Statistical Overview of WordNet from 1.6 to 2.0

Jiangsheng Yu*, Zhenshan Wen, Yang Liu, and Zhihui Jin

Institute of Computational Linguistics, Peking University

Abstract. We defined several discrete random variables and made their statistical comparisons between different versions of WordNet, by which the macroscopical evolution of WordNet from 1.6 to 2.0 is explored. And at the same time, the examples of extreme data will be enumerated during the experimental analysis.

Keywords *WordNet, distribution, Kolmogorov-Smirnov test*

1 Introduction

For a complex machine-readable dictionary like WordNet [3], it is difficult to compare versions by all the modifications in details [12]. Yet, sometimes we indeed feel a stable trend with more updatings. In the following sections, we will define several discrete random variables and explore their statistical properties in WordNets. For convenience, only the noun and verb parts are considered.

Table 1. Amount of SynSets and words in WordNet

Amount	#NounSynSet	#VerbSynSet	#Noun	#Verb
WN1.6	66,025	12,127	94,474	10,319
WN1.7	75,804	13,214	109,195	11,088
WN2.0	79,689	13,508	114,648	11,306

The first random variable (rv), say F, is the amount of instant hypernyms that a SynSet has, whose distribution indicates the uniqueness of induction along the hypernymy tree. The second rv M describes the polysemia of English words. The third rv W measures the size of SynSet, i.e., how many words a SynSet contains. The fourth rv S depicts the amount of hyponyms a SynSet has, by which we are able to learn about the reification of concepts. Lastly, we will show the distribution of category, associated with which the distribution of category depth is studied. The examples of extreme data are enumerated during the experimental analysis and some further work will be mentioned in the conclusion.

A nonparametric method named *two-sample Kolmogorov-Smirnov goodness-of-fit test* is used in the version comparison.

* This research is supported by Beijing Natural Science Foundation, No. 4032013 and National Project 973, No. G1998030507-4. All the data we used in the paper are available at `http://icl.pku.edu.cn/yujs`.

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 352–357.

2 Uniqueness of Induction

In WordNet, concept is represented by a SynSet formally. Among the SynSets various relations are defined, where the hypernymy one is the most important. It is very convenient to make induction along the hypernymy tree, which provides us an easy way of reasoning based on the semantic distances. By the fact that a SynSet may have several father-nodes in the net despite of the categories, we surveyed the random variable F, the amount of instant hypernyms each SynSet has, and summarized the data in Table 2.

Table 2. Observations of F in noun and verb SynSets

	#NounSynSet in		
F	WN1.6	WN1.7	WN2.0
0	9	9	9
1	65,144	73,997	77,594
2	852	1,751	2,016
3	18	40	54
4	2	6	12
5	0	1	3
6	0	0	1

	#VerbSynSet in		
F	WN1.6	WN1.7	WN2.0
0	617	626	554
1	11484	12557	12923
2	26	31	31

In WordNets, the noun concept that has the most hypernyms is {*Ambrose, Saint Ambrose, St. Ambrose*}, and then $\{atropine\}$.

The two-sample Kolmogorov-Smirnov goodness-of-fit test (the usual nonparametric approach to testing whether two samples are from the same population when the underlying distributions are unknown, abbreviated by *KS-test*, see [2,6]) denies that the cumulative distribution function (cdf) of F in the noun part of WordNet invariably keeps along the version updatings except from WN1.7 to 2.0 ($ks = 0.0036$ and p-value$= 0.9957$). Apropos of verb concepts, the percentage of roots is much bigger than that of noun concepts. The fact of few instances of multiple hypernyms predicates that the verb concepts are well congregated. For example, the sense 4 of *warm up* is verb concept with two hypernyms. By the KS-test, the distribution of verb hypernym varies much in every version updating. From WN1.6 to 1.7, many verb SynSets with single hypernym were added. And in the latest updating, quite a few roots have been merged. The mean of noun and verb hypernyms is 1.027 and 0.9613, respectively.

3 Polysemia

The cardinality of the meanings of each word in WordNet is a random variable, say M, that can imply the polysemia of English words [10]. The noun with the most meanings in WordNets is *head*, then *line*, and the most meaningful verb is *break*, then *make*.

The KS-test predicates that the polysemia of nouns changes little only from WN1.7 to 2.0 ($ks = 0.007$ and p-value$= 1$), and same thing happens to the verbs ($ks = 0.005$, $p -$ value $= 0.9989$). Additionally, the mean of senses can be found in Table 6. A further work includes the estimation of sense distribution of the frequent words in practice.

Table 3. Polysemia of nouns and verbs in WordNet

	#Noun in		
M	WN1.6	WN1.7	WN2.0
1	81,910	94,714	99,365
2	8,345	9,416	9,912
3	2,225	2,710	2,859
4	873	1,027	1,113
5	451	535	565
⋮	⋮	⋮	⋮
32	0	1	1

	#Verb in		
M	WN1.6	WN1.7	WN2.0
1	5,752	5,948	6,110
2	2,199	2,499	2,508
3	979	1,085	1,094
4	502	580	604
5	318	357	360
⋮	⋮	⋮	⋮
63	1	1	0

4 Size of SynSet

The size of a SynSet, written by W, is the amount of words it contains, which provides us a cue of word substitution and corpus extension. The largest SynSet in WordNets is {*buttocks, nates,* $\cdots$*, ass*}, and then {*dohickey, dojigger,* $\cdots$*, thingummy*}. The Sense 4 of *love* is the largest verb SynSet, then the senses of *botch* and *bawl out*.

Table 4. Observations of W in noun and verb SynSets

	#NounSynSet in		
W	WN1.6	WN1.7	WN2.0
1	33,926	38,576	40,753
2	21,214	24,158	25,160
3	6,640	8,126	8,502
4	2,551	2,984	3,159
5	973	1,099	1,178
⋮	⋮	⋮	⋮
28	0	1	1

	#VerbSynSet in		
W	WN1.6	WN1.7	WN2.0
1	7,032	7,630	7,855
2	2,782	3,047	3,106
3	1,181	1,271	1,264
4	539	600	608
5	270	318	314
⋮	⋮	⋮	⋮
24	0	0	1

The KS-test detects the diverse distributions of SynSet size in WordNets, except the verb parts of WN1.7 and 2.0 ($ks = 0.0051$ and p-value$= 0.9938$). This conclusion does not contradict with that in Section 2, since SynSet size has nothing to do with the hypernymy relation. From the statistical facts of F and W, we are able to comprehend their distinct functions in lexicographic analysis. In addition, the mean size of SynSets in distinct WordNets is calculated in Table 6.

5 Reification of Concepts

The hyponyms (or troponyms) are usually used as the extension of retrieval word. For instance, the hyponyms of *disaster* $\in$ {*calamity*, *catastrophe*, *disaster*, $\cdots$} include

Table 5. Observations of S in noun and verb SynSets

	#NounSynSet in		
S	WN1.6	WN1.7	WN2.0
0	51,446	59,693	62,870
1	5,214	5,800	6,069
2	3,003	3,297	3,410
3	1,808	1,930	1,994
4	1,080	1,178	1,229
5	701	782	833
⋮	⋮	⋮	⋮
619	0	0	1

	#VerbSynSet in		
S	WN1.6	WN1.7	WN2.0
0	9,069	9,986	10,234
1	1,355	1,426	1,444
2	568	595	593
3	338	328	338
4	212	234	235
5	124	138	148
⋮	⋮	⋮	⋮
393	0	0	1

{*plague*}, {*famine*}, etc. The amount of hyponyms (or troponyms) a SynSet has is a random variable of our interest, denoted by S in this paper.

The noun concept in WordNet that has the most hyponyms is {*city, metropolis, urban center*}, then {*bird genus*}, {*writer, author*}, {*mammal genus*}. Sense 2 and 1 of *change* has the most troponyms, and then {*be*}. The KS-test verifies that the distribution of S in verb SynSets is unaltered from WN1.7 to 2.0 ($ks = 0.0034$ and p-value$= 1$). The SynSets with no hyponyms are leaves of the hypernymy trees, whose complement is the set of inner concept nodes. For the leaves are useless for the extension of retrieval word, we examined the inner nodes and found the same result as the forenamed ($ks = 0.0092$ and p-value$= 0.9987$). For the cdf of S, the similarity between WN1.7 and 2.0 is larger than that between WN1.6 and 1.7. The data in the parentheses are the means of inner hyponyms, as a comparison of those without restrictions: see Table 6.

Table 6. Mean of senses, mean size of SynSets, and mean of (inner) hyponyms

WordNet version	Noun senses	Verb senses	Noun SynSets	Verb SynSets	Noun hyponyms	Verb hyponyms
1.6	1.231	2.138	1.762	1.820	1.013 (4.589)	0.9513 (3.772)
1.7	1.234	2.180	1.777	1.829	1.024 (4.820)	0.9550 (3.909)
2.0	1.236	2.179	1.778	1.824	1.027 (4.867)	0.9613 (3.966)

6 Distribution of Category

The amount of noun SynSets that each category contains is a random variable of interest, whose distribution represents an ontology of semantics. Although the KS-test concludes that the distribution of category varies much during the version updatings, but the shape of distribution keeps well that means the ontology of WordNet develops consistently. The numeralization of ontology and its application makes the evaluation possible.

The deepest path of induction in each category is called the *category depth*. It is not the case that the more SynSets a category has the deeper it is, e.g., category 6 and 30. Intuitively,

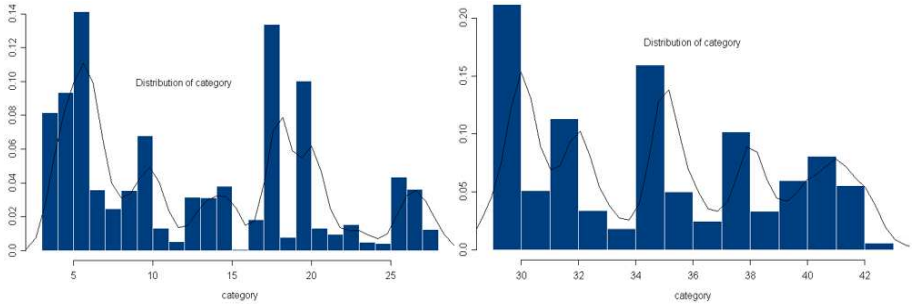

Fig. 1. Distributions of noun and verb categories

the depth of verb category varies less than that of noun category. The noun (verb) category depth reaches the maximum at category 5 (category 41), where 1, 2, 3 denotes WN1.6, 1.7, 2.0 respectively.

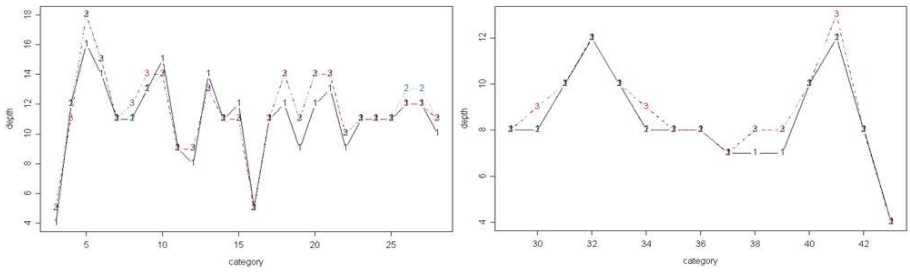

Fig. 2. Scatter plots of noun and verb category depth

Comparing the scatter plots with the histograms of category, there is no obvious relationship between the depth and the distribution. A heuristic explanation of those counterexamples is that the knowledge representation in WordNet by hypernymy tree is notable in width sometimes.

Conclusion

As a linguistic comparison, the statistical survey of Chinese Concept Dictionary (CCD, see [8,13]), the Chinese WordNet, is under consideration. Also, the similar research of EuroWordNet [11] is still worthwhile.

To improve WordNet and its widespread applications (e.g., WSD in [1], text clustering ini [5], semantic indexing in [4,9]), there is still a lot of work to do. For instance, the more advanced coding of offset, the regular patterns of frequent words and concepts, the reasonable definition of semantic distance between concepts in WordNet, co-training between WordNet and its application (e.g., information retrieval, text categorization, attitude identification), etc.

Acknowledgement

We appreciate the persistent enthusiasm of all participants in the seminar of Machine Learning at Peking University. Special thanks are due to Prof. Shiwen Yu who is concerned about Chinese WordNet all the time.

References

1. S. Banerjee and T. Pedersen (2002), *An adapted Lesk algorithm for word sense disambiguation using WordNet*. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.
2. P. J. Bickel and K. A. Doksum (2001), *Mathematical Statistics – Basic Ideas and Selected Topics* (Second Edition). Prentice-Hall, Inc.
3. C. Fellbaum (ed) (1999), *WordNet: An Electronic Lexical Database*. The MIT Press.
4. C. Fellbaum, et al (2001), *Manual and Automatic Semantic Annotation with WordNet*. In Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations.
5. A. Hotho, S. Staab and G. Stumme (2003). *Wordnet improves text document clustering*. Submitted for publication.
6. E. L. Lehmann (1975), *Nonparametrics: Statistical Methods based on Ranks*. Hoden-Day, San Francisco.
7. Y. Liu, J. S. Yu and S. W. Yu (2002), *A Tree-structure Solution for the Development of ChineseNet*. The First Global WordNet Conference, Mysore, India, pp. 51–56.
8. Y. Liu, S. W. Yu and J. S. Yu (2002), *Building a Bilingual WordNet: New Approaches and Algorithms*. COLING 2002, Taiwan, pp. 1243–1247.
9. R. Mihalcea and D. I. Moldovan (2000), *Semantic indexing using WordNet senses*. In Proceedings of ACL Workshop on IR & NLP, Honk Kong.
10. W. Peters (2000), *Lexicalized Systematic Polysemy in WordNet*. Proc Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, pp. 1391–1396.
11. P. Vossen (1999), *Euro WordNet General Document*. University of Amsterdam. Available online at `http://www.hum.uva.nl/-ewn`.
12. J. S. Yu (2002), *Evolution of WordNet-like Lexicon*. The First Global WordNet Conference, Mysore, India, pp. 134–142.
13. J. S. Yu, Y. Liu and S. W. Yu (2003), *The Specification of Chinese Concept Dictionary*. Journal of Chinese Language and Computing, Vol. 13 (2), pp. 176–193.