Graph-Based, Supervised Machine Learning Approach to (Irregular) Polysemy in WordNet

Bastian Entrup

Applied and Computational Linguistics,
Justus-Liebig Universität Gießen, Germany
bastian.entrup@germanistik.uni-giessen.de

Abstract. This paper presents a supervised machine learning approach that aims at annotating those homograph word forms in WordNet that share some common meaning and can hence be thought of as belonging to a polysemous word. Using different graph-based measures, a set of features is selected, and a random forest model is trained and evaluated. The results are compared to other features used for polysemy identification in WordNet. The features proposed in this paper not only outperform the commonly used CoreLex resource, but they also work on different parts of speech and can be used to identify both regular and irregular polysemous word forms in WordNet.

1 Introduction

In [1, p. 16], regular polysemy is defined as follows:

> Polysemy of the word A with the meanings a_i and a_j is called regular if, in the given language, there exists at least one other word B with the meanings b_i and b_j, which are semantically distinguished from each other in exactly the same way as a_i and a_j and if a_i and b_i, a_j and b_j are nonsynonymous.

Often mentioned is the so-called grinding rule: the name of an animal can often also be used to refer to products gained from it. Irregular polysemy on the contrary covers those cases that do not exhibit such patterns. When for example animal names are used to denote humans, this can be done referring to different properties of animals. Calling someone a *lion* is mostly due to strength and courage, while *chicken* may refer to a lack of courage. Since the productive patterns of regular polysemy can be identified and used by computer systems while irregular cases are harder to identify, most computational approaches to polysemy based on WordNet (WN) [11] are focused on regular polysemy.

Since WN represents word senses rather than words, the classic definitions of polysemy cannot be applied to WN. Instead of looking at the binary decision of whether a *word* is polysemous or homonymous, the method proposed in this paper will look at the sense level and try to identify those (homograph) word forms that actually are related. Taking the word *bank* and all its word forms in

A. Przepiórkowski and M. Ogrodniczuk (Eds.): PolTAL 2014, LNAI 8686, pp. 84–91, 2014.

WN, the approach taken here is to connect all those forms that are related to the field of finance, while not connecting them to those forms that are related to a slope of any kind.

To achieve this goal, WN's network topology is exploited: measures such as the geodesic paths or properties of nodes (e.g., the closeness or betweenness) are used. Since it could be assumed that homonymous words – or more exactly in this context word forms – have no semantic similarity to each other, measures of semantic similarity based on paths between two nodes are taken into consideration to distinguish related and unrelated word forms.

Although the approach proposed here, especially the graph-based features, can be applied to all parts of speech (POSs) in WN, this paper is restricted to those word forms that are either nouns or potentially connected to nouns. This restriction is necessary to compare the proposed feature set to the CoreLex features.

2 Related Work

2.1 Regular Polysemy Detection in WordNet

The *CoreLex* resource [5] defines a set of 39 basic types (BTs), i.e., semantic classes of words that subsume a number of word senses in WN (e.g. *food* or *animal*). Taking advantage of the hierarchical order of nouns in WN, the BTs are assigned to anchor nodes in WN that are identified as the hypernym of the word senses belonging to the given semantic class. When looking at words, one lexeme is likely to have different word senses and hence different word forms in WN. Each word form is assigned to at least 1 of the 39 BTs, resulting in a list of BTs related to a word. This list should display patterns of regular polysemy. For example, a word like *lamb* has a meaning that belongs to the BT *animal* as well as one belonging to *food*, etc. This pattern can be found in other words as well. It therefore satisfies the definition of regular polysemy given above.

The approaches described in [2] and [16] are based on the CoreLex resource. [16] calculate a ratio of polysemy for words based on the BTs they are related to. The more words share the same pattern of BTs, the more likely those words are polysemous. Polysemy and homonymy are considered "two points on a gradient, where the words in the middle show elements of both" [16, p. 268].

[5], [16], and [2] can only be used to detect regular polysemy. The great number of irregular polysemous forms cannot be found and is, as done in [16], regarded as homonymous.

[17] takes a different approach and calculates lexical similarity of glosses of potentially related word senses sharing a common lemma. If the glosses are similar, their meaning is considered to be similar as well. This approach is applicable to other POSs than nouns since it does not rely on the hypernymy/hyponym relation – which in WN does not exist for adjectives or adverbs.

Table 1. Existing measures of semantic similarity/relatedness

Abbreviation	Source and description	POS
Resnik and Yarowsky	[14], implemented by [13]	N, V
Lin	[10], implemented by [13]	N, V
Jiang and Conrath	[8], implemented by [13]	N, V
Hirst and St. Onge	[7], implemented by [13]	N
Leacock, Miller, and Chodorow	[9], implemented by [13]	N, V
Wu and Palmer	[18], implemented by [13]	N, V
distance	geodesic path between the two nodes	all

2.2 Computing Semantic Similarity in WordNet

The most intuitive measure of semantic similarity in WN is to calculate the geodesic path (i.e., the distance) between two nodes.

A number of semantic similarity features that can be applied to WN have been proposed (see Table 1).[1] These are mainly based on the geodesic path between the two nodes in question. Most of these measures are restricted to the noun and verb subset of WN, since they rely on the hierarchical order of the noun or verb network.[2]

The features in Table 1 and others have been used in [15] to find synsets that are related and could be merged to make WN more coarse grained and thereby raise the accuracy in wore sense disambiguation based on WN. Furthermore, [15] propose calculating the distance of both word senses to their closest common hypernym.

3 Feature Set

The question of whether two word senses are related in meaning can be answered by calculating the semantic similarity of the word forms and by looking at the network toponymy (i.e., local and global features of the network). These include information on the degree of the nodes being examined, the nodes representing the synsets as well as the nodes representing the word forms connected to them. Furthermore, different centrality measures for these nodes are calculated. The centralities are thought of as giving insight to the position a node has within the network. For example, a node with a high closeness centrality [6] can be expected to show shorter geodesic paths to any other node, not only to those it is semantically related to. The betweenness [6] indicates the node's position on geodesic paths of other nodes. The eigenvector centrality [3] and the PageRank [12] are two further centrality measures that are likely to better indicate centrality than the degree of the nodes.

[1] Measures based on gloss overlap have been excluded.

[2] For information on the single features, see the sources given in Table 1.

Table 2. Proposed graph-based feature set

Abbreviation	Source or description	POS
isA-Rel	*is-A* relation between word forms	N,V
isDerivedFrom	is one word form derived from the other?	all
closeness	the closeness value of the node	all
betweenness	the betweenness value of a node	all
POS	the part of speech of word form	all
word sense degree	degree of the word sense nodes	all
synset degree	degree of the synset nodes	all
eigenvector centrality	the eigenvector centrality values of the nodes	all
page rank	the page rank values of the nodes [12]	all
POS	the part of speech of word form	all
sharedLemmas	number of lemmas shared by the synsets	all
minDist2SharedHypernym	proposed in [15]	N,V

Although this paper focuses on nouns, noun word forms are often homograph to word forms of other POSs. Since these homographs are also considered, the information on the POS of a node is used as a feature.[3]

The number of lemmas two synsets share is a further feature. Also the information of whether one word sense is a direct hyponym or hypernym of the other (e.g., the synset {human, man} subsumes {man} (male human being)) was considered. An overview of the graph-based features for the noun subset of WN is given in Table 2.

4 Evaluation

4.1 Evaluation of the Model and Feature Set

Each noun word sense sharing its word form with any other word sense in WN is extracted.[4] An instance consists of the features for the two word forms as reported in Tables 1 and 2. Each pair of the kind w1 : w2 or w2 : w1 is only considered once. A subset of 2,511 pairs was manually classified as either sharing a similar/common meaning, class {*yes*}, or as being just arbitrarily homograph, class {*no*}. In the set, 1,237 pairs have been classified as being related, while 1,274 have been classified as being unrelated. Using a simple prediction assigning the most common class to each instance (i.e., *no*), a model has to top a baseline of 50.74% correctly classified instances.[5]

[3] Especially when looking at other POSs than nouns, one can find the tendency of some POSs (e.g., adverbs and adjectives) to be more likely to share meaning with adverbs or adjectives than other POSs.

[4] These include pairs of two noun word forms as well as pairs of noun word forms that are homographic to an adjective, verb, or adverb.

[5] This number might seem high, since polysemy is expected to be more frequent than homonymy. But here word forms belonging to different word senses are considered, not words.

Table 3. Precision difference obtained by removal of the single feature

Feature	Loss	Feature	Loss
word 1 degree	0.52	word 2 pos	0.48
word 1 closeness	5.98	word 2 degree	−0.24
word 1 betweenness	1.19	word 2 closeness	0.16
word 1 eigenvector centrality	0.68	word 2 betweenness	0.24
word 1 page rank	1.23	word 2 eigenvector centrality	0.32
word 1 synset degree	3.46	word 2 page rank	0.36
distance	−1.95	word 2 synset degree	0.52
is-A rel.	0.78	Lin	0.36
Hirst and St. Onge	0.36	Resnik and Yarowsky	0.36
Leacock and Chodorow	0.48	Jiang and Conrath	0.84
Wu and Palmer	0.28	isDerivedFrom	0.52
sharedLemmas	0.92	minDis2SharedHypernym	0.20

Different algorithms have been proposed for classification tasks. Here, the best results are obtained using the random forest model [4].[6] Based on 100 random trees, each constructed while considering 17 random features and 10 seeds, using a 10-fold cross-validation, the model reaches a precision of 0.861 and a recall of 0.877 out of 1. The F-measure is thus 0.87; 86.98% of the instances were correctly classified. The model outperforms the baseline by 36.24 points.

Unfortunately, the random forest algorithm is a black box when it comes to evaluating the impact of a single feature on the overall performance. Unlike decision trees the random forest model selects the features randomly. To evaluate the contribution of a feature, an ablation study was performed. One feature is sequentially deleted and the algorithm evaluated again. The gain or loss in accuracy is shown in Table 3. A negative number indicates a gain.

An ablation study, however, does not evaluate the impact of the feature but rather its contribution to the trained model. The combination of different features has more influence on the model than the information content of the single feature. This is especially true for the random forest model, as [4] shows, and can be explained by the randomly chosen features.

Interestingly enough, removing the distance from the feature set results in a considerable gain of accuracy of nearly 2 points. The distance was thought to be a good indicator of the class $\{yes\}$: Almost all instances with a geodesic path shorter than 6 are of this class. It does not reliably predict the other class though. Even infinite paths are no indicator of class $\{no\}$.

The closeness of the first word form has the biggest impact on the model. The closeness indicates the mean length of the geodesic paths from this node to any other node in the graph. The degree of the first word form has a high

[6] Other classifiers (e.g., support vector machines and Bayesian models) were evaluated as well. The full results cannot be presented here. Still, the findings are very comparable to the ones that will be presented here, but the precision, recall, and F-measures are considerably lower.

impact on the performance of the model. Also the page rank and betweenness can contribute a lot to the overall accuracy. The measures of semantic similarity that were thought to indicate close relations between word senses have less impact. This is likely related to the fact that these measures only work on pairs where both word forms are nouns. Only 1,322 of the total 2,511 available instances are of this kind. The local and global network measures that were proposed in this paper as features of semantic relatedness between word forms in WN contribute the greater part to the accuracy of the model.

Leaving out all semantic similarity measures and the geodesic path shows this even more clearly. Using only the here proposed graph-based features results in an even higher accuracy of 90.12% of correctly classified instances. The model is trained using 16 random features.[7] This relatively small set of easy to use and to calculate graph measures contains enough information to suit the classification task and reach high precision, recall, and F-measure (all three 0.9). This is very well balanced: No class has a significantly higher precision or recall. In the following comparison to CoreLex BTs, the measures of semantic similarity and the geodesic path will not be considered and only the features given in Table 2 will be used.

4.2 Comparison to Using CoreLex BTs as Features

To compare the graph-based features to the CoreLex BTs, every noun word form was annotated the appropriate BT assigned by the CoreLex resource. The assumption is that those BTs show patterns of regular polysemy that the classification algorithm should be able to identify. The actual rules proposed in [5,16,2] were not used. Also, this is not a direct comparison to those results. But it should give insight into the quality of the features used.

The CoreLex approach can only be used to identify regular polysemy. The ratio of correctly classified instances can therefore be expected to drop compared to the graph-based approach. All instances of a noun word form and one of another POS will not show any significant patterns. This again can be expected to result in a drop of accuracy.

Using only the BTs to train a model and evaluating it as before results in 64.99% accuracy.[8] This is 25.13 points less accurate than the method proposed in this paper.

The BTs are not fit to identify cases involving other POSs than nouns. Using only instances containing just nouns and no other POSs, and only the BTs as features results in 66.77% accuracy. Using the graph-based measures instead results in 82.9%. Thus, they are still 16.13 points more accurate. Combining

[7] As the number of available features drops so does the optimal number of features used in the random forest model.

[8] Different models were trained using different algorithms. Still, the random forest model was the most accurate one. The numbers given in the following are always the highest possible rates of accuracy of a random forest model of 100 trees. The number of randomly selected features varies.

all measures further increases the accuracy up to 93.31% on the set containing instances of nouns and other POSs and 87.9% when only noun–noun pairs are used.

5 Outlook

The approach presented in this paper aims at connecting those homograph word forms in WN that share some common meaning and can hence be assumed to belong to a polysemous word. Both regular and irregular polysemy are meant to be covered. This distinguishes the proposed method from other approaches. Earlier efforts were solely focused on regular polysemy and thereby ignored the apparently quite high number of irregular cases found in WN.

The procedure described here is not limited to only nouns but can be straightforwardly adapted to other POSs as well. By taking homographs of different POSs into account and by handling irregular polysemy as well, it outperforms models trained on the BTs proposed in [5] by far. Using the proposed features and the mentioned measures of semantic similarity including the geodesic path results in 86.86% accuracy. Using only the graph-based measures further increases the accuracy up to 90.12%. Combining the network-based features and the BTs results in 93.31% accuracy. Although it was assumed that the decision whether two homograph word forms belonging to different synsets share a common meaning was a question of semantic similarity, the geodesic path, although when < 6 a good indicator for the class *yes*, and other measures of semantic similarity did not improve the performance.

Following [16, p. 268], polysemy and homonymy are "two points on a gradient, where the words in the middle show elements of both". The method presented in this paper allows measuring the degree of homonymy a word exhibits by looking at the sense level and connecting the different senses of a word by relating the corresponding word forms. The word forms of a word like *bank* are not all connected, only those that actually share a common meaning.

The next steps will be to find fitting features for other POSs based on a deep analysis of the networks structure, to manually annotate a test and training set, and to train appropriate models on this data.

References

1. Apresjan, J.U.D.: Regular Polysemy. Linguistics 12, 5–32 (1974)
2. Boleda, G., Padó, S., Utt, J.: Regular polysemy: A distributional model. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics – vol. 1: Proceedings of the Main Conference and the Shared Task, and vol. 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, SemEval 2012, vol. 1, pp. 151–160. Association for Computational Linguistics, Stroudsburg (2012)
3. Bonacich, P.: Power and Centrality: A family of Measures. The American Journal of Sociology 92(5), 1170–1182 (1987)

4. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
5. Buitelaar, P.: Corelex: An ontology of systematic polysemous classes. In: Proceedings of the 1st International Conference on Formal Ontology in Information Systems (FOIS 1998), June 6-8. Frontiers in Artificial Intelligence and Applications, vol. 46, pp. 221–235. IOS Press, Trento (1998)
6. Freeman, L.C.: Centrality in Social Networks Conceptual Clarification. Social Networks 1(1978), 215–239 (1978)
7. Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, pp. 305–332. MIT Press (1998)
8. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR cmp-lg/9709008 (1997)
9. Leacock, C., Miller, G.A., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. Comput. Linguist. 24(1), 147–165 (1998)
10. Lin, D.: An information-theoretic definition of similarity. In: Shavlik, J.W. (ed.) Proceedings of the 15th International Conference on Machine Learning (ICML 1998), Madison, WI, USA, July 24-27, pp. 296–304. Morgan-Kaufman Publishers, San Francisco (1998)
11. Miller, G.A.: WordNet: A lexical database for english. Commun. ACM 38(11), 39–41 (1995)
12. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 161–172 (1998)
13. Pedersen, T., Patwardhan, S., Michelizzi, J.: Wordnet: Similarity – measuring the relatedness of concepts. In: McGuinness, D.L., Ferguson, G. (eds.) AAAI, pp. 1024–1025. AAAI Press / The MIT Press (2004)
14. Resnik, P., Yarowsky, D.: Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. Natural Language Engineering 5(2), 113–133 (1999)
15. Snow, R., Prakash, S., Jurafsky, D., Ng, A.Y.: Learning to merge word senses. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, pp. 1005–1014. ACL (2007)
16. Utt, J., Padó, S.: Ontology-based distinction between polysemy and homonymy. In: Proceedings of the Ninth International Conference on Computational Semantics, IWCS 2011, pp. 265–274. Association for Computational Linguistics, Stroudsburg (2011)
17. Veale, T.: Polysemy and category structure in wordnet: An evidential approach. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004. European Language Resources Association (2004)
18. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, ACL 1994, pp. 133–138. Association for Computational Linguistics, Stroudsburg (1994)