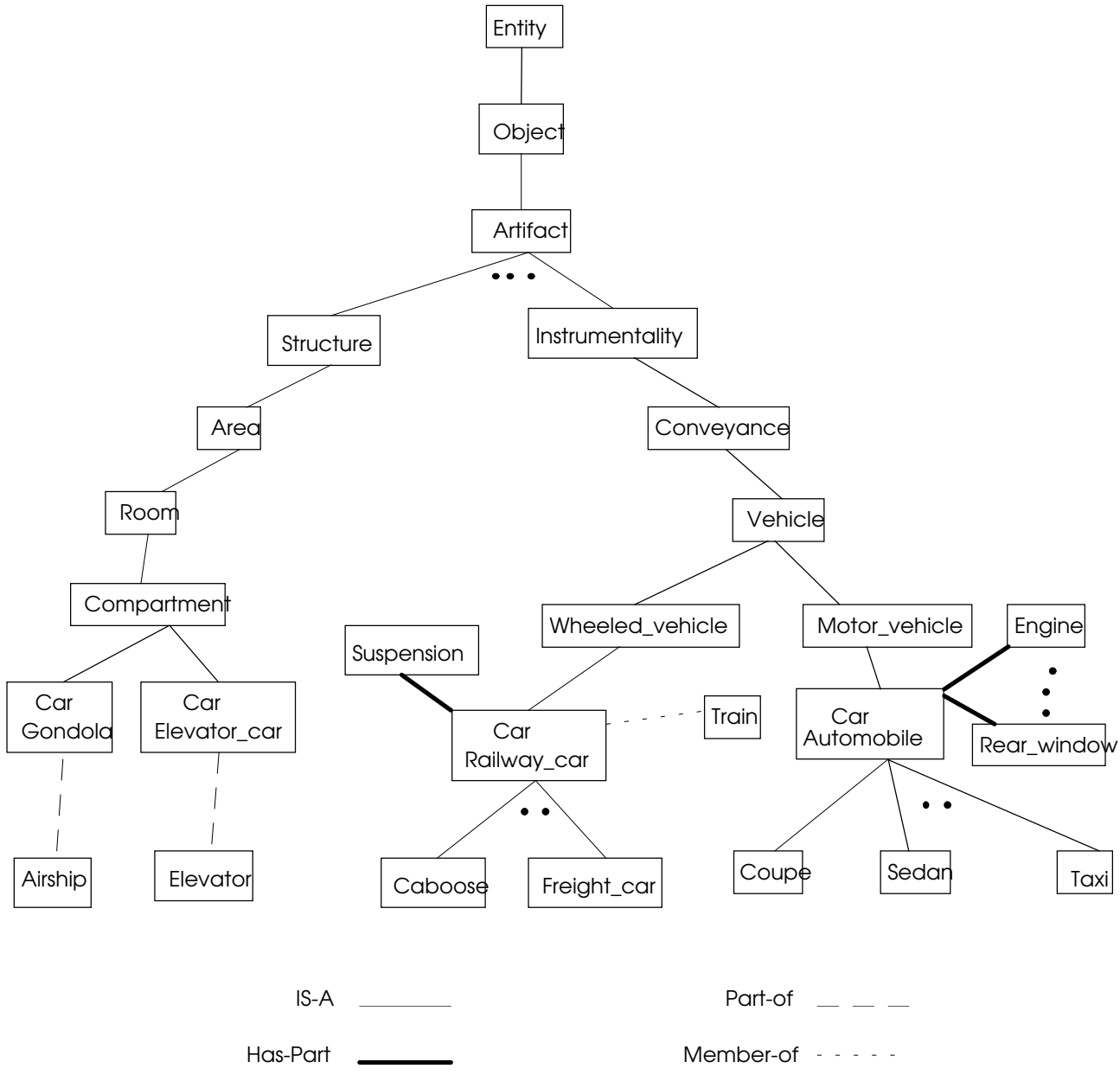


Figure 1 : WordNet Extract for the concept 'car'

3 Conceptual Similarity

A number of approaches to measuring conceptual similarity have been taken in the past. Tversky's feature based similarity model, [Tvers77], is arguably the most powerful similarity model to date. However, its applicability in WordNets situation

would require a much richer knowledge base than is actually the case. Although a WordNet derived knowledge base is quite thorough in its coverage of concepts the number of semantic relation types connecting these concepts is considerably less than would be required for use by a feature based similarity model.[2] As such, we employ a combination of a conceptual distance and an information based approach for estimating semantic similarity. The conceptual distance approach is based on the work of [Rada89, Kim90, and Lee93] and uses edge weights, (refer to section 4 for a discussion on how edges are weighted), between adjacent nodes as an estimator of semantic similarity.

The information based approach to measure semantic similarity is based on work carried out by Philip Resnick, [Resn93a, Resn93b]. Resnick views noun synsets as a class of words, the class is made up of all words in a synset as well as words in all directly or indirectly subordinate synsets. Conceptual similarity is considered in terms of class similarity. The similarity then between two classes is approximated by the information content of the first class in the noun hierarchy that subsumes both classes. The information content of a class is approximated by estimating the probability of occurrence of the class in a large text corpus[3] , (see appendix A for a discussion on class probabilities). As such, the similarity of two classes can be expressed as :

$$Sim(c_1,c_2) = \max_{c_i}[\log \frac{1}{P(C_i)}] \qquad (1)$$

where {Ci} is the set of classes dominating both C1 and C2, P(Ci) is the class probability of class Ci, and $\log 1/P(C_i)$ is the information content of class Ci.

---

[2] It is intended in the future to extend WordNet to include relation types of the form ATTRIBUTE-OF and FUNCTION-OF which will connect WordNet's adjective and verb collections with its noun collection. These developments should considerably enhance WordNets applicability to feature based similarity models.

[3] Implemented using 11 million noun occurrences from the Wall Street Journal with special handling for collocations

The methodology could probably be best illustrated by example[4]. If we assume we wish to discover the similarities between the following classes : 'car', 'bicycle', 'banana', and 'fork'. Taking first Sim(car, bicycle), we see that WordNet has six classes to which both 'car' and 'bicycle' are subordinate :

Synset	**Info_Content**
< vehicle >	2.500
<conveyance >	2.433
<instrumentality>	1.338
< artifact >	0.980
< object >	0.763
< entity >	0.565

If one takes the similarity measure as being the maximum information content value amongst the set of classes that subsume both synsets then SIM(car,bicycle) = 2.5. Notice that, as would be expected, classes grow more frequent and as such less informative as one moves higher in the hierarchy. Since 'car' and 'bicycle' have some specific (therefore informative) classes in common, one can conclude that they are similar. In contrast, the other examples yield the following :

Sim(car,fork)		**Sim(car,banana)**	
<instrumentality>	1.338	< object >	0.763
< artifact >	0.980	< entity >	0.565
< object >	0.763		
< entity >	0.565		

As such, cars and forks seem considerably less similar than cars and bicycles, however they are more similar than cars and bananas. This can be explained in that forks and cars are objects that people make (artifacts), whereas all that can be said in terms of the similarity of cars and bananas is they are both nonliving things (object).

Following some informal experimentation with the use of the conceptual distance measure, we found some general concerns with regard to the sole use of this measure as an estimator of conceptual similarity. Due to the comparatively broad

[4] This replicates one of Resnicks examples using version 1.4 of WordNet

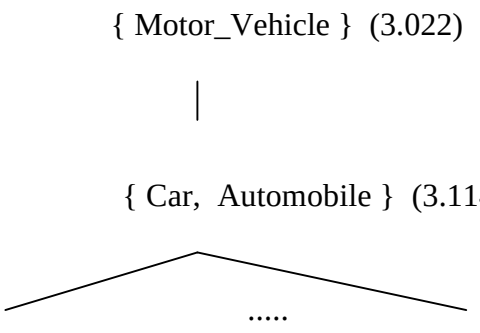
domainedness of the WordNet derived HCGs, (as compared with those of Rada who worked solely in the medical domain), the conceptual distance measures were less accurate than expected. The situation was improved to a large degree when it was decided to include the non-hierarchical link[5] types in the distance calculation. However, the conceptual distance measure is still particularly susceptible to vagaries of the builders of WordNet. In particular the organisation of concepts within WordNet can often be puzzling. The irregular densities of links between concepts results in unexpected conceptual distance measures. These are typically as a result of expected links between concepts not being present. Also due to the general operation of the conceptual distance similarity estimator, most concepts in the middle to high sections of the HCG, being geographically close to each other, would therefore be deemed to be conceptually similar to each other. Although the depth scaling factor in the link weighting mechanism softens the overall effect in many cases, sometimes the general structure of the WordNet derived HCGs cannot be overcome by link weighting without causing serious side effects elsewhere in the KB.

We believe the weaknesses of conceptual distance as an estimator of conceptual similarity can be addressed to a certain degree by the inclusion of Resniks conceptual similarity measure. It is expected that a combined conceptual similarity measure would not suffer as severely from the absence of expected links. Of course, the question arises of why combine, why not just use Resniks measure ? The answer can be found in the fact Resnik's proposed measure is not itself without its weaknesses. Perhaps foremost is the fact that his technique ignores information in WordNet that may be useful. Only the synonym and IS-A relations are used, the other relation types, which are used effectively by the adapted Rada-like conceptual distance approach, are overlooked. A second weakness is apparent in the method of calculating the information content of classes. Many polysemous words and multi-worded synsets will have an exaggerated information content value. If one takes for instance the word 'bank', the information content for this word will include all occurrences of bank in the corpus, regardless of meaning. This gives the same

---

[5] The PART-OF, MEMBER-OF and SUBSTANCE-OF links are described as the non-hierarchical links.

(exaggerated) information content value to a 'commercial bank' and a 'river bank'. Also, due to the fact information content values are calculated for synsets as opposed to individual words, it is possible for the information content value to be over exaggerated in situations where synsets are made up of a number of commonly occurring ambiguous words. If one takes for example the synset *{ yield, fruit },* the information content value of this synset is calculated both from the frequencies of the word *'fruit'* and the word *'yield'*. Given the fact that the information content of a class is defined in terms of the information contents of its subordinate classes, super classes of classes containing polysemous words are similarly over-valued. This disregard of ambiguous words is a particular problem given the fact that classes in WordNet refer to particular senses. A final caveat apparent with the information theoretic approach to semantic similarity is the fact two different concepts can be more similar to each other than another concept is to itself. The effect of this can be more clearly seen with the following example :



Above is an extract from the KB, the numbers in brackets after the synsets are the information content values. From here we can see the information based estimate of the similarity between an ambulance and a convertible (car), 3.114, is closer than the estimated similarity between a motor vehicle and itself, 3.022.

It is thus proposed in our research to more fully evaluate both measures and to investigate the possibility of combining both measures so as to take advantage of the stronger aspects of each approach and to compensate for individual weaknesses.

4 Weighting a HCG

The conceptual distance estimator of semantic similarity requires the edges between concepts in the KB to be weighted. However, unlike the concept graphs of other researchers, ([Gins93], [Rada89], [Kim90], and [Lee93]), those created for our research are very large, containing of the order of tens of thousands of nodes. For this reason, the usual process of hand weighting each link is not viable and a method of automatically weighting each link had to be developed. Initial research in this area was based on Botafogo's work on node metrics in hierarchical hypertexts, [Bota92]. However, our research was subsequently considerably influenced by that of Sussna, [Suss93].

Certain observations can be made with regard to conceptual distance in HCGs which can aid in the process of automatically determining the weight of edges. For instance, the value for the weight of a link is affected by the following :

(a) the density of the HCG at that point
(b) the depth in the HCG
(c) the strength of connotation between parent and child nodes.

With regard to the width, it can easily be seen that different parts of a HCG are denser than others. It is commonly held that a link in a dense part of the hierarchy represents a smaller conceptual distance than a link in a less dense region. In terms of the depth it can be said that distance shrinks as one descends down a HCG. To explain, suppose there are two only sibling relations, one near the top of the hierarchy and one deep down in the detailed portion of the HCG. For example, suppose the node 'Living Thing' is high in the hierarchy and it only has two children nodes, 'Plant' and 'Animal'. These two siblings are far apart conceptually when compared against the two siblings 'Wolfhound' and ‘Foxhound' under the parent 'Hound' deep down in the HCG. Finally, Figure 2 illustrates the point regarding the local strength of connotation.

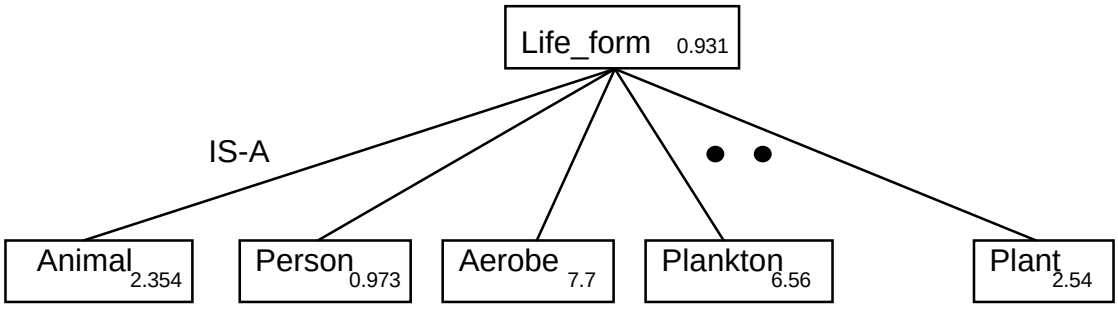


Figure 2 : KB Extract

. It can be argued the parent node *Life_Form* is more strongly connotated with the child nodes *Animal*, *Plant*, and *Person*, than with the nodes *Aerobe* and *Plankton*.

At present, the density of a HCG for a specific link type is estimated by counting the number of links of that type. The strength of connotation of the link being weighted is estimated as a function of its information content value, and those of its sibling and parent node, (the numbers in Figure 2 are the nodes information content values). The result of these two operations is then scaled by dividing by the depth of the link in the HCG[6] .

5 Conclusions and Future Work

The research described in this paper is ongoing. At present facilities exists for constructing HCGs, weighting HCGs under a number of different strategies, computing semantic similarity using both conceptual similarity measures and a process

---

[6] An investigation is currently underway into improving this scaling factor by using the information content values of HCG synsets. One of the inconsistencies left by the builders of WordNet is the fact that two concepts that could be thought of as being at the same level of abstraction are at different levels from the root because of the order of the hierarchy in WordNet. If you take for example 'horse' and 'cow', (both being regarded as being of the same level of abstraction), the node for 'horse' is 10 levels from the root, taking 'entity' as the root concept, and one for 'cow' is 13 levels deep. As such, because there is a large body of information in WordNet for one concept relative to another, the weightings for the link from that concept are unfairly penalized.

which performs word sense disambiguation of text. Evaluation of various configurations of the system are currently underway. The application chosen for the evaluation is document retrieval using the Wall Street Journal text corpus along with TREC, [Harm94], queries and evaluation procedures. Basically the system is given a natural language query and is expected to rank the top thousand or so documents from the set of 153,000 WSJ articles with respect to their similarity to this query. The querying strategy is to compare each query term against all the index terms of an article and to aggregate all comparisons to give an overall score for the relevance of that article to the query. The comparison mechanisms are the information based and conceptual distance semantic similarity estimators. A traditional pattern matching IR system using tf/IDF term weightings is used a baseline to compare overall results.

An unfortunate prerequisite to this application is the assumed existence of a sense disambiguator which can automatically tag words from the WSJ articles with their appropriate KB meanings. Given the fine sense distinctions WordNet makes, the semantic tagger had to be particularly perceptive and accurate. However, an informal analysis of the semantic tagger has shown very promising results. Also, sample runs of the query matcher have been likewise, very promising.

Future work includes a complete analysis of results from the large scale evaluation, an investigation into an appropriate method of combining both semantic similarity estimators, a more rigorous evaluation of the semantic tagger, and further development of the automatic HCG weighting strategy.

References

[Bate86] : Bates M. (1986). Subject Access in Online Catalogs: A Design Model, *Journal of the American Society for Information Science*, 11, 357 - 376.

[Beck92] : Beckwith R. and Miller G. A. (1992). Implementing a Lexical Network, Report No. 43, *Princeton University*.

[Beck93] : Beckwith R., Miller G. A., and Tengi R. (1993). Design and Implementation of the WordNet Lexical Database and Searching Software, Working Paper, *Princeton University*.

[Bota92] : Botafogo A., Rivlin E., and Shneiderman B.(1992). Structural Analysis of Hypertexts: Identifying Hierarchies and useful Metrics, *ACM Transactions on Information Systems*, 10, (2), 142 - 180.

[Gins93] : Ginsberg A. (1994). A Unified Approach to Automatic Indexing and Information Retrieval, *IEEE Expert*, 8, (5), 46-56.

[Harm94] : Harman D. K. (1994). Overview of the Second Text Retrieval Conference (TREC-2), *The Second Text Retrieval Conference (TREC-2)*, Gaithersburg, Maryland, Aug 31 - Sept 2, 1993, 1-20.

[Kim90] : Kim Y. W. and Kim J. H. (1990). A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph, *Journal of Documentation*, 46, (2), 113-137.

[Lee93] : Lee J. H., Kim M. H., and Lee Y. J. (1993). Information Retrieval Based on Conceptual Distance in IS-A Hierarchies, *Journal of Documentation*, 49, (2), 113 - 136.

[Mill90a] : Miller G. A., Beckwith R., Felbaum C., Gross D., and Miller K., (1990). Introduction to WordNet : An On-line Lexical Database, *International Journal of Lexicography*, 3, (4), 235 - 244.

[Mill90b] : Miller G. A. (1990). Nouns in WordNet : A Lexical Inheritance System, *International Journal of Lexicography*, 3, (4), 245 - 264.

[Rada89] : Rada R., Mili H., Bicknell E. , and Blettner M., (1989). Development and Application of a Metric on Semantic Nets, *IEEE Transactions on Systems, Man, and Cybernetics*, 19, (1), 17-30.

[Rich94] : Richardson R., Smeaton A. F. and Murphy J. (1994). Using WordNet for Conceptual Distance Measurement, *BCSIRSG colloquim*, Glasgow, March 22-23, 1994.

[Resn93a] : Resnik P. (1993), Selection and Information : A Class based Approach to Lexical Relationships, *PhD. dissertation at the University of Pennsylvania*. Also appears as Technical Report 93-42, November.

[Resn93b] : Resnik P., (1993). Semantic Classes and Syntactic ambiguity", *ARPA Workshop on Human Language Technology*, Princeton, March.

[Suss93] : Sussna M. (1993). Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, *Proceedings of CIKM*.

[Tver77], : Tversky A., (1977). Features of Similarity, *Psychological Review*, 84, (4), 327 - 352.

Appendix A

Calculation of Class Probabilities

Class probabilities are used in the determination of the information content or specificity of WordNet classes. The specificity of a class can be defined in terms of its class probability as follows :

$$Specificity(C_i) = -\log(P(C_i))$$

where P(Ci) is the class probability of class *i*.

In order to define the probability of a class we must first define *words(c)* and *class(w)*. *Words(c)* is defined as the set of words in all directly or indirectly subordinate classes of the class c. For example *Words(cloister)* consists of *religious residence, convent, abbey, friary, monastery, nunnery,* and *priory*. *Classes(w)* represents the set $\{c|w \in words(c)\}$, i.e. this includes all the classes in which the word *w* is contained, regardless of the particular sense of *w.* From these two definitions we can define the frequency of a class as :

$$Freq(C_i) = \sum_{w \in words(c)} \frac{1}{|classes(w)|} \times Freq(w)$$

where Freq(w) is the frequency of occurrence of word *w* in a large text corpus. The class probabilities can be estimated from such a distribution using maximum likelihood estimation (MLE) :

$$P(c) = \frac{Freq(c)}{N}$$

where N is defined as $\sum_{c'} Freq(c')$, i.e. the total size of the sample.