International Journal of Applied Information Systems (IJAIS) ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 4 - No. 5, October 2012 - www.ijais.org

Semantic Similarity Measure for Pairs of Short Biological Texts

Olivia Sanchez Graillet
Research Institute of Applied Mathematics and Systems (IIMAS)
National Autonomous University of Mexico (UNAM)
Ciudad Universitaria, Coyoacan, 04510, Mexico City, Mexico

ABSTRACT

Finding the semantic similarity between biological texts, specially short texts, such as article abstracts and experiment descriptions of microarrays, may throw important information for experts in that field. To date, these methods have not been widely explored. In this paper, a comparison of different measures to calculate the semantic similarity of pairs of short biological texts is presented. An existing method for semantic similarity between general texts was adapted to be used in the biological context by employing the UMLS ontology. An evaluation of the methods was carried out and it was found that the adapted method works well for short biological texts.

General Terms:

Semantic Similarity in Biological Texts, Bio-text Mining

Keywords:

Semantic Similarity, Ontology, Knowledge Discovery, Text Processing

1. INTRODUCTION

Existing methods to compute semantic similarity or semantic relatedness between biomedical concepts have been developed [1, 2], which are based on general similarity methods [3].
Methods to calculate semantic similarity between concepts in a general domain [4, 5] use statistics obtained from a corpus or transverse an ontology graph to find the smallest path between concepts [3, 6].
In this paper, an automatic method to calculate the semantic similarity between short biological texts is presented. This method is based on a general text metric adapted to the biological context. The calculation of the semantic similarity value between concepts uses the UMLS ontology. The objective of this method is to help biologists to automatically find important evidence contained in short texts.
The remainder of this paper is structured as follows. In Section 2, the UMLS resources are reviewed. Section 3 describes the method to obtain the semantic similarity values between concepts. Section 4 describes semantic similarity measures for pairs of texts. Section 5 presents the adapted method. Section 6 presents the evaluation of the method and a discussion of the results. Section 7 outlines the conclusions of the work here presented.

2. UMLS RESOURCES

The Unified Medical Language System (UMLS) consists of three knowledge sources that can work together or separately. These components are briefly described in the following paragraphs. For more information refer to the UMLS tutorial webpage [1]. For the present study, version 2012AA of the UMLS resources is used.

2.1 Metathesaurus

The Metathesaurus is a huge, multi-purpose, multi-lingual vocabulary database. It contains information on more than 2.7 million concepts and 10.8 million unique concept names from over 160 source vocabularies from over 100 vocabularies, terminologies and code sets in 17 languages.

2.1.1 SNOMED-CT. The Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [2] is an ontological resource included in the UMLS metathesaurus. SNOMED-CT is an organised collection of medical terms, synonyms and definitions covering diseases, findings, and procedures. It is used for indexing electronic medical records, information retrieval, data aggregation and exchange, etc. SNOMED-CT includes the semantic relationship *is_a* among other relationships. In this study, the version of SNOMED-CT included in UMLS-2012AA is used.

2.2 Semantic Network

The Semantic Network is a set of broad subject categories (Semantic Types) and a set of relationships between them (Semantic Relationships). Semantic Types are assigned to Metathesaurus concepts. The Semantic Network contains 133 semantic types and 54 relationships. Its primary relation is *is_a*, which establishes the hierarchy of types within the network and is used for deciding on the most specific semantic type available for assignment to a Metathesaurus concept [3].
The set of non-hierarchical relations between the semantic types are grouped into five categories and are themselves relations: *physically related to*, *spatially related to*, *temporally related to*, *functionally related to*, and *conceptually related to*.

2.3 The SPECIALIST Lexicon and Lexical Tools

The SPECIALIST Lexicon contains lexical information for over 300 thousand common English words and biomedical vocabularies. The lexical entry for each word or term includes syntax information, morphological information (e.g. inflection, derivation, and composition information), and spelling information.
The SPECIALIST lexical tools are programs for assisting natural language processing. These programs are: a lexical variant generator (LVG), a normalized string generator (Norm) and a word Index generator (Wordind). The lexical tools also include the programs SemRep and MetaMap.

[1] www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_003.htm
[2] www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
[3] www.ncbi.nlm.nih.gov/books/NBK9675/

2.3.1 MetaMap. MetaMap v.10 [4] is part of the lexical tools provided by UMLS. It maps arbitrary terms to concepts in the UMLS Metathesaurus from free texts.
The MetaMap options -yYIc are used by the method to obtain the respective UMLS concepts from biological texts. The meanings of these options are:

—y (word sense disambiguation): MetaMap attempts to disambiguate among concepts scoring equally well in matching input text.

—Y (prefer multiple concepts): MetaMap scores mappings with more concepts higher than those with fewer concepts. For example, the input text "lung cancer" will score the mapping to the two concepts 'Lung' and 'Cancer' higher than the mapping to the single concept 'Lung Cancer'.

—I (show cuis): shows the UMLS CUI for each concept displayed.

—c (hide candidates): disables the displaying of the list of Metathesaurus candidates

3. MEASURES OF CONCEPT/WORD SEMANTIC SIMILARITY

There are different methods to measure the semantic similarity between two words or concepts. Some methods exclusively use the structure of a given taxonomy; others are based on the frequencies of the words found in a large corpus; and others combine both corpus and the taxonomy structure. In the next paragraphs, some of these measures are briefly described.

3.1 Corpus-based measures

Two popular corpus-based methods are the latent semantic analysis (LSA) [7] and the PMI-IR method [8]. Both methods are based on word co-occurrence. The PMI-IR method uses word co-occurrence calculated with the counts collected from a corpus where the corpus can be the web. The co-occurrence of words in LSA is obtained by a singular value decomposition (SVD) on a term-by-document matrix which represents the corpus. These methods have shown to be effective but highly computationally expensive.

3.2 Taxonomy-based measures

These measures rely on the distances between concepts linked by a type of relationship between them. The *is_a* relationship is the most commonly used. The basic method consists of getting the shortest path between two concepts ($path$). The simplest way to calculate the similarity between two concepts c_1 and c_2 is defined by formula (1):

$$Sim_{path}(c_1, c_2) = \frac{1}{length} \tag{1}$$

Where $length$ is the shortest path between two concepts using node-counting.
Applications of this principle were implemented by Rada et al. [3] in the MeSH ontology and by Caviades and Cimino [9] in the UMLS ontology. Variations of the $path$ measure are those of Leacock and Chodorow [10] (lch), and that of Wu and Palmer [6] (wup).
The lch equation is shown in (2):

$$Sim_{lch}(c_1, c_2) = -log\frac{length}{2 * D} \tag{2}$$

Where D is the maximum depth of the taxonomy.

The wup similarity score is calculated with equation (3).

$$Sim_{wup}(c_1, c_2) = \frac{2 * depth_{LCS}}{depth_{c1} + depth_{c2}} \tag{3}$$

where $depth_{LCS}$ is the depth of the least common subsumer (LCS).
The advantages of the taxonomy-based measures are that they are simple and not computationally expensive.

3.3 Taxonomy and corpus-based measures

These measures use the information obtained from the taxonomy combined with the information content (IC), which is the amount of information provided by the probability of a word/concept to appear in a corpus $p(c)$. IC is calculated with equation (4).

$$IC(c) = -log\, p(c) \tag{4}$$

The similarity measure introduced by Resnik [5] is calculated with (5):

$$Sim_{res}(c_1, c_2) = IC(LCS) \tag{5}$$

Where LCS is the least common subsumer of the two concepts.
Lin [11] added a normalisation factor to the Resnik measure, calculating the semantic similarity in the following way:

$$Sim_{lin}(c_1, c_2) = \frac{2 * IC(LCS)}{IC(c_1) + IC(c_2)} \tag{6}$$

These measures use the implicit information contained in a corpus. Therefore, they depend on the coverage and size of the corpus used.

4. METHODS FOR TEXT SEMANTIC SIMILARITY

Mihalcea et al. [12] and Banerjee and Padersen [13] provide two different methods to calculate the semantic similarity between two texts. The Mihalcea's method consists in the use of different semantic similarity measures between the words found in the text being compared in an equation that weights and normalises the final similarity measure. The equation to get the similarity between texts T_1 and T_2 is:

$$Sim(T_1, T_2) = \tfrac{1}{2}\Bigg(\frac{\sum_{w\in(T_1)}(maxSim(w,T_2)*idf(w))}{\sum_{w\in(T_1)} idf(w))} + \frac{\sum_{w\in(T_2)}(maxSim(w,T_1)*idf(w))}{\sum_{w\in(T_2)} idf(w))}\Bigg) \tag{7}$$

Where idf is the inverse document frequency [14] of a word w, which defines its specificity. The idf measure is calculated as shown in (8).

$$idf(w, D) = log\frac{|D|}{|d \in D : w \in d|} \tag{8}$$

Where D is the number of documents in the corpus and the denominator is the number of documents where the word w appears.
The Milhecea's method has given good results when comparing general texts.
The method of Banerjee and Pedersen [13] (bap) looks for overlaps of the words of the texts being compared. The metric can be obtained by using a Perl module[5] that measures the

[4] metamap.nlm.nih.gov/

[5] www.d.umn.edu/ tpederse/text-similarity.html

similarity of two given files or strings by comparing the number of overlapping (shared) words, scaled by the lengths of the files.

5. METHOD TO OBTAIN THE SEMANTIC SIMILARITY VALUE BETWEEN TWO BIOLOGICAL SHORT TEXTS

The presented method for biological texts is based on the Mihalcea and Strapparava's method. However, while Mihalcea uses words, the adopted method uses concepts contained in the UMLS ontology. Metamap is used to get the concepts from text. Once the concepts are identified, equation (9) is used to determine the similarity score.

$$sim(T_1,T_2) = \frac{1}{2}\left(\frac{\sum_{c\in(T_1)}(maxSim(c,T_2)*idf(c))}{\sum_{c\in(T_1)} idf(c))} + \frac{\sum_{c\in(T_2)}(maxSim(c,T_1)*idf(c))}{\sum_{c\in(T_2)} idf(c))}\right) \quad (9)$$

The idf can be calculated from a corpus or from the two files being compared. The semantic similarity value between two concepts is obtained by using the taxonomy-based measures presented in Subsection 3.2 and implemented in the Perl module created by the Ted Pedersen's team [6]. The semantic similarity values lie in the range from zero to one.

5.0.1 Example. To illustrate the way the method works, a query taken from the OHSUMED corpus is taken, which is also used in the evaluation, as well as a relevant answer and an irrelevant answer for it which are contained in two different PubMed abstracts.

EXAMPLE 1. *(a) Query: "Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy"*

(b) Relevant document: "Changes in lipids and lipoproteins with long-term estrogen deficiency and hormone replacement therapy"

(c) Non-relevant document: "Nausea and vasopressin [editorial]"

After running MetaMap, the concepts obtained for text (a) are shown in Table 1, for text (b) in Table 2 and for text (c) in Table 3.

Table 1. Query concepts

C0001688:adverse effects (aspects of adverse effects)
C0523744:Lipids NOS (Lipids measurement)
C0033308:Progesterone
C1947971:Give (Give - dosing instruction imperative)
C0014939:Estrogen (Estrogens)
C0559956:Replacement
C1363945:Therapy (Therapy Object (animal model)

The query text (a) is compared with the two texts (b) and (c). For simplicity, only the comparison between (a) and (c) is presented. The maximum similarity values obtained with the lch method concepts in (a) with respect to concepts in (c) are presented in Table 4 and for concepts in (c) with respect to concepts in (a) in Table 5. The value -1 is given in cases when there is not any path between two concepts in the UMLS taxonomy.

After applying the text similarity measure lch, the values obtained are: $semSim_{a,b} = 0.63$ and $semSim_{a,c} = 0.26$. Taking a threshold of 0.5 for relevant documents, it can be seen that document (b) is relevant for the query while document (c) is not.

Table 2. Relevant document concepts

C1705241:Change (Delta (difference))
C0523744:Lipids NOS (Lipids measurement)
C0023820:Lipoproteins
C0205166:Long
C1515273:Term (Term (temporal))
C0014939:Estrogen (Estrogens)
C1623416:deficiency (deficiency aspects)
C0019932:Hormone (Hormones)
C0559956:Replacement
C1363945:Therapy (Therapy Object (animal model))

Table 3. Non-relevant document concepts

C1963179:Nausea (Nausea Adverse Event)
C0003779:Vasopressin (Argipressin)
C0282412:Editorial

Table 4. Max similarity values of concepts in (a) with respect to concepts in (c)

c1 in text2	maxSim
C0001688	-1
C0523744	1.4955
C0033308	2.1145
C1947971	1.5755
C0014939	2.2687
C0559956	1.5755
C1363945	-1

6. EVALUATION AND DISCUSSION OF THE RESULTS

The OHSUMED 91 corpus [15, 16] was used for the evaluation of the method. The OHSUMED 91 corpus was created for the information retrieval competition TREC 9 [7]. This corpus contains 63 queries and their corresponding relevant and non-relevant documents. The queries have been classified by experts who had agreed about their relevance. One query was selected and a test dataset was formed with a total of 50 documents: 14 relevant documents and 36 non-relevant documents for that query.

The method based on Milhecea's was run calculating the idf from a corpus as well as from the texts being compared. The evaluation was also run with the bap method. Here, a corpus was formed by 49,302 abstracts collected from PubMed, using the keywords "microarray" or "genechip". The taxonomy-based metrics of wup, lch and $path$ were employed to achieve concept semantic similarity. For the bap method, the score of the texts files compared with the query ($score_1$) and the score of the text files compared with themselves ($score_2$) were obtained, which resulted in a final score: $final\_score = \frac{score_1}{score_2}$. A stoplist was considered for the calculation of the bap score.

Then, recall (number of correct answers divided by the number of instances) and precision (number of correct answers divided by the number of answers reported) in the text classification context [17] were calculated, using a threshold of 0.5 to determine whether they were relevant or not. Table 6 shows the respective results.

The best F-score was obtained when employing the wup measure and the idf from a corpus in the adapted method (0.81). A similar value (0.80) was obtained when using lch and

[6] www.d.umn.edu/ tpederse/text-similarity.html

[7] http://trec.nist.gov/data/t9_filtering.html

Table 5. Max similarity values of concepts in (c) with respect to concepts in (a)

c2 in text1	maxSim
C1963179	-1
C0003779	2.2687
C0282412	-1

Table 6. Evaluation Results

Metric	Precision	Recall	F-score
Adapted method (*idf* from corpus)			
wup	0.92	0.73	**0.81**
lch	0.90	0.64	0.75
path	1.0	0.14	0.25
Adapted method (*idf* from files)			
wup	0.90	0.60	0.72
lch	0.91	0.71	**0.80**
path	1.0	0.14	0.25
bap method			
	1	0.07	0.13

calculating the idf from the files being compared. The measures using $path$ were too low in both cases. These results show that the simple $path$ measure that relies only on node-counting is not enough when texts are compared. The lowest score was given by the bap method due to the fact that it looks for the overlapping of words rather than for conceptual relationships.

The similarity method using the wup and lch measures performed well in terms of finding the semantic similarity between two short biological texts. It was demonstrated that the use of taxonomy-based methods has given good results, whilst involving comparatively low computational cost and time.

Similarity values can vary due to errors in text processing, mainly provoked by wrong word-sense-disambiguation or because there is not any path between two concepts in the UMLS ontology. In future work, these aspects will be improved by using WordNet as an alternative option to get a path between concepts and to double-check word meanings. Another aspect to be considered in order to improve the accuracy of the method is the presence of prepositions, which can significantly alter the meaning of the texts. To illustrate this point, (2) shows two sentences taken from different texts. The first sentence contains the word *without*, while the second sentence contains *followed by*. Even if the concepts of the two texts are quite similar, the respective prepositions entail opposite meanings.

EXAMPLE 2. (*1*) ... *leukocyte isolation* without *enrichment for malignant blasts.*

(*2*) ... *leukocyte isolation* followed by *enrichment for malignant blasts by non-malignant cell depletion.*

The methods based on the use of biological ontologies also deal with several irregularities since they depend on the quality and completeness of both the ontology and the annotated corpus they use. Often times, these irregularities are also due to the ambiguous status of biomedical knowledge [18].

7. CONCLUSIONS

This paper has presented a method to calculate the semantic similarity between pairs of short biological texts. The proposed procedure extends the Milhecea's method for general texts that employs words and WordNet to the use of concepts and the public specialised UMLS ontology.

In contrast to existing methods for text similarity, the method proposed in this paper adds more meaning to the words employed in biological contexts. The measures of similarity between concepts used in the proposed method consider the degree of similarity between concepts according to the level in which they are located within the ontology. Furthermore, the method also takes into account the implicit information contained in a corpus which can be easily obtained from the biological literature. Through this combination of features, the proposed method offers a sound balanced between time effectiveness and system performance. However, even though the method has proven to be able to perform well in biological contexts, its overall performance still depends on factors like the accuracy of the ontology or the word-sense-disambiguator used. These aspects will be improved in future work.

8. ACKNOWLEDGEMENTS

The author is supported by a grant from the Mexican National Council for Science and Technology (CONACYT).

9. REFERENCES

[1] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic similarity measures as tools for exploring the gene ontology. In *Pac Symp Bio-comput Proc.*, pages 601–612, 2003.

[2] T. Pedersen, S.V. Pakhomov, S. Patwardhan, and C.G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

[3] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989.

[4] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, pages 19–33, 1997.

[5] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[6] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.

[7] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to latent semantic analysis. discourse. *Discourse Processes*, 25:259–284, 1998.

[8] Peter D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *Proceedings of the Twelfth European COnference on Machine Learning ECML.2001*, pages 491–502, 2001.

[9] J.E. Caviedes and J.J. Cimino. Towards the development of a conceptual distance metric for the umls. *J. of Biomedical Informatics*, 37(2):77–85, 2004.

[10] C. Leacock and M. Chodorow. *Combining local context and WordNet similarity for word sense identification*, pages 305–332. In C. Fellbaum (Ed.), MIT Press, 1998.

[11] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 296–304, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[12] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press, 2006.

[13] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 805–810, 2003.

[14] K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[15] W.R. Hersh, C. Buckley, T.J. Leone, and D.H. Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual ACM SIGIR Conference*, pages 192–201, 1994.

[16] W.R. Hersh and D.H. Hickam. Use of a multi-application computer workstation in a clinical setting. In *Bulletin of the Medical Library Association*, volume 82, pages 382–389, 1994.

[17] D.L. Olson and D. Delen. *Advanced Data Mining Techniques*. Springer, 2008.

[18] C. Pesquita, D. Faria, A.O. Falco, P. Lord, and F.M. Couto. Semantic similarity in biomedical ontologies. *PLoS Compututational Biology*, 5(7), 2009.