A semantic relatedness metric based on free link structure

Sander Wubben
TiCC
Tilburg University
s.wubben@uvt.nl

Antal van den Bosch
TiCC
Tilburg University
antal.vdnbosch@uvt.nl

Abstract

While shortest paths in WordNet are known to correlate well with semantic similarity, an *is-a* hierarchy is less suited for estimating semantic relatedness. We demonstrate this by comparing two free scale networks (ConceptNet and Wikipedia) to WordNet. Using the Finkelstein-353 dataset we show that a shortest path metric run on Wikipedia attains a better correlation than WordNet-based metrics. ConceptNet attains a good correlation as well, but suffers from a low concept coverage.

1 Introduction

In this paper we propose a new estimate metric for semantic relatedness, by finding the shortest path between two concepts in a semantic network. The semantic networks that we exploit are the extracted free link structures of Wikipedia and the ConceptNet 3 database, a commonsense knowledgebase [3]. Various metrics of calculating semantic similarity have been developed for WordNet [1] and applied to Wikipedia's category graph [7]. These measures tend to perform well on semantic similarity (how synonymous two words are), but not very well on semantic relatedness (how related two words are).

2 Free-link pathfinding

The graph extracted from Wikipedia contains over 2 million nodes and 55 million edges, from respectively the number of articles and the number of internal links. ConceptNet contains over 18 thousand usable concepts (nodes)

and over 254 thousand useful assertions (edges). To find paths between concepts in these massive networks, they need to be indexed well and and the algorithms need to be efficient.

For the current study, the English Wikipedia dump dated 12 March 2008 is downloaded[1]. To form the network, only the article names and hyperlinks between the articles are extracted, constituting the link structure. ConceptNet 3 can be freely downloaded [2]. Due to the database being in N3 format, the links can be extracted straightforwardly. The Porter stemmer[3] is used for normalization of the concepts.

Because of the richness and scale of the links in these networks, which both can be argued to have scale free characteristics [6, 8], a breadth-first search is the best option. As we index both incoming and outgoing links, forward and backward chaining can be used simultaneously. Therefore, we choose to utilize a bidirectional breadth-first search algorithm.

3 Experiments

To evaluate the task we compare our results to how humans would do given the same task. Most available datasets focus on semantic similarity rather than semantic relatedness. Examples of these are the Rubenstein and Goodenough and Miller and Charles wordpairs. The Finkelstein-353[4] test collection is a dataset that does contain semantic relatedness scores. It contains 353 wordpairs, among which the 30 wordpairs from the Miller & Charles dataset. The collection contains English word pairs along with human-assigned relatedness judgements.

The breadth-first free-link pathfinding search algorithm is applied to both the Wikipedia and the ConceptNet data, generating semantic relatedness estimates for the Finkelstein pairs. To compare our metric against path-based metrics based on WordNet, the `WordNet::Similarity` package[5] is used [5] on WordNet 3.0 [4].

[1] `http://download.wikimedia.org/enwiki/20080312/enwiki-20080312-pages-articles.xml.bz2`

[2] `http://conceptnet.media.mit.edu/conceptnet\_en\_20080605.n3.bz2`

[3] `http://www.ldc.usb.ve/~vdaniel/porter.pm`

[4] `http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html`

[5] `http://search.cpan.org/dist/WordNet-Similarity/`

WordNet measures	0.18 - 0.45
WikiRelate!	0.19 - 0.48
ESA Wikipedia	0.75
Free-link Wikipedia, directed	0.45
Free-link Wikipedia, undirected	**0.53**
Free-link ConceptNet	0.35
Free-link ConceptNet, non-missing	0.47

Table 1: Spearman's ρ rank order coefficients of different measures with human judgements.

4 Results

As displayed in Table 1 (displaying also the results from ESA and WikiRelate! taken from [2]), the basic directed free-link pahfinding measure applied on Wikipedia shows a Spearman's $\rho = 0.45$, while the undirected measure shows a $\rho = 0.53$ with human judgements on the full Finkelstein dataset. Only the word *defeating* was not found in Wikipedia.

In ConceptNet, 58 of the Finkelstein pairs were not found, significantly impairing its score: $\rho = 0.35$. When we look only at those word pairs for which a path was found, the measure applied to ConceptNet performs sightly better than the directed Wikipedia measure, but worse than the more comparable undirected Wikipedia measure: $\rho = 0.47$.

5 Discussion

With a correlation of 0.53 for undirected search versus 0.43 for directed search with the human judging scores of the Finkelstein-353 dataset, the former appears the better suited way of exploiting the Wikipedia network for a semantic relatedness metric. It also outperforms undirected search in ConceptNet, which only scores a correlation of 0.35. This lower score can be mainly attributed to the lack of coverage of ConceptNet. But even when only non-missing wordpairs are considered, ConceptNet performs worse than Wikipedia with a correlation of 0.47 with human judgements.

The method introduced here outperforms any other existing pathfinding method for calculating semantic relatedness. It also outperforms methods that makes use of WordNet's extended gloss vectors. This cannot be explained by coverage: both in WordNet 3.0 and in the Wikipedia dump that was used only one wordpair was not found. Arguably, the results show that

free link structure in conceptual networks is better suited for finding semantic relatedness than hierarchical structures organized along taxonomic relations as WordNet.

While ESA's performance ($\rho = 0.75$) is even higher, the comparison is off. ESA is a fully integrated machine learning architecture that makes extensive use of Wikipedia's free text. Our free link measure is a simple and portable method that uses only the link structure of any given free associative conceptual network. We aim to improve the method and use it in various applications, such as automated ontology building and memory based paraphrasing.

References

[1] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[2] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[3] H. Liu and P. Singh. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

[4] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[5] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts, 2004.

[6] S. Spek. Wikipedia: organisation from a bottom-up approach. In *WikiSym '06: Proceedings of the 2006 international symposium on Wikis*, New York, NY, USA, Nov 2006. ACM.

[7] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *21. AAAI / 18. IAAI 2006*. AAAI Press, july 2006.

[8] S. Wubben. Using free link structure to calculate semantic relatedness. Master's thesis, Tilburg University, the Netherlands, 2008.