Semantic Relatedness from Automatically Generated Semantic Networks

Pia-Ramona Wojtinnek and Stephen Pulman
Oxford University Computing Laboratory
{pia-ramona.wojtinnek,stephen.pulman}@comlab.ox.ac.uk

Abstract

We introduce a novel approach to measuring semantic relatedness of terms based on an automatically generated, large-scale semantic network. We present promising first results that indicate potential competitiveness with approaches based on manually created resources.

1 Introduction

The quantification of semantic similarity and relatedness of terms is an important problem of lexical semantics. Its applications include word sense disambiguation, text summarization and information retrieval (Budanitsky and Hirst, 2006). Most approaches to measuring semantic relatedness fall into one of two categories. They either look at distributional properties based on corpora (Finkelstein et al., 2002; Agirre et al., 2009) or make use of pre-existing knowledge resources such as WordNet or Roget's Thesaurus (Hughes and Ramage, 2007; Jarmasz, 2003). The latter approaches achieve good results, but they are inherently restricted in coverage and domain adaptation due to their reliance on costly manual acquisition of the resource. In addition, those methods that are based on hierarchical, taxonomically structured resources are generally better suited for measuring semantic similarity than relatedness (Budanitsky and Hirst, 2006). In this paper, we introduce a novel technique that measures semantic relatedness based on an automatically generated semantic network. Terms are compared by the similarity of their contexts in the semantic network. We present our promising initial results of this work in progress, which indicate the potential to compete with resource-based approaches while performing well on both, semantic similarity and relatedness.

2 Similarity and Relatedness from semantic networks

In our approach to measuring semantic relatedness, we first automatically build a large semantic network from text and then measure the similarity of two terms by the similarity of the local networks around their corresponding nodes. The semantic network serves as a structured representation of the occurring concepts, relations and attributes in the text. It is built by translating every sentence in the text into a network fragment based on semantic analysis and then merging these networks into a large network by mapping all occurrences of the same term into one node. Figure 1(a) contains a sample text snippet and the network derived from it. In this way, concepts are connected across sentences and documents, resulting in a high-level view of the information contained.

Our underlying assumption for measuring semantic relatedness is that semantically related nodes are connected to a similar set of nodes. In other words, we consider the context of a node in the network as a representation of its meaning. In contrast to standard approaches which look only at a type of context directly found in the text, e.g. words that occur within a certain window from the target word, our network-based context takes into account indirect connections between concepts. For example, in the text underlying the network in Fig. 2, *dissertation* and *module* rarely co-occurred in a sentence, but the network shows a strong connection over *student* as well as over *credit* and *work*.

2.1 The Network Structure

We build the network incrementally by parsing every sentence, translating it into a small network fragment and then mapping that fragment onto the main network generated from all previous sentences. Our translation of sentences from text to network is based on the one used in the ASKNet system (Harrington and Clark, 2007). It makes use of two NLP tools, the Clark and Curran parser (Clark and Curran, 2004) and the semantic analysis tool Boxer (Bos et al., 2004), both of which are part of the C&C Toolkit[1]. The parser is based on Combinatory Categorial Grammar (CCG) and has been trained on 40,000 manually annotated sentences of the WSJ. It is both robust and efficient. Boxer is designed to convert the CCG parsed text into a logical representation based on Discourse Representation Theory (DRT). This intermediate logical form representation presents an abstraction from syntactic details to semantic core information. For example, the syntactical forms *progress of student* and *student's progress* have the same Boxer representation as well as *the student who attends the lecture* and *the student attending the lecture*. In addition, Boxer provides some elementary co-reference resolution.

The translation from the Boxer output into a network is straightforward and an example is given in Figure 1(b). The network structure distinguishes between object nodes (rectangular), relational nodes (diamonds) and attributes (rounded rectangles) and different types of links such as subject or object links.

Students *select* **modules** *from* the published **list** and *write* a **dissertation**. **Modules** usually *provide* 15 **credits** each, but 30 **credits** are *awarded for* the **dissertation**. The **student** must *discuss* the **topic** *of* the final **dissertation** with their appointed **tutor**.

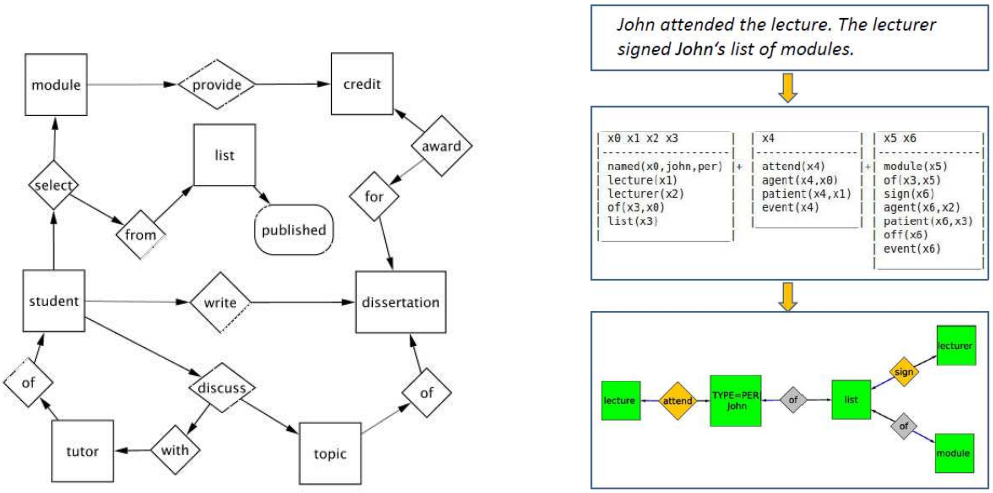

Figure 1: (a) Sample text snippet and according network representation. (b) Example of translation from text to network over Boxer semantic analysis

The large unified network is then built by merging every occurrence of a concept (e.g. object node) into one node, thus accumulating the information on this concept. In the second example (Figure **??**), the *lecture* node would be merged with occurrences of *lecture* in other sentences. Figure 2 gives a subset of a network generated from a few paragraphs taken from Oxford Student Handbooks. Multiple occurrences of the same relation between two object nodes are drawn as overlapping.

2.2 The Vector Space Model

We measure the semantic relatedness of two concepts by measuring the similarity of the surroundings of their corresponding nodes in the network. Semantically related terms are then expected to be connected to a similar set of nodes. We retrieve the network context of a specific node and determine the level

[1] http://svn.ask.it.usyd.edu.au/trac/candc

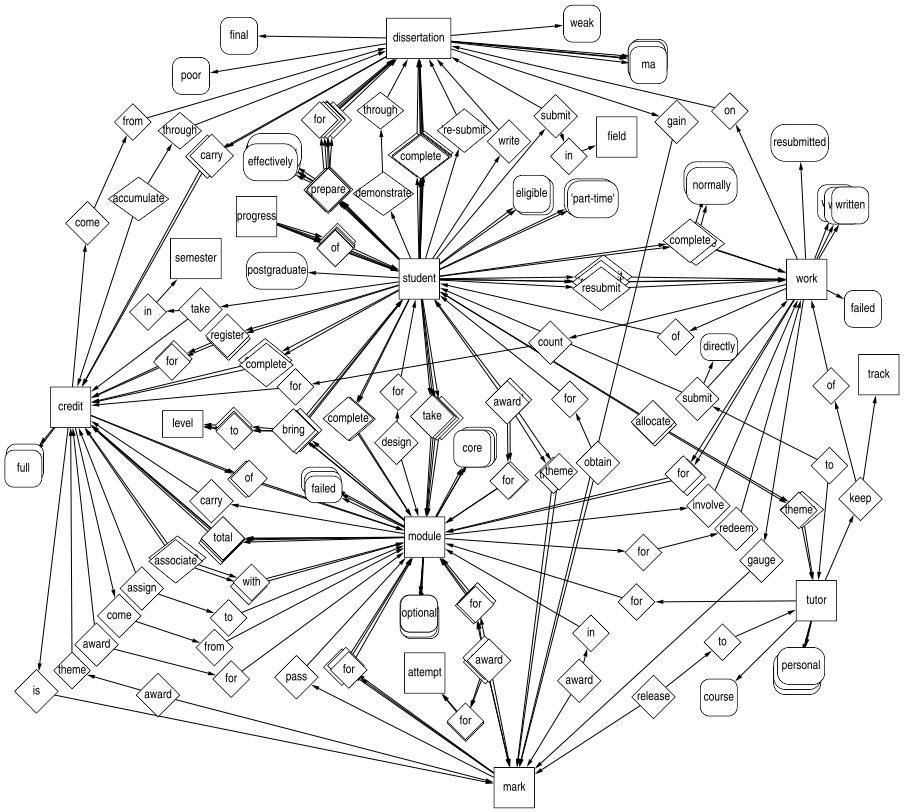

Figure 2: Subgraph displaying selected concepts and relations from sample network.

of significance of each node in the context using spreading activation[2]. The target node is given an initial activation of $a_x = 10{*}\textit{numberOfLinks}(x)$ and is fired so that the activation spreads over its out- and ingoing links to the surrounding nodes. They in turn fire if their received activation level exceeds a certain threshold. The activation attenuates by a constant factor in every step and a stable state is reached when no node in the network can fire anymore. In this way, the context nodes receive different levels of activation reflecting their significance.

We derive a vector representation $\vec{v}(x)$ of the network context of x including only object nodes and their activation levels. The entries are

$$v_i(x) = act_{x,a_x}(n_i) \qquad\qquad n_i \in \{n \in nodes \mid type(n) = object\_node\}$$

The semantic relatedness of two target words is then measured by the cosine similarity of their context vectors.

$$sim\_rel(x, y) = cos(\vec{v}(x), \vec{v}(y)) = \frac{\vec{v}(x) \cdot \vec{v}(y)}{\|\vec{v}(x)\| \, \|\vec{v}(y)\|}$$

As spreading activation takes several factors into account, such as number of paths, length of paths, level of density and number of connections, this method leverages the full interconnected structure of the network.

3 Evaluation

We evaluate our approach on the WordSimilarity-353 (Finkelstein et al., 2002) test collection, which is a commonly used gold standard for the semantic relatedness task. It provides average human judgments scores of the degree of relatedness for 353 word pairs. The collection contains classically similar word

[2]The spreading activation algorithm is based on Harrington (2010)

Approach		Spearman
(Strube and Ponzetto, 2006)	Wikipedia	0.19-0.48
(Jarmasz, 2003)	Roget's	0.55
(Hughes and Ramage, 2007)	WordNet	0.55
(Agirre et al., 2009)	WordNet	0.56
(Finkelstein et al., 2002)	Web corpus, LSA	0.56
(Harrington, 2010)	Sem. Network	0.62
(Agirre et al., 2009)	WordNet+gloss	0.66
(Agirre et al., 2009)	Web corpus	0.66
(Gabrilovich and Markovitch, 2007)	Wikipedia	0.75
Network (all pairs)		0.38
Network (>100 freq: 293 pairs)		0.46
Network (>300 freq: 227 pairs)		**0.50**

	Similarity	Relatedness
all pairs	0.19 (100 pairs)	0.36 (250 pairs)
>300 freq	**0.50** (60 pairs)	**0.52** (171 pairs)

Table 1: (a) Spearman ranking correlation coefficient results for our approach and comparison with previous approaches. (b) Separate results for similarity and relatedness subset.

pairs such as *street - avenue* and topically related pairs such as *hotel - reservation*. However, no distinction was made while judging and the instruction was to rate the general degree of *semantic relatedness*.

As a corpus we chose the British National Corpus (BNC)[3]. It is one of the largest standardized English corpora and contains approximately 5.9 million sentences. Choosing this text collection enables us to build a general purpose network that is not specifically created for the considered work pairs and ensures a realistic overall connectedness of the network as well as a broad coverage. In this paper we created a network from 2 million sentences of the BNC. It contains 27.5 million nodes out of which 635.000 are object nodes and the rest are relation and attribute nodes. The building time including parsing was approximately 4 days.

Following the common practice in related work, we compared our scores to the human judgements using the Spearman rank-order correlation coefficient. The results can be found in Table 1(a) with a comparison to previous results on the WordSimilarity-353 collection.

Our first result over all word pairs is relatively low compared to the currently best performing systems. However, we noticed that many poorly rated word pairs contained at least one word with low frequency. Excluding these considerably improved the result to 0.50. On this reduced set of word pairs our scores are in the region of approaches which make use of the Wikipedia category network, the WordNet taxonomic relations or Roget's thesaurus. This is a promising result as it indicates that our approach based on automatically generated networks has the potential of competing with those using manually created resources if we increase the corpus size.

While our results are not competitive with the best corpus based methods, we can note that our current corpus is an order of magnitude smaller - 2 million sentences versus 1 million full Wikipedia articles (Gabrilovich and Markovitch, 2007) or 215MB versus 1.6 Terabyte (Agirre et al., 2009). The extent to which corpus size influences our results is subject to further research.

We also evaluated our scores separately on the semantically similar versus the semantically related subsets of WordSim-353 following Agirre et al. (2009) (Table 1(b)). Taking the same low-frequency cut as above, we can see that our approach performs equally well on both sets. This is remarkable as different methods tend to be more appropriate to calculate either one or the other (Agirre et al., 2009). In particular, WordNet based measures are well known to be better suited to measure similarity than relatedness due to its hierarchical, taxonomic structure (Budanitsky and Hirst, 2006). The fact that our system achieves equal results on the subset indicates that it matches human judgement of semantic relatedness beyond specific types of relations. This could be due to the associative structure of the network.

4 Related Work

Our approach is closely related to Harrington (2010) as our networks are built in a similar fashion and we also use spreading activation to measure semantic relatedness. In their approach, semantic relatedness

[3]http://www.natcorp.ox.ac.uk/

of two terms a and b is measured by the activation b receives when a is fired. The core difference of this measurement to ours is that it is path-based while ours is context based. In addition, the corpus used was retrieved specifically for the word pairs in question while ours is a general-purpose corpus.

In addition, our approach is related to work that uses personalized PageRank or Random Walks on WordNet (Agirre et al., 2009; Hughes and Ramage, 2007). Similar the spreading activation method presented here, personalized PageRank and Random Walks are used to provide a relevance distribution of nodes surrounding the target word to its meaning. In contrast to the approaches based on resources, our network is automatically built and therefore does not rely on costly, manual creation. In addition, compared to WordNet based measures, our method is potentially not biased towards relatedness due to similarity.

5 Conclusion and Outlook

We presented a novel approach to measuring semantic relatedness which first builds a large-scale semantic network and then determines the relatedness of nodes by the similarity of their surrounding local network. Our preliminary results of this ongoing work are promising and are in the region of several WordNet and Wikipedia link structure approaches. As future work, there are several ways of improvement we are going to investigate. Firstly, the results in Section 3 show the crucial influence of corpus size and occurrence frequency on the performance of our system. We will be experimenting with larger general networks (e.g. the whole BNC) as well as integration of retrieved documents for the low frequency terms. Secondly, the parameters and specific settings for the spreading activation algorithm need to be tuned. For example, the amount of initial activation of the target node determines the size of the context considered. Thirdly, we will investigate different vector representation variants. In particular, we can achieve a more fine-grained representation by also considering relation nodes in addition to object nodes. We believe that with these improvements our automatic semantic network approach will be able to compete with techniques based on manually created resources.

References

Agirre, E., E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *NAACL '09*.

Bos, J., S. Clark, M. Steedman, J. R. Curran, and J. Hockenmaier (2004). Wide-coverage semantic representations from a ccg parser. In *COLING'04*.

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

Clark, S. and J. R. Curran (2004). Parsing the wsj using ccg and log-linear models. In *ACL'04*.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin (2002). Placing search in context: the concept revisited. *ACM Trans. Inf. Syst. 20*(1), 116–131.

Gabrilovich, E. and S. Markovitch (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI'07*.

Harrington, B. (2010). A semantic network approach to measuring semantic relatedness. In *COLING'10*.

Harrington, B. and S. Clark (2007). Asknet: automated semantic knowledge network. In *AAAI'07*.

Hughes, T. and D. Ramage (2007). Lexical semantic relatedness with random graph walks. In *EMNLP-CoNLL'07*.

Jarmasz, M. (2003). Roget's thesaurus as a lexical resource for natural language processsing. Master's thesis, University of Ottawa.

Strube, M. and S. P. Ponzetto (2006). Wikirelate! computing semantic relatedness using wikipedia. In *AAAI'06*.