SEBSem: simple and efficient biomedical semantic relatedness measure

Maciej Rybinski✉, **José Francisco Aldana-Montes**

Departmento de Lenguajes y Ciencias de la Computacion, Universidad de Málaga, Spain

Received 31 July 2013; **Accepted** 6 September 2013; **Published** 14 October 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

Calculating semantic relatedness between terms is crucial in numerous knowledge and information processing tasks highly relevant to the biomedical domain. Examples include semantic search and automated processing of scientific texts. Most available methods rely heavily on highly specialised resources, which substantially limits their reusability in various applications within the domain. In this work we present a simple semantic relatedness measure that relies only on very general resources and its design features allow minimising the costs of online computations. The relatedness is computed through comparing automatically extracted key-phrases relevant to respective input terms. This simple strategy provides a method that gives promising early test results, comparable to those of human annotators and state-of-the-art methods, on a well established benchmark.

Motivation and Objectives

Calculating semantic relatedness between terms is vital in numerous knowledge and information processing tasks of much relevance to the biomedical domain, such as named entity disambiguation (Hoffart *et al.*, 2012), ontology population (Shen *et al.*, 2012), word sense disambiguation (McInnes *et al.*, 2011). Being able to relate entities of interest is crucial in processes of semantic search, information extraction from texts and in building similarity databases. Many successful methods use specialised knowledge bases and lexicon-style resources (Lin, 1998), preparation of which is very tiresome and time consuming. Moreover, in many domains it is not possible to create resources that capture all the possible relationships between the entities of interest. Typically, it is much easier to assemble a fairly large repository of possibly relevant documents that span the domain with their implicit knowledge.

There are also corpus based approaches, whose downsides are often related with computational intensity (Pedersen *et al.*, 2007) and heavy dependence on a specific corpus and its specific features. This paper presents a simple measure designed to provide solutions to those problems, while still being able to produce results comparable with state-of-the art methods.

The design goal was to design a measure that could be used successfully in less-than -ideal availability of knowledge-rich resources. The method takes advantage of a fairly general document corpus of medium size (several orders of magnitude less than Web scale) with very limited use of background knowledge without depending on specific structural features of the knowledge base. Following those design goals should increase robustness and applicability of the new measure, which in terms of quality provides results comparable to those of a human annotator.

Methods

The measure relies on the idea of computing relatedness based on comparing key-phrases related to respective terms. The outline of the method is presented in Figure 1, along with key components used for the similarity computation.

Most relevant documents for the input terms are chosen from the public subset of PubMed articles[1]. Linked Data (Bizer *et al.*, 2009) flavored version of Wikipedia, DBPedia[2], is used as a complementary knowledge base (KB) in the process of query expansion. These are very general resources, selection of which is aimed at obtaining a robust and flexible tool for resolving the relatedness calculation within the whole biomedical domain.

Relatedness of two terms is defined as an overlap measure between key-phrase sets of those terms, as shown in Formula 1. Key-phrases are extracted from K most relevant documents, where document relevance for a term is defined as cosine distance between the document vector and the vector of an expanded query formed around the term. Vectors are defined for a TF-IDF weighted Vector Space Model. For each document N most frequent key-phrases are extracted with a one-pass sliding window T-GSP algorithm

1 http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
2 http://dbpedia.org/sparql

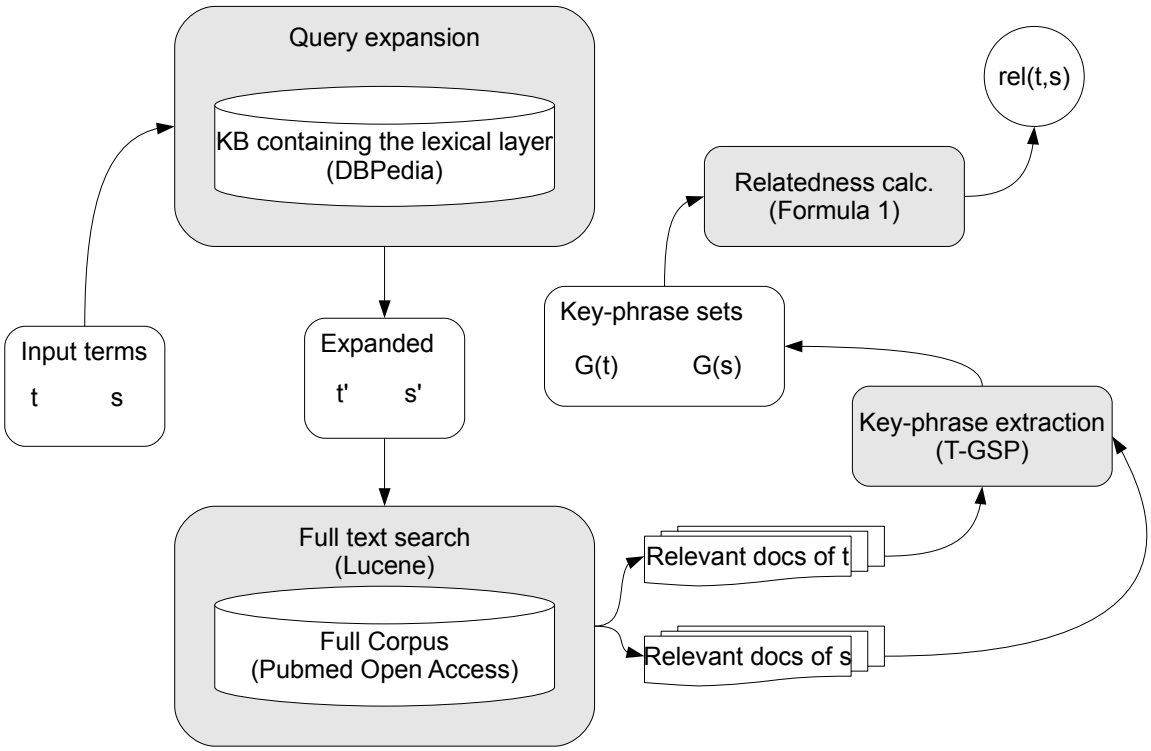

Figure 1. General vision of the system for computing semantic relatedness.

(Protaziuk *et al.*, 2007) that uses several common grammatical patterns to take advantage from shallow parsing information. Query extraction is an optional step that uses the general KB in order to provide wider query context.

Formula 1:

$$\rho(s,t,K,N) = \frac{|G^{sKN} \cap G^{tKN}|}{max(|G^{sKN}|, |G^{tKN}|)}$$

where ρ(s,t,K,N) is semantic relatedness of string s to string t under given parameters K and N and G^{xKN} denotes a set of key-phrases related to string x, with K being the number of documents and N being the max number of most common phrases extracted from a single document.

Query expansion for input terms is executed on the fly and aggregates the results of a DBPedia query that retrieves disambiguation/redirect labels and other synonyms.

For better clarity, the flow of the method is presented in its 'on-demand' implementation, where all computations are executed on-demand for a pair of arguments and additional parameters (*K* and *N*). Nonetheless, it is worth taking notice that the steps of identifying relevant documents for a term and T-GSP analysis of article contents are independent. This means that the key-phrase extraction can be done efficiently off-line for the entire corpus (given the *K* and *N* parameters) in order to speed up the on-line portion of the process. In this case the actual relatedness computations are very simple and are limited to a reduced vector space, which makes the method suitable for large-scale processing, very much present in the biomedical domain.

Results and Discussion

The presented method has been tested on a small benchmark presented in (Pedersen *et al.*, 2007). The benchmark consists of 30 term pairs annotated by a group of physicians and the same pairs annotated by a group of medical coders. Our method was tested with *K* and *N* pa-

rameters ranging from 1 to 10 (integers only), and in it achieved best results for the physicians case with K=6 and N=7, giving correlation rp=0.68 with respect to the average answers, and with K=6 and N=10 it achieved its best result for coders set, achieving the correlation rc=0.77 with respect to the average answers. Those scores were achieved for documents matched to terms through expanded queries, a method that provides 95% coverage for the terms included in the benchmark.

The method without query expansion achieves best case scores of rp=0.39 and rc=0.46 respectively, while also providing 95% coverage for single terms. Best scores without expansion are achieved for a yet another parameter pair (K=5, N=2), which shows the need for further testing in order to fine-tune the algorithm. The version with query expansion seemed more robust in terms of parameter dependence, as it would generally score reasonably high for higher values of K and N parameters.

Additionally, shifting the key-phrase extraction to offline processing will allow involving whole-corpus statistics in the relatedness computation process without excessive additional cost. Doing so should also improve the algorithm in terms of its parameter sensitivity, which shows in the preliminary tests, especially for the evaluations without query expansion.

In general, the presented method in its optimal settings (with query expansion) achieves promising results, which are comparable to those of a human annotator (correlation higher than inter-annotator agreement). Additionally, according to a comparative of various measures presented in (Zhang *et al.*, 2011), our method ranks well against other state-of-the-art related measures for biomedicine, while using very general knowledge sources. During the evaluation it has also been established that the presented measure provides much better (correlation-wise) results than a commonly used Wikipedia-based measure, which relies on comparing class labels of objects most related to the input terms. Furthermore, the automatically extracted key-phrases seem to be more meaningful in terms of semantic relatedness than the actual keywords associated with the documents. Using PubMed keywords as input for the relatedness computation would tend to cause a substantial decrease in result quality.

As a part of the future work the measure should be tested against a much larger (by at least two orders of magnitude) benchmark, possibly with various subdomains. Such an evaluation would certainly help in terms of validating the robustness of the approach and showing the real importance of tuning the K and N parameters. It would also certainly be beneficial for further development of the measure formula itself, as a large benchmark will provide more significant answers.

Acknowledgements

Part of this work was financed under project grants TIN2011-25840 (Spanish Ministry of Education and Science) and P11-TIC-7529 (Innovation, Science and Enterprise Ministry of the regional government of the Junta de Andalucía).

References

Bizer C, Heath T, Berners-Lee T (2009) Linked data-the story so far. Int. *J. Semant. Web Inf. Syst.* **5**(3), 1-22. doi:10.4018/jswis.2009081901

Hoffart J, Seufert S, *et al.* (2012) KORE: keyphrase overlap relatedness for entity disambiguation. In: *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*. ACM, New York, pp. 545-554. doi:10.1145/2396761.2396832

Lin D (1998) An information-theoretic definition of similarity. In: *ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 296-304.

McInnes BT, Pedersen T, *et al.* (2011) Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In: *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, Vol. 2011, pp. 895–904.

Pedersen T, Pakhomov SVS, *et al.* (2007) Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics* **40**(3), 288-299. doi:10.1016/j.jbi.2006.06.004

Protaziuk G, Kryszkiewicz M, *et al.* (2007) Discovering compound and proper nouns. In: *Rough Sets and Intelligent Systems Paradigms*. Springer, Berlin, pp. 505-515.

Shen W, Wang J, *et al.* (2012) A graph-based approach for ontology population with named entities. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. ACM, pp. 345-354. doi:10.1145/2396761.2396807

Zhang Z, Gentile AL, Ciravegna F (2011) Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 991-1002.