Improving Semantic Similarity for Pairs of Short Biomedical Texts with Concept Definitions and Ontology Structure

Olivia Sanchez Graillet
Posgrado en Ciencias e Ingeniería de la Computación

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
Ciudad Universitaria, Coyoacán, 04510, México D.F., México

ABSTRACT

Finding semantic similarity between short biomedical texts, such as article abstracts or experiment descriptions, may provide important information for health researchers. This paper presents a method for calculating text similarity in the biomedical context. The method implements a pairwise concept semantic similarity measure that uses concept definitions and ontology structure. The respective results have demonstrated an improved performance in comparison with a previous version of the method using lexical-based measures as similarity function, as well as with other alternative tools for measuring text similarity.

General Terms:

Semantic text similarity in biomedicine, text mining

Keywords:

semantic text similarity, knowledge discovery, text mining

1. INTRODUCTION

Methods for comparing biomedical texts have been developed for different purposes such as discovering plagiarism in specialised literature [20, 4] or text similarity searching, in which a piece of text is supplied and similar texts are returned [23]. The majority of these methods consider words rather than concepts.

The present paper builds on an earlier method for measuring text similarity (here called SimText) [24] adapted from Mihalcea et al. [16] from general to biomedical context. SimText employs concepts rather than words and has used taxonomy-based methods as similarity functions. To elaborate on this previous procedure and to improve its respective results, a method based on ontological hierarchy and ontological concept definitions [25] (here called SemSim) has been implemented as concept similarity function in SimText.

In what follows, Section 2 describes existing methods for obtaining semantic similarity between concepts, as well as methods for obtaining similarity between texts. Section 3 reviews the data resources used and introduces the proposed improvement of SimText. Section 4 presents the evaluation of the proposed version of SimText and the discussion of the respective results. Section 5 revisits some central points by way of conclusion.

2. RELATED WORK

2.1 Methods for semantic similarity between concepts

There are several techniques for computing semantic similarity between biomedical concepts [15, 19, 21, 30, 8, 13, 27, 2] that can generally be grouped into three categories: corpus-based, taxonomy-based, and taxonomy/corpus-based methods. The subsequent paragraphs outline examples of methods in these three categories.

2.1.1 Corpus-based measures. Large corpus are used to obtain word co-occurrences. Latent semantic analysis (LSA) [9], the pointwise mutual information-information retrieval (PMI-IR) method [28], and the context vector method [18] are based on this technique. In LSA word co-occurrences are obtained by applying a singular value decomposition (SVD) on a term-by-document matrix which represents the corpus, in order to reduce its dimensionality. The resulting vector space is then measured with the cosine similarity function. The PMI-IR method calculates the statistical dependencies between two given words by obtaining their probabilities from a large corpus, such as the web. Both LSA and PMI-IR have shown to be effective, but highly computationally expensive. The context vector method relies on the idea that similar words are surrounded by similar contexts. A fixed window is used to obtain word co-occurrences from corpus, and semantic relatedness is calculated as the cosine of the angle between the context vectors of the two words being compared. Pre-processing is applied to text in order to clean noise and redundancy. The results depend on the availability of suitable corpora, an efficient data-cleaning process, and the amount of text used.

2.1.2 Taxonomy-based measures. A taxonomy (or ontology) where concepts are commonly connected with “Is_a” relationships is used. For example, the *path* method consists of the inverse of the shortest path (length) between two concepts in the taxonomy. Rada et al. [21] applied this idea to a taxonomy where concepts were connected by “broader than” relationships, while Caviades

and Cimino [3] applied it to the UMLS ontology.
Leacock and Chodorow [10] (lch) and Wu and Palmer [30] (wup) developed variations of the $path$ measure. The lch measure divides the shortest path between two concepts ($length$) by twice the maximum depth of the "Is_a" hierarchy ($depth$) and smooths it with $-log$, as shown in (1).

$$Sim_{lch}(c_1, c_2) = -log \frac{length}{2 \cdot depth} \quad (1)$$

While the wup similarity score is calculated with equation (2).

$$Sim_{wup}(c_1, c_2) = \frac{2 \cdot depth_{LCS}}{depth_{c1} + depth_{c2}} \quad (2)$$

Where $depth_{LCS}$ is the depth of the least common subsumer (LCS) of concepts c_1 and c_2.
More recently, Batet et al. [2] proposed a method that considers all superconcepts and not only the minimal paths between two concepts. Cases with a small number of shared superconcepts are penalised. The final measure is the ratio between the non-shared superconcepts and the sum of non-shared and shared superconcepts smoothed by $-log_2$.
The advantage of taxonomy-based measures is their simplicity and low computational cost.

2.1.3 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 (3).

$$IC(c) = -log\, p(c) \quad (3)$$

An example of this measure is the one of Resnik [22] that is calculated as $IC(LCS)$. Where LCS is the least common subsumer of concepts c_1 and c_2.
Lin [14] developed a variation of the Resnik measure, in which a normalisation factor is added, as shown in (4).

$$Sim_{lin}(c_1, c_2) = \frac{2 \cdot IC(LCS)}{IC(c_1) + IC(c_2)} \quad (4)$$

The results given by these measures depend on the coverage and size of the corpus used.

2.1.4 Other semantic similarity methods. Clustering methods group similar concepts according to given features. Clusters are defined for each branch in the hierarchy with respect to the root node. The common node specificity, given by the LCS of the two concepts, states that lower level concept pairs are more similar than higher level concept pairs. For example, the method of Al-Mubaid and Nguyen [1] includes features such as cross-modified path length, common specificity of two concepts, and local granularity of the clusters.
The method recently proposed by Sanchez-Graillet [25] (SemSim), uses the "Is_a" hierarchy of the SNOMED-CT ontology and the logical definitions (in OWL format) of the concepts. The semantic similarity value between two OWL defined concepts C and D is calculated with (5).

$$SemSim(C, D) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} sim(c_i, d_j)}{n \cdot m} \quad (5)$$

Where:

—$sim(c_i, d_j)$ is the similarity between concept $c_i \in T_C$ and concept $d_j \in T_D$

—T_C is either the set of concept names (classes) contained in the *intersectionOf* and *someValuesFrom* declarations in the definition of C (i.e., C is a combined concept) or C itself, if C has only one parent in its *subClassOf* declaration (i.e., C is a general concept)

—T_D is either the set of concept names (classes) contained in the *intersectionOf* and *someValuesFrom* declarations in the definition of D (i.e., D is a combined concept) or D itself, if D has only one parent in its *subClassOf* declaration (i.e., D is a general concept)

—n and m are the number of concepts in T_C and T_D respectively

$sim(c_i, d_j)$ is calculated in (6) according to the SNOMED-CT ontology.

$$sim(c, d) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \quad (6)$$

Where:

—λ_1 is the number of shared ancestors between c and d

—λ_2 is the number of unshared ancestors between c and d, counting c and d

In addition to the ontological information of the concepts being compared, this method considers the implicit relations derived from the respective concept definitions.

2.2 Methods for text similarity

In this section, SimText and two freely available tools to calculate text similarity are described. Based on Mihalcea et al.'s method, SimText compares two texts by adding the highest semantic similarity scores ($maxSim$) between the concepts contained in both texts, and then weights this value with idf, and normalises the final similarity measure. Similarity between texts T_1 and T_2 is calculated with (7).

$$SimText(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{c \in (T_1)} (maxSim(c, T_2) \cdot idf(c))}{\sum_{c \in (T_1)} idf(c))} + \frac{\sum_{c \in (T_2)} (maxSim(c, T_1) \cdot idf(c))}{\sum_{c \in (T_2)} idf(c))} \right) \quad (7)$$

Where idf is the inverse document frequency [26] of a concept c, which defines its specificity.
idf corresponds to log(number of documents in the corpus / number of documents where concept c appears).
The Text::Similarity tool (v0.08)[1] is based on the Lesk value used for word sense disambiguation [11] that relies on the idea that the higher the number of overlapping words between two files, the more related those files are. Text::Similarity counts the number of overlapping (shared) words of two given files or strings, without taking into account word order, and (optionally) normalises the obtained score by the lengths of the files.
The eTBlast [12] text-pair comparison tool[2] receives pairs of small text, such as paragraphs or abstracts as input. Then the cosine

[1] available at http://text-similarity.sourceforge.net
[2] http://etest.vbi.vt.edu/etblast3/index/paircompare

coefficient [23] is used as similarity function with vectors $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n)$, where n is the number of unique words in the library (set of Medline documents), $x_i = 1$ if word i is in the query, otherwise $x_i = 0$, and $y_i = 1$ if word i is in the library text, otherwise $y_i = 0$. The cosine similarity function is weighted by idf with $log_{1.6}$ as shown in (8).

$$cosine\, coefficient = \frac{\sum_{i=1}^{n} x_i \cdot y_i \cdot idf_i}{\sqrt{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}} \qquad (8)$$

Where $idf = log_{1.6}$(number of documents in the library / number of documents with term i).

$log_{1.6}$ was chosen in order to down-weight the score of words in user queries and Medline abstracts, since it was proved that it did not significantly alter the weight of words occurring up to four times, while it did so more significantly when words occurred more than five times. Pre-processing to remove stop-words from text is applied before forming the respective vectors.

The eTBlast comparison tool outputs a similarity ratio (range 0-100%) calculated as the eTBLAST similarity score of the two texts over the eTBlast similarity score of the first text against itself. Different similarity score ratios might occur depending on the order in which the two texts are queried. The evaluation in Section 4 uses the highest similarity eTBlast scores obtained for each text pair.

3. MATERIALS AND METHODS

3.1 UMLS MetaMap

The Unified Medical Language System (UMLS)[3] compiles several health and biomedical vocabularies and standards to enable interoperability between computer systems. UMLS also contains tools for accessing such data resources.

MetaMap[4] is part of the lexical tools provided by UMLS. It maps terms to concepts in the UMLS Metathesaurus from free texts by using a knowledge-intensive approach based on symbolic, NLP (Natural Language Processing) and computational linguistic techniques. In the present work, MetaMap has been configured with the following options:

—y: attempts to disambiguate among concepts scoring equally well

—Y: mappings with more concepts are scored 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 : shows the UMLS CUI for each concept displayed

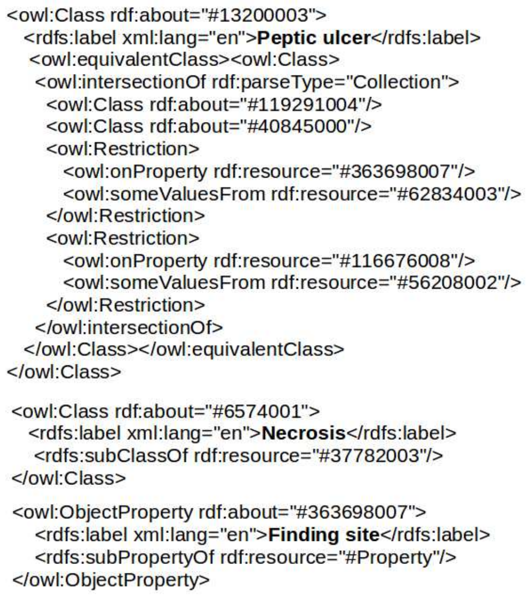
—c: disables the displaying of the list of Metathesaurus candidates

3.2 The SNOMED-CT ontology

The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT)[5] is an organised collection of medical terms, synonyms and definitions covering diseases, findings, and procedures. The SNOMED-CT vocabulary includes “Is_a” relationships that link concepts within a hierarchy and attribute relationships that link concepts across hierarchies [29]. The “Is_a” relationship relates a concept to its more general concepts. For example, “viral pneumonia” has an “Is_a” relationship to the more general concept “pneumonia”. Attribute relationships on the contrary, represent other aspects of the definition of a concept. For example, “viral pneumonia” has a “causative agent” relationship to “virus” and a “finding site” relationship to “lung”[6].

An OWL ontology has been derived from the SNOMED-CT vocabulary [25]. The ontology contains 297,327 classes (concept definitions) organised into top-level hierarchies joint together by a root node, and attribute relationships that correspond to 62 OWL object properties. There are more than 890,000 logically-defined relationships among all concepts[7].

As an example, the respective OWL definitions of concepts “Peptic ulcer” and “Necrosis”, and property “Finding site” are shown in Fig. 1.



```
<owl:Class rdf:about="#13200003">
  <rdfs:label xml:lang="en">Peptic ulcer</rdfs:label>
   <owl:equivalentClass><owl:Class>
    <owl:intersectionOf rdf:parseType="Collection">
      <owl:Class rdf:about="#119291004"/>
      <owl:Class rdf:about="#40845000"/>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#363698007"/>
        <owl:someValuesFrom rdf:resource="#62834003"/>
      </owl:Restriction>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#116676008"/>
        <owl:someValuesFrom rdf:resource="#56208002"/>
      </owl:Restriction>
    </owl:intersectionOf>
  </owl:Class></owl:equivalentClass>
</owl:Class>

<owl:Class rdf:about="#6574001">
  <rdfs:label xml:lang="en">Necrosis</rdfs:label>
   <rdfs:subClassOf rdf:resource="#37782003"/>
</owl:Class>

<owl:ObjectProperty rdf:about="#363698007">
   <rdfs:label xml:lang="en">Finding site</rdfs:label>
   <rdfs:subPropertyOf rdf:resource="#Property"/>
</owl:ObjectProperty>
```

Fig. 1. Example of OWL classes (Adopted from [25])

3.3 Text similarity method

In [24], texts were parsed with MetaMap in order to retrieve files containing the UMLS CUIs (concept unique identifiers) of the corresponding words. The obtained files were input into SimText using the taxonomy-based methods $path$, wup and lch as similarity functions.

In the current study, UMLS CUIs are mapped by a Perl program to SNOMED-CT CUIs according to the UMLS metathesaurus database. Concepts without a corresponding SNOMED-CT CUI are ignored. The two files containing SNOMED-CT CUIs are compared with equation (7) to determine their similarity text value. For example, the corresponding UMLS and SNOMED-CT CUIs obtained with MetaMap and the Perl program for (a), (b), and (c) are shown in Table 1. Where (a) is a OHSUMED query, (b) is a relevant answer for (a), and (c) is an irrelevant answer for (a).

[3] www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_003.htm
[4] metamap.nlm.nih.gov/
[5] www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

[6] http://www.ihtsdo.org/snomed-ct/snomed-ct0/snomed-ct-components/
[7] ihtsdo.org/fileadmin/user_upload/doc/download/doc_SnomedCT ReleaseNotes_Current-en-US_INT_20130731.pdf

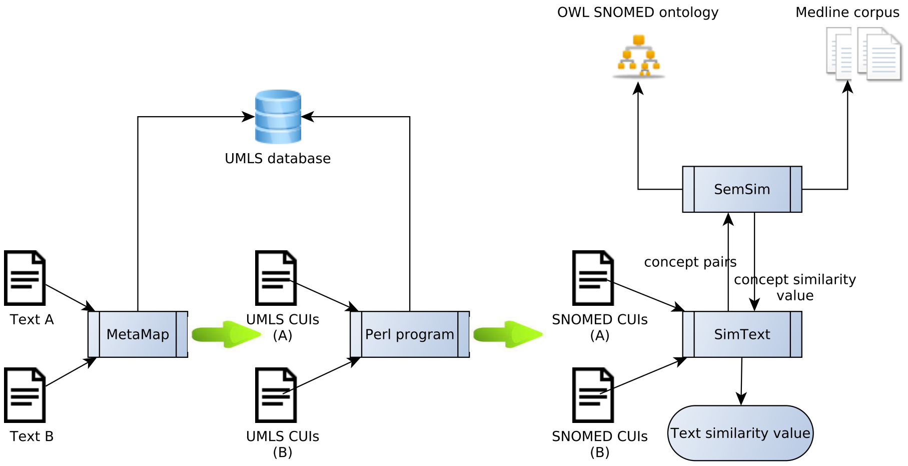


Fig. 2. Procedure for calculating text similarity

(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) Irrelevant document: "Nausea and vasopressin [editorial]"

Table 1. Corresponding CUIs for concepts

Concept	UMLS CUI	SNOMED CUI
Query (a)		
Effects	C1280500	253861007
Lipid	C0023779	70106000
Progesterone	C0033308	16683002
Estrogen	C0014939	41598000
Replacement	C0559956	282089006
Therapy	C0087111	276239002
Relevant document (b)		
Lipid	C0023779	70106000
Lipoproteins	C0023820	301861005
Long	C0205166	255511005
Estrogen	C0014939	41598000
Hormone	C0019932	87568004
Replacement	C0559956	282089006
Therapy	C0087111	276239002
Irrelevant document (c)		
Nausea	C0027497	422587007
Vasopressin	C0003779	420773001

In the current study, SemSim is used as similarity function in the text similarity method presented in [24]. SemSim obtains the ancestors and definitions of the concepts contained in the corresponding texts from the OWL ontology; while idf is calculated for each concept based on the corpus formed by 49,302 Medline abstracts used in [24] that are parsed to the corresponding UMLS CUIs.

Fig. 2 illustrates the flow of the proposed procedure for calculating text similarity. In the next section, the method's performance evaluation is presented.

4. EVALUATION AND DISCUSSION

In order to evaluate the performance of the proposed improved method, the system has been used to classify a set of texts as relevant or irrelevant according to a given threshold. The chosen threshold reflects the classification made by human judges about the relevance of the documents in relation to a given query. The evaluation is based on the idea that the more relevant a text is in relation to the given query text, the more similar these two texts are.

As in [24], the OHSUMED-91 corpus [6, 7] was used as baseline and test data for the corresponding evaluation. The OHSUMED-91 corpus was created for the TREC9-IR competition[8]. This corpus contains 63 queries and their corresponding relevant and irrelevant documents. The queries were classified by experts who agreed about their relevance.

For the evaluation of the proposed method, one query was selected from the OHSUMED-91 corpus with the corresponding test dataset formed by 50 documents: 14 relevant documents and 36 irrelevant documents in the context of this particular query.

SimText using SemSim was compared with SimText using wup, which obtained the best performance in the comparison made in [24]), as well as with Text::Similarity, and eTBlast.

The following metrics were used in order to measure the performance of the methods in the text classification context [17]:

—True-positive rate (also called recall): true positives / (true positives + false negatives)

—False-positive rate: false positives / (false positives + true negatives)

—Precision: true positives / (true positives + false positives)

—F-score: 2 · (precision · recall / precision + recall)

To determine answer relevance, thresholds 0.3, 0.5 and 0.7 were used as evaluation criteria. Table 2 contains the respective results.

The corresponding ROC space depicted in Fig. 3 shows the relation between TP-rates and FP-rates of the evaluated methods according to the respective columns in Table 2 for the different thresholds.

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

Table 2. Text similarity results using thresholds 0.3, 0.5 and 0.7

Method	TP-rate	FP-rate	Precision	F-score
Threshold 0.3				
SemSim-SimText	1.00	0.12	0.70	0.82
SimText-wup	1.00	0.20	0.61	0.76
Text::Similarity	0.29	0.00	1.00	0.44
eTBlast	0.14	0.00	1.00	0.25
Threshold 0.5				
SemSim-SimText	0.86	0.05	0.86	0.86
SimText-wup	0.73	0.03	0.92	0.81
Text::Similarity	0.07	0.00	1.00	0.13
eTBlast	0.00	0.00	0.00	0.00
Threshold 0.7				
SemSim-SimText	0.21	0.00	1.00	0.35
SimText-wup	0.14	0.00	1.00	0.25
Text::Similarity	0.00	0.00	0.00	0.00
eTBlast	0.00	0.00	0.00	0.00

ROC (Receiver Operating Characteristics) graphs have been used for visualising the performance of classifiers in areas like machine learning and data mining [5].

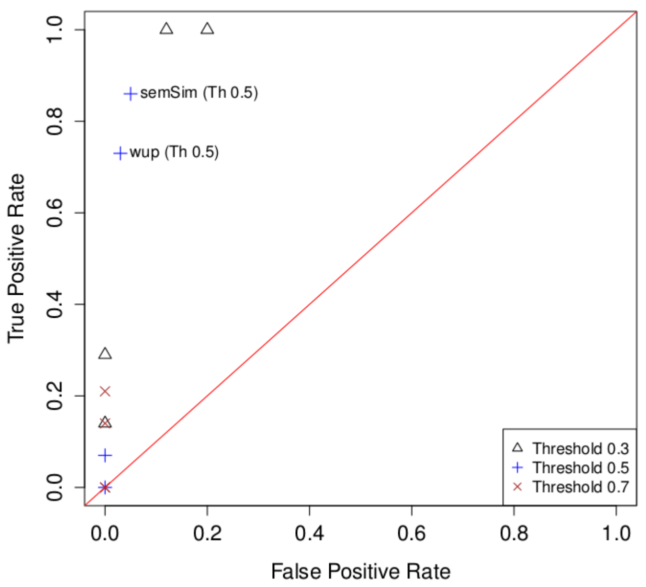


Fig. 3. ROC space showing classification points

In addition, Spearman correlation coefficients among the respective results were computed in order to study the behaviour of the methods with respect to each other (See Table 3).

4.1 Discussion

As the above evaluations illustrate, SimText using SemSim outperforms the other methods in relation to the three chosen thresholds, followed by SimText using *wup*. On the other hand, Text::Similarity and eTBlast perform poorly regarding the three thresholds.

Table 3. Correlation among results of text similarity methods

	SemSim	wup	Text::Sim	eTBlast
SemSim	1.00			
wup	0.80	1.00		
Text::Sim	0.74	0.74	1.00	
eTBlast	0.74	0.73	0.95	1.00

Threshold 0.5 is considered the most accurate relevance measure of the three thresholds tested, since it reflects the answers closest to the ones of humans to asses the relevance of query answers. At this threshold, SimText using SemSim shows high precision, recall (TP-rate), and F-score (0.86 each) as well as low FP-rate (0.05). These results represent a good performance of the method.

In an ROC space, one point is better than another if it is located to the northwest of the first point (i.e., TP-rate is higher, FP-rate is lower, or both). Classifiers on the left side of an ROC graph near the X axis may classify positive only with strong evidence, so FP-errors are lower, but often TP-rates are also low. On the other hand, classifiers on the upper right side of an ROC graph may classify positive with weak evidence, so TP-rate is high, but often FP-rates are also high [5].

Based on these observations, it can be seen in Fig. 3 that SimText using SemSim followed by SimText using *wup* have the best performance with thresholds 0.5 and 0.3 (crosses and triangles, respectively).

The fact that both Text::Similarity and eTBlast are based on words rather than on conceptual relationships might be responsible for their low performance, since lexical comparison involves a lower level of abstraction than semantic comparison of concepts.

Table 3 shows a strong correlation between SimText using *wup* and SemSim (0.80), and a strong correlation between Text::Similarity and eTBlast (0.95). These correlations indicate that the two knowledge-based methods behave comparably, while the two lexical-based methods behave comparably. Furthermore, the correlations between knowledge-based and lexical-based methods are high (about 0.74). In general, it is worth noting that the results of all methods correlate highly with each other, perhaps due to the specialised context in which they perform.

5. CONCLUSIONS

In this paper, a previous procedure for calculating semantic similarity between concepts (SemSim), which is based on a given ontological hierarchy and concept definitions, has been used as similarity function of a novel method for calculating similarity between two short biomedical texts (SimText). SemSim considers the degree of similarity between concepts according to the number of common and uncommon ancestors between them in the specialised SNOMED-CT ontology as well as the logical definitions of the concepts.

SimText using SemSim has been compared with SimText using a taxonomy-based semantic similarity method (*wup*), as well as with other tools for calculating text similarity (Text::Similarity, eTBlast). SimText using SemSim has shown the best performance among the methods tested. SimText together with SemSim involves a higher level of abstraction than lexical-based methods for text similarity in the specialised context of biomedicine. Since the present work is on short texts, grammatical structures are not taken into account. In future work, the effect of including such structures in the proposed method will be analysed. However, the overall performance of SimText-SemSim still depends on factors such as text preprocessing, accuracy of the mappings from words to concepts, completeness of the ontology, and the respective corpus or database used. In future work, possible ways of overcoming these problems need to be addressed.

6. ACKNOWLEDGEMENTS

The authorship of this work was supported by a grant from the Mexican National Council for Science and Technology (CONACYT).

7. REFERENCES

[1] H. Al-Mubaid and H.A. Nguyen. A cluster-based approach for semantic similarity in the biomedical domain. In *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, pages 2713–17, 2006.

[2] M. Batet, D. Sánchez, and A. Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of Biomedical Informatics*, 44(1):118–125, 2011.

[3] 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.

[4] J. Chen, R. Chau, and Ch-H. Yeh. Discovering parallel text from the world wide web. In *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*, volume 32, pages 157–161. Australian Computer Society, Inc., 2004.

[5] T. Fawcett. Roc graphs: notes and practical considerations for data mining researchers (hpl-20034). Technical report, HP Laboratorie, 2003.

[6] 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.

[7] 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.

[8] J.J. Jiang and D.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.

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

[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] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26. ACM, 1986.

[12] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H.R. Garner. Text similarity: An alternative way to search medline. *Bioinformatics*, 22(18):2298–304, 2006.

[13] Y. Li, Z.A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. on Knowl. and Data Eng.*, 15(4):871–882, 2003.

[14] D. 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.

[15] 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.

[16] R. Mihalcea, C. Corley, and C. 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.

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

[18] S. Patwardhan and T. Pedersen. Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 workshop, making sense of sense: Bringing computational linguistics and psycholinguistics together*, pages 1–8, 2006.

[19] 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.

[20] A. Pertsemlidis and H.R. Garner. Text comparison based on dynamic programming. *IEEE Engineering in Medicine and Biology Magazine*, 23(6):66–71, 2004.

[21] 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.

[22] Ph. 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.

[23] G. Salton. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[24] O. Sanchez-Graillet. Semantic similarity measure for pairs of short biological texts. *International Journal of Applied Information Systems*, 4(5):1–5, 2012.

[25] O. Sanchez-Graillet. Using concept definitions and ontology structure to measure semantic similarity in biomedicine. *International Journal of Applied Information Systems*, 4(5):1–5, 2014.

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

[27] I. Spasic and S. Ananiadou. A flexible measure of contextual similarity for biomedical terms. In *Pacific Biocomputing Symposium*, pages 197–208, 2005.

[28] P.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.

[29] G. Wade. SNOMED CT: The Clinical Data Standard. Overview and Application to eHRs, 2013.

[30] Z. Wu and M. 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.