Identifying Poorly-Defined Concepts in WordNet with Graph Metrics

John P. McCrae and Narumol Prangnawarat

Insight Centre for Data Analytics, National University of Ireland, Galway
john@mccr.ae, narumol.prangnawarat@insight-centre.org

Abstract. Princeton WordNet is the most widely-used lexical resource in natural language processing and continues to provide a gold standard model of semantics. However, there are still significant quality issues with the resource and these affect the performance of all NLP systems built on this resource. One major issue is that many nodes are insufficiently defined and new links need to be added to increase performance in NLP. We combine the use of graph-based metrics with measures of ambiguity in order to predict which synsets are difficult for word sense disambiguation, a major NLP task, which is dependent on good lexical information. We show that this method allows use to find poorly defined nodes with a 89.9% precision, which would assist manual annotators to focus on improving the most in-need parts of the WordNet graph.

Keywords: wordnet, language resource, data quality, graph metrics, lexical resources

1 Introduction

Princeton WordNet [1] is the most widely used lexical resource and even with the recent rise in deep learning and machine learning approaches to NLP, it has been shown [2, 3], that the best solutions (such as at SemEval 2016 [4]) to many tasks in natural language processing still rely on this resource. As such WordNet is one of the most vital resources for knowledge extraction and integration. However, there have also been many criticisms of WordNet as an unreliable and error-prone resource and there were significant quality issues ranging from misspellings and cycles in the hypernym graph[1] to issues with poor definitions [5]. Moreover, WordNet is a resource whose principal aim is to use a graph in order to describe the concepts of a language and the methods that build on it normally do not rely on textual descriptions of a concept but only its graph relationships. This is problematic as the average degree of the WordNet graph is only 2.43, which is significantly less than that of similar knowledge graphs such as DBpedia [6], which has an average degree of 6.39[2]. For some concepts this may be sufficient

[1] For example: https://lists.princeton.edu/cgi-bin/wa?A2=ind1509&L=wn-users&P=R2&1=wn-users&9=A&I=-3&J=on

[2] This is calculated as usual as the number of links (triples) divided by the number of nodes (entities) in the graph

to describe the meaning of the word, for example the concept 'Slovenian' is described only by the fact that it 'pertains to' the concept 'Slovenia', which in this case is a sufficient description, but for a more complex concept, in particular adverbial concepts such as 'fairly' many links would be required, yet WordNet frequently contains one or even zero links for adverbs.

Princeton WordNet is a manually-developed resource and its value as a gold-standard resource is one of the main reasons that it is so widely used in natural language processing. As such, a fully-automatic approach, such as [7] to the extension or improvement of this resource would not create a resource that can be applied with the reliability of WordNet. Due to recent changes instantiated by the Global WordNet Association to found an interlingual index [8], WordNet is developing from a resource that is developed by one institute for one language into a collaborative project considering multiple language and contributors. As such, it is wise to consider where this collaborative effort is best directed, and this paper's main contribution is to provide a function that can rank every node in the WordNet graph according to whether the description is sufficient for NLP tasks.

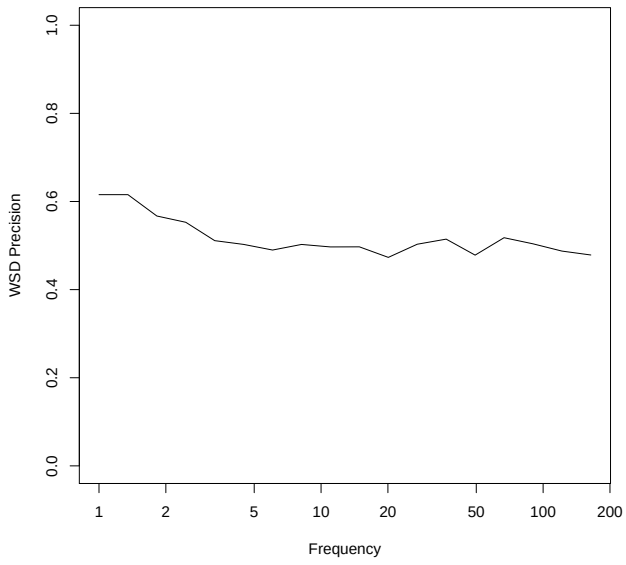

Fig. 1. Comparison of WSD precision against frequency

In order to estimate the quality of the graph, we use word sense disambiguation (WSD) as a proxy task for representing quality. This is for several reasons,

firstly that this has been established by other authors [9] as a suitable task for this purpose. Moreover, it is our intuition that the 'bad' nodes in the WordNet graph are those for which the graph does not provide sufficient information to describe the concept, and thus it follows that a WSD algorithm would also have problem with such concepts. Finally, there have been several methods identified recently [10], which can perform WSD, using only the WordNet graph and without any supervision, while still providing state-of-the-art WSD performance. As such, WSD seems to be the ideal task for the measurement of the quality of individual WordNet nodes.

This work is focussed on WordNet as a particular knowledge graph, as it is the most widely used graph and as it is manually constructed then we have a clear idea of how this analysis can directly help in the lexicon construction process. However, we note that this work is applicable to other forms of knowledge graphs such as DBpedia, and could help in the process of integrating automatically extracted taxonomies, with manually constructed lexicons. Moreover, many of the metrics here generically describe the structure of the graph and could be adapted for semantic similarity or even cross-lingual linking, which is of particular importance for the development of interlingual wordnets.

2 Related Work

The quality of a language resource, such as WordNet, affects its applicability for many tasks and has thus been the focus of many studies. One particular aspect has been a focus on the technical quality of the resource such as Lohk et al. [11], who looked at the quality of a graph by looking at existing patterns within the graph structure, which may be erroneous, or similar work by Liu et al. [12]. Other work on technical quality has focused on detecting empty, duplicate and logically unsound structures in wordnets [13]. Nadig et al. [14] examined the semantic correctness issue, in particular looking to validate if links in the graph could be validated based on corpus, definition or structural information. Another corpus-based approach to evaluating the quality of a wordnet was followed by Krsteve et al. [15]. These works differ from this paper crucially in that they detect where information is likely incorrect, whereas we focus on where data is absent.

Another aspect of quality has been fitting a second taxonomy, especially that of an upper-level ontology to WordNet, such as the work of Gangemi et al. [16], where WordNet was fitted to the DOLCE ontology, which was said to improve the hierarchy of WordNet. Similar to this Kaplen et al. [17] worked on examining the logical errors in wordnet in particular issues such as multiple inheritance and transitive inference of properties. It has however not been clearly shown that these structural issues impact actual applications, however a significant issue that has been detected is to do with *sense granularity*, that is the distinction between similar meanings. It has been shown [18], that a less fine-grained sense distinction is better for WSD and as such a more coarse-grained sense distinction has been used for the construction of wordnets in other languages [19].

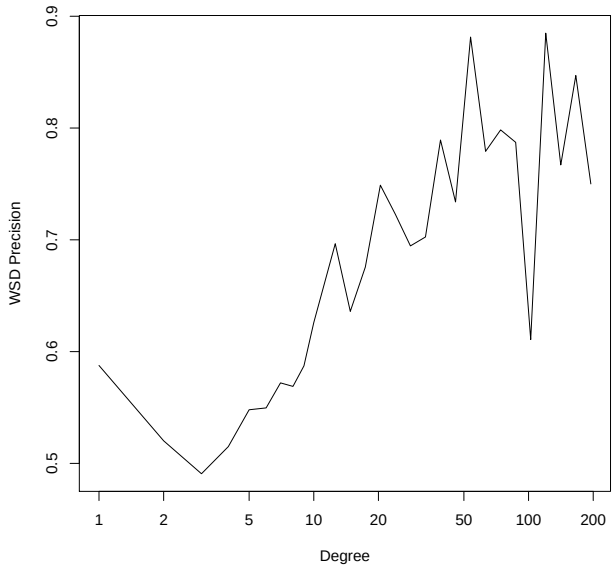

Fig. 2. Comparison of WSD precision against node degree

3 Methodology

3.1 Word Sense Disambiguation

Number of synsets	117,791
Number of links	285,688
Annotations in Brown corpus	234,136
Synsets at least once in Brown	31,755

Table 1. Statistics about Princeton WordNet 3.0 and the Brown corpus

In order to learn the quality of a single node in a WordNet graph, we need a proxy task in order to understand the effectiveness of the graph around a given concept. For this we use WSD and in particular we used the Personalized PageRank (PPR) algorithm developed by Agirre et al. [10]. We ran the standard mode of the PPR algorithm for every sentence in the Brown corpus, based on the sense annotation given in SemCor[3]. For each synset in the graph we aggregated

[3] http://web.eecs.umich.edu/∼mihalcea/downloads.html#semcor

the results of the output per synset, that is we counted the precision as how many times out of its occurrences in the gold-standard Brown corpus it was correctly identified by the PPR algorithm. For synsets that did not occur in the Brown corpus, we treated the precision as a missing value and did not use to learn the quality estimator.

In Figure 1, we see the comparison between the frequency of the synsets in the Brown corpus and the precision that was obtained in the WSD task. We observe that there is very little difference in performance for the higher-frequency concepts than for the low-frequency concepts. Instead in Figure 2, we compare the frequency to the node degree and in this case we see a very different result, suggesting that for the particular method of PPR, the degree of the node is a key predictor for the quality of a WordNet node. Both these graphs we generated by taking the precision of the WSD for synsets grouped by their degree or frequency.

3.2 Graph-based Metrics

Wordnet graph is constructed as a directed typed[4] graph $G = (V, E)$. V is a set of nodes where each node represents a synset s and E is a set of edges where each edge e_{ij} connects synset i and synset j that have any semantic relations. In other words, $e_{ij} \notin E$ if no semantic relation between synset i and synset j.

We observed (Figure 2) that the higher degree of the Wordnet synset, the better precision of WSD is, however the actual correlation of degree and WSD precision is very low. We would like to combine other graph properties to increase precision of WSD and introduce graph measures that we analyzed as features in this task.

Degree Degree is the simplest way to measure the importance of nodes in the network by counting edges connecting to the nodes. In this measure, all neighbors of a node are equivalent.

$$d(s) = |\{(s, s_j)\ inE\}|$$

Network Centralities One of the most important measures to rank the importance of nodes in a graph are centrality measures. We measured the following network centralities[5]:

Betweenness centrality measures a node by considering the shortest path from a node to itself.

Closeness centrality measures how far is a node to any other node in the network by considering the average distance from the node to every other node in the graph.

Eigenvector centrality [20] measures centrality of a node based on the centrality of its neighbors from the idea that a node becomes more important if

[4] The type of the links, such as 'hypernym', are ignored in this work

[5] We use the implementations provided by NetworkX (https://networkx.github.io) for our analysis.

it is connected to the important nodes, which can be found from the eigenvector of the adjacency matrix.

PageRank [21] normalizes centrality by dividing the centrality of a node by the number of the nodes it points to and distributing equally to them. The idea is that an important nodes may point to many different nodes but all its neighbors are not necessarily considered as high important nodes. A node with high centrality that points to many other nodes will pass a small amount of its centrality to the others.

Average degree of neighbors The average degree of neighbors of synset, where n is the number of the neighbors of synset s, is:

$$avg(d) = \frac{1}{n}\Sigma d(s_i)$$

The higher neighbor degree, the more information we can acquire from the synset neighbors.

Cycles and Triangles A cycle is a sequence of edges where the first node and the last node are the same node.

In particular, we analyzed the cycles of length 3 which are called triangles.

$$triangle(s) = \{(s, s_1, s_2); (s, s_1) \in E \wedge (s, s_2) \in E \wedge (s_1, s_2) \in E\}$$

Triangles can reveal how many synset neighbors have semantic relations with nearby neighbors whereas cycles can include distant neighbors.

Cluster coefficient The cluster coefficient measures the likelihood that the neighbors of each node will connect with each other. This measure is also used to find which nodes tend to be clustered together as relevant synsets.

We analyzed the cluster coefficient of a synset s as the following equation:

$$clust_s = \frac{2 \times triangle(s)}{d(s)(d(s) - 1)}$$

where $triangle(s)$ is a number of triangle of article s and $d(s)$ is the degree of the article s.

Many features, such as degree, exhibit a power-law distribution, therefore we experimented with applying log to all features and took the best performing version of the metric.

3.3 Word-based Metrics

In addition to graph-based metrics, much of the precision of WSD depends on whether the synset is ambigiuous. To this end, we developed features that decide

how ambiguous a particular synset is. The first measure is the log synset size defined for a synset $s = \{w_1, \ldots, w_n\}$

$$\text{log-size}(s) = \log(|s|)$$

Next we define the ambiguity of a word w as the number of synsets that w is part of, e.g.,

$$\text{ambig}_w(w) = |\{s : w \in s\}|$$

We then use log average ambiguity as follows:

$$\text{ambig}_s(s) = \log\left(\frac{\sum_{w_i \in s} \text{ambig}_w(w_i)}{|s|}\right)$$

Let $f(w, s)$ denote the frequency of word w with sense s in the Brown corpus. We denote such a sense as a most frequent sense, mfs, as

$$s \in \text{mfs}(w) \Leftrightarrow f(w, s) = \max_{s'} f(w, s')$$

Finally we define the MFS score for a synset as the percentage of senses for which it is the MFS

$$\text{mfs-score}(s) = \frac{|\{w \in s \wedge s \in \text{mfs}(w)\}|}{|s|}$$

4 Results

	Precision	Log Degree
Log Degree	0.081	1.000
Closeness centrality	-0.096	0.463
Log Average Neighbor Degree	-0.166	-0.164
Log Number of Cycles	-0.006	0.540
Log Number of Triangles	0.014	0.572
Log Eigenvector Centrality	-0.028	0.429
Log PageRank Centrality	0.142	0.980
Log Betweenness Centrality	-0.015	0.770
Log Clustering Coefficient	-0.012	0.420
Synset Size	0.019	0.208
MFS Score	-0.526	0.181
Ambig_s	0.436	-0.134

Table 2. Correlation of individual features with precision and log degree

In Table 2 we see the correlations between the individual features and the precision as predicted by the word sense disambiguation task, evaluated using

10-fold cross-validation. For these features we see that in general there is low correlation across all the features. This implies that no single measure of quality can be used to estimate whether a node will perform well at predicting precision for the WSD task.

Features	Classifier	Correlation	Absolute Error	Mean Squared Error
Graph Features	Linear	0.2341	0.4367	0.4600
Word Features	Linear	0.5556	0.3258	0.3934
Both Features	Linear	0.6319	0.3018	0.3667
Graph Features	Tree	0.3964	0.3910	0.4344
Word Features	Tree	0.5725	0.3154	0.3879
Both Features	Tree	0.6795	0.2569	0.3472

Table 3. Prediction of WSD precision based on features

Following this evaluation we combined all these features using two classifiers: A linear regression model, and the M5P decision tree algorithm [22] (all implementations were those provided by Weka[6]), the results are presented in Table 3. We present the Pearson's correlation (higher is better) as well as both the average absolute error and the mean squared error (lower is better). We also analyzed the effects of just the graph-based features (Section 3.2) and the word-based features (Section 3.3). We see that the word-based features are more important for predicting precision, however this is unsurprising as these features directly measure the ambiguity of a particular word. The purely graph-based features, however, still show strong performance and for all classifiers the combination of graph-based features and word-based features significantly outperforms other features.

Finally, to evaluate whether this achieves the goal of identifying low and high quality node, an expert on WordNet evaluated the top 50 highest scoring and top 50 lowest scoring entities. This was performed as a double-blind experiment, where the annotator had to rate the entries as "Completely lacking" if there were no semantic relations, "Majorly lacking" if there were only one or two semantic relations, "Slightly lacking" if was either a diverse set of relations or there were many relations of the same type (typically only 'hypernym'/'hyponym' relations) and "Sufficient" if there were many links of different types. The results are presented in Table 4 and show that our system can with high precision detect those nodes in need of improvement. The table also shows the average predicted precision score given by our system for each of the categories indicating a correlation between our systems evaluation and the annotator's opinion.

[6] http://www.cs.waikato.ac.nz/ml/weka/

	Top 50	Bottom 50	Average predicted precision
Completely lacking	1	36	0.07
Majorly lacking	0	5	0.05
Slightly lacking	34	8	0.80
Sufficient	14	1	0.92

Table 4. Ranking by WordNet expert of top 50 and bottom 50 synsets

5 Conclusion

We have presented a system for identifying nodes that have insufficient description in Princeton WordNet. We followed a model where we regressed a number of features to the per-synset precision on WSD. Two sets of features were examined: firstly, graph-based features looking at the structure of the wordnet graph around the node and secondly, word-based features, which measured the ambiguity of the synset. We found that both features were complementary and that the combination of these features was effective at predicting the quality of nodes. Our features do not yet consider the actual type of links in the wordnet graph and as future work, we plan to include these into our evaluation.

References

1. Fellbaum, C.: WordNet. Wiley Online Library (1998)
2. Rothe, S., Schütze, H.: Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). (2015)
3. Rychalska, B., Pakulska, K., Chodorowska, K., Walczak, W., Andruszkiewicz, P.: Samsung Poland NLP team at SemEval-2016 Task 1: Necessity for methods to measure semantic similarity. In: Proceedings of the 10th International Workshop on Semantic Evaluation. (2016) 614–620
4. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation. (2016) 509–523
5. Bond, F., Vossen, P., McCrae, J.P., Fellbaum, C.: CILI: the Collaborative Interlingual Index. In: Proceedings of the Global WordNet Conference 2016. (2016)
6. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al.: DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web **6**(2) (2015) 167–195
7. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193** (2012) 217–250
8. Vossen, P., Bond, F., McCrae, J.P.: Toward a truly multilingual Global Wordnet Grid. In: Proceedings of the Global WordNet Conference 2016. (2016)

9. Cuadros, M., Rigau, G.: Quality assessment of large scale knowledge resources. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. (2006)
10. Agirre, E., Soroa, A.: Personalizing Pagerank for word sense disambiguation. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2009) 33–41
11. Lohk, A., Fellbaum, C., Vohandu, L.: Tuning hierarchies in Princeton WordNet. In: Proceedings of the Global WordNet Conference. (2016)
12. Liu, Y., Yu, J., Wen, Z., Yu, S.: Two kinds of hypernymy faults in wordnet: the cases of ring and isolator. In: Proceedings of the Second Global WordNet Conference. (2004) 347–351
13. Smrž, P.: Quality control for wordnet development. In: Proceedings of the Second International WordNet conference. (2004)
14. Nadig, R., Ramanand, J., Bhattacharyya, P.: Automatic evaluation of wordnet synonyms and hypernyms. In: Proceedings of ICON-2008: 6th International Conference on Natural Language Processing. (2008) 831
15. Krstev, C., Pavlović-Lažetić, G., Obradović, I., Vitas, D.: Corpora issues in validation of Serbian WordNet. In: International Conference on Text, Speech and Dialogue, Springer (2003) 132–137
16. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A.: Sweetening WORDNET with DOLCE. AI magazine **24**(3) (2003) 13
17. Kaplan, A.N., Schubert, L.K.: Measuring and improving the quality of world knowledge extracted from wordnet. University of Rochester, Rochester, NY (2001)
18. Yong, C., Foo, S.K.: A case study on inter-annotator agreement for word sense disambiguation. In: Proceedings of the ACL SIGLEX Workshop on Standardizing Lexical Resources (SIGLEX99), College Park, MD. (1999)
19. Carpuat, M., Ngai, G., Fung, P., Church, K.: Creating a bilingual ontology: A corpus-based approach for aligning WordNet and HowNet. In: Proceedings of the 1st Global WordNet Conference. (2002) 284–292
20. Bonacich, P.: Power and centrality: A family of measures. American Journal of Sociology **92**(5) (1987) 1170–1182
21. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab (1999)
22. Quinlan, J.R., et al.: Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence. Volume 92., Singapore (1992) 343–348