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

Bastian Entrup, Justus-Liebig-Universität Gießen @ PolTal 2014, Warsaw

This poster 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 the features shown in the Table below, a random forest (Breiman, 2001) model is trained and evaluated on a hand-crafted, manually classified list of homograph word senses. The results are compared to a model based on CoreLex basic types (Buitelaar, 1998). The features presented here are applicable to all parts-of-speech. Since CoreLex can only be applied to nouns, the model and evaluation presented here is restricted to nouns, too.

Not only can the features presented in the Table below and similar features be applied to other POS as well, they also outperform the commonly used CoreLex resource, and can be used to identify both regular and irregular polysemous word forms in WordNet.

Interestingly enough, adding measures of semantic similarity commonly used with WordNet, most based on the geodesic path between two nodes, to the feature set has been found to result in a drop of precision and recall. The network measures used here seem to be better indicators of semantic similarity than measures based on the geodesic path.

Table Proposed graph-based feature set.

Abbreviation	Source or description	POS
closeness	the closeness value of the node	all
betweenness	the betweenness value of a node	all
distance	geodesic path between the two nodes	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	all
isA-Rel	is-A relation in WN	N,V
isDerivedFrom	derived-from relation in WN	all
POS	the part of speech of the word form	all
sharedLemmas	number of lemmas shared by the synsets	all
minDist2SharedHypernym	minimum path length to next common hypernym	N,V

(Ir)regular cases of Polysemy of *chicken*

chicken, poulet, volaille (the flesh of a chicken used for food) *(regular)*

chicken, Gallus gallus (a domestic fowl bred for flesh or eggs; believed to have been developed from the red jungle fowl) *(regular)*

chicken, wimp, crybaby (a person who lacks confidence, is irresolute and wishy-washy) *(irregular)*

chicken (a foolhardy competition; a dangerous activity that is continued until one competitor becomes afraid and stops) *(irregular)*

Algorithm

RandomForest (Breiman, 2001)
no. of trees: 100
seed: 10
random features: 17

Dataset

classes: {yes,no}
set-size: 2,511
instances yes: 1,237
instances no: 1,274

Evaluation:

baseline: 50.74%
10-fold cross evaluation

this model:
precision: 0.9
recall: 0.9
f-measure: 0.9

CoreLex:
precision: 0.62
recall: 0.62
f-measure: 0.62

both combined:
precision: 0.93
recall: 0.93
f-measure: 0.93

Reference:

Breiman, L(2001): Random Forests. Machine Learning (45), 5–32

Buitelaar, P. (1998): Corelex: An ontology of systematic polysemous classes. In: Proceedings of the 1st International Conference on Formal Ontology in Information Systems (FOIS'98), June 6–8. Frontiers in Artificial Intelligence and Applications, vol. 46, pp. 221–235. IOS Press, Trento, Italy

Bastian Entrup
bastian.entrup@germanistik.uni-giessen.de
Justus Liebig University Giessen
Applied and Computational Linguistics
Otto-Behaghel-Str. 10 D
35394 Giessen, Germany

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

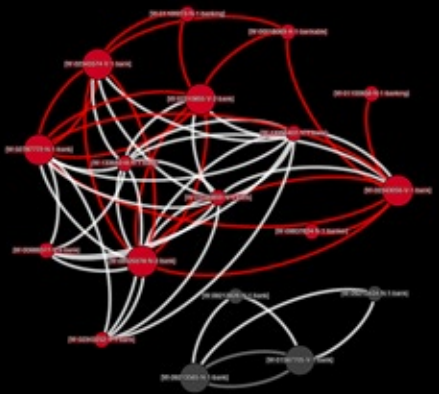

Fig. WordNet before and after: grey and red edges are those found in WordNet. White edges are those added by the model presented on this poster.