An Effective TF/IDF-based Text-to-Text Semantic Similarity Measure for Text Classification

Shereen Albitar, Sébastien Fournier, Bernard Espinasse

Aix-Marseille University, LSIS UMR CNRS 7296
Domaine universitaire de St Jerome, F-13397 Marseille Cedex 20, France.
{first_name.last_name@lsis.org}

Abstract. The use of semantics in tasks related to information retrieval has become, in recent years, a vast field of research. Considering supervised text classification, which is the main interest of this work, semantics can be involved at different steps of text processing: during indexing step, during training step and during class prediction step. As for class prediction step, new text-to-text semantic similarity measures can replace classical similarity measures that are traditionally used by some classification methods for decision-making. In this paper we propose a new measure for assessing semantic similarity between texts based on TF/IDF with a new function that aggregates semantic similarities between concepts representing the compared text documents pair-to-pair. Experimental results demonstrate that our measure outperforms other semantic and classical measures with significant improvements.

Keywords: Classification, Semantics, Text-to-Text Semantic Similarity.

1 Introduction

Supervised text classification is currently a challenging research topic, particularly in areas such as information retrieval, recommendation, personalization, user profiles etc. Generally, supervised text classification methods use syntactical and statistical models for text document representation. This applies to the most popular text classification methods such as: Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio, and k Nearest Neighbors (kNN). These representation models ignore all semantics that reside in the original text that can help in text classification.

However, it is possible to use semantic resources to take into account meaning of the words in text representation in order to improve classification effectiveness. Thus, resulting text representation models can take into account synonyms, relations between words and also can resolve some ambiguities. Many researchers reported that using semantics in text classification improves its effectiveness in specific domains especially by deploying domain specific semantic resources [1].

There are several possibilities for involving semantics during the process of supervised texts classification. In this work, we are interested in involving semantics in class prediction step using, text-to-text semantic similarity measures. Hence, we pro-

adfa, p. 1, 2014.

pose a new text-to-text semantic similarity measure (TF/IDF based), called in this article SemTFIDF, and we present an experimental study to evaluate it in the context of text classification. In addition, we compare it with another text-to-text semantic similarity measure proposed in the literature (IDF based) called semIDF in this article, and also with the well-known classical similarity measure Cosine that is usually deployed in the Vector Space Model. These experiments are carried out in the biomedical domain using the Ohsumed corpus and domain specific knowledge base Unified Medical Language System (UMLS®) and Rocchio with Cosine as the baseline [2].

Second section reviews state of the art methods deploying semantics in classification or other tasks related to information retrieval or data mining. Third section focuses on the use of semantics during class prediction step and presents our new measure (SemTFIDF) based on TF/IDF and suitable for supervised text classification. Fourth section presents experimental setup that we used to evaluate our new measure. Fifth section analyses the experimental results obtained with Cosine classical similarity measure and these two text-to-text similarity measures (SemIDF and SemTFIDF). Finally, we conclude and present our perspectives for future works.

2 Involving Semantics in Supervised Text Classification

Typically, most of supervised text classification techniques are based on statistical and probabilistic hypothesis in both training and classification procedures. As for text representation or indexing, the importance of a term to a document is assessed using the frequency of its occurrences in the document. So far, the intended meaning of terms and the relations among them are not treated or used in text classification. In other words, semantics and relatedness behind literally occurring words are missing in classical text classification techniques. However, last few years have seen different approaches seeking to introduce semantics during indexing, training and prediction.

Involving Semantics in Indexing. Semantics can be used during indexing for a semantic text representation. Indeed, vector-based (binary or TF/IDF) representations, used by these classical supervised classification methods, enable semantic integration or "conceptualization" that enriches document representation model using background knowledge bases [1, 3]. To involve semantic features in indexing, state of the art approaches used either implicit semantics through topic modeling [4] or explicit semantics derived from structured resources and used as new features for text representation [1, 6]. Other approaches use either type in semantic kernels to support some supervised classification techniques [5].

Involving Semantics in Training. In these approaches, concepts replace words in text representation. In addition, the hierarchy and the relations among the added concepts are taken into consideration in the training step which affects the learned model, so the classification model is either the entire ontology or part(s) of its hierarchy. Both works [7, 8] used the hierarchical structure of semantic resources to involve related concepts in text representation. Authors in [8] used propagation algorithm to propagate the weights of identified concepts in patents to their superconcepts. Furthermore, authors in [9] used similar concepts in order to enriched text representation and proposed the approach Enriching vectors. Similarities among concepts are assessed using relations between concepts in the semantic resource. Both Generalization [7, 8] and Enriching vectors [9] involve semantics in the classification model implicitly.

Involving Semantics in Class Prediction. According to the literature, most research focused on enriching text representation with semantics and used classical techniques for prediction [10]. Only few works tried to involve semantics in class prediction by proposing new Text-To-Text Semantic Similarity Measures like Semantic Trees or Concept Forest in [10, 11]. Both works involve explicitly the hierarchy of ontology in text representation and training as a classification model. As for assessing the similarity between two documents, authors chose to use a relatively simple formula inspired from the classical cosine measure and reported significant improvement in classifying web documents according to Yahoo! categories.

New semantic approaches for assessing text-to-text similarities seem to be feasible using semantic similarities among concepts pair-to-pair. In fact, such approaches involve semantics in document comparison and in class prediction as well by discovering similarities between texts considering semantically similar terms in addition to lexically similar ones. According to the literature, assessing the semantic similarity between concepts of semantic resources has attracted the attention of many researchers which resulted in proposing numerous semantic similarity measures [12].

3 Text-to-Text Semantic Similarity Measures

In this section, we are interested in involving semantics in the prediction step of text classification process, particularly, through Text-To-Text Semantic Similarity Measures. In fact, some classifiers in the vector space like Rocchio use this kind of measures in class prediction as the criterion with which they choose the most similar class for a treated document. We propose a new measure for assessing semantic similarity between two Bag of Concepts (BOCs) representing two text documents (or a document and a centroïd in the case of a Rocchio classifier). First, we present some related work on text-to-text semantic similarity measures. Then, we present a new text-to-text semantic similarity measure based on a new aggregation function based of TF/IDF weighting scheme.

3.1 Related Works on Semantic Text-to-Text Similarity

In [13], authors proposed an aggregation function that assesses the semantic similarity between two groups of concepts using the mean of similarities of all combinations of pairs of concepts between these groups. Azuaje, Wang [14] proposed a similar aggregation function that takes into consideration maximum semantic similarities between each concept of g_1 and all concepts from g_2 and *vice versa*. Authors in [8] proposed a propagation algorithm to attribute weights to subsumers involving them in text representation. Furthermore, authors proposed a new text-to-text similarity measure based on these weights as well as the semantic similarity between concepts pair-to-pair. This new similarity measure is the prediction criterion that replaces classical text-to-text similarity of the vector space model like Cosine. Authors reported better clustering of patents using semantic similarities [8].

Authors in [15, 16] developed a different aggregation function (that we refer to later by SemIDF) for comparing short texts or phrases using semantic similarities and Inverse Document Frequency – Idf of text concepts. This function improved significantly text-to-text similarity on Microsoft paraphrase corpus [17] as compared to the classical Cosine similarity measure [15]. It demonstrated high accuracy when applied to automatic short answer grading [16]. The main drawback is that this approach ignores all dependencies between words in sentences.

In the context of text classification, we tested most of the similarity measures presented in this section using the tools and resources presented in section 4. According to results of our preliminary tests, only the measure proposed by Mihalcea et al. 2006 [15] demonstrated some satisfactory results.

3.2 A TF/IDF-based Text-To-Text Semantic Similarity Measure (SemTFIDF)

We propose a new aggregation function for assessing text-to-text semantic similarity that adapts the previous measure to text classification by using TF/IDF weights instead of IDF weights. In fact, TF/IDF reflects how a feature is important to a document in a corpus. Thus, our measure takes into consideration the importance and the specificity of a feature to each of the compared documents instead of its importance to the corpus in general.

This measure is applied on indexed conceptualized documents represented as BOCs. This measure aggregates semantic similarities between concepts of the compared documents pair-to-pair. An aggregation function calculates the semantic similarity between the compared documents using their representation, and the semantic similarities between their concepts pair-to-pair that are stored in the semantic proximity matrix. This measure can be used in decision-making in order to involve semantics in class prediction of supervised text classification.

Given two text documents represented in the same feature space as BOCs and weighted using TF/IDF scheme, we propose a new measure for assessing the semantic similarity between these documents according to the following formula:

$$SemSim(T_1, T_2) = \frac{1}{2}\left(\frac{\sum_{c \in T_1} maxSim(c, T_2) * TFIDF_1(c)}{\sum_{c \in T_1} TFIDF_1(c)} + \frac{\sum_{c \in T_2} maxSim(c, T_1) * TFIDF_2(c)}{\sum_{c \in T_2} TFIDF_2(c)}\right) \qquad (1)$$

where: $\mathrm{maxSim}(\mathrm{c}, \mathrm{T}_2)$ is the maximum similarity between the concept (c) and all concepts representing (T_2), and TF/IDF$_1$(c) is the weight of the concept c in document $T_{1.}$

In fact, this measure is a TF/IDF-weighted average of maximum similarities between each concept from the first document with all concepts representing the second one and vice versa. In the following, we present an experimental study in the context of supervised text classification. Next section presents the platform used in this study.

4 Experimental Setup

In order to assess the effect of Text-To-Text Semantic Similarity Measures, we use the experimental platform illustrated in Fig. 1. This platform uses Rocchio for training and prediction as the classification technique. This technique deploys TFIDF as a weighting scheme and *Cosine*, *SemIDF* and *SemTFIDF*, our contribution, as similarity measures. This section presents resources and tools used in this experimental study.

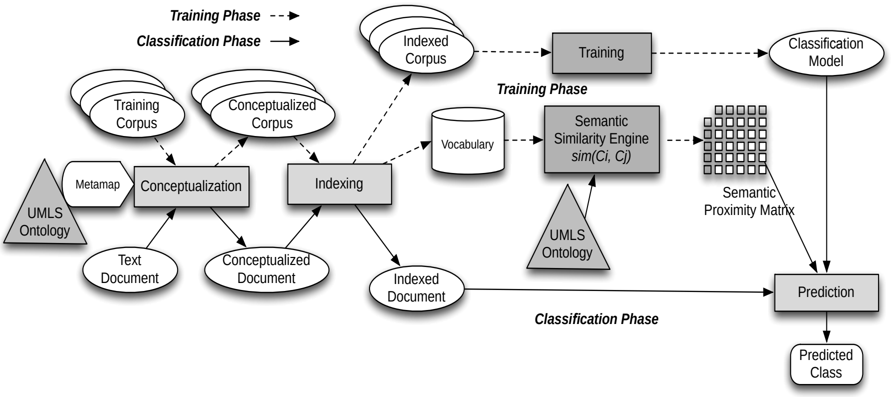

Fig. 1. Platform for supervised text classification deploying Semantic Similarity Measures

Unified Medical Language System *(UMLS®)* was developed at the National Library of Medicine (NLM) in the intent to model the language of biomedicine and health and to help computers understand the language of medicine. It organizes concepts of the various source vocabularies (like MeSH, SNOMED-CT, etc.) according to their senses grouping common concepts together. We choose to use SNOMED-CT exclusively as it provides a large nomenclature on clinical terms.

Ohsumed corpus *[18]* is composed of abstracts of biomedical articles of the year 1991 retrieved from the Medline database indexed using MeSH (Medical Subject Headings). The corpus is divided into Training and Test sets, so experiments are done in two phases: Training and Test. In this work, we restricted this corpus to the five most frequent classes listed in in Table 1.

Category	***Training***	***Test***
C04	972	1251
C23	976	1181
C06	588	632
C14	1192	1256
C20	502	664
Total	4230	4984

Table 1. Ohsumed Corpus

MetaMap. The major goal of MetaMap [19] developers at the NLM was to improve medical text retrieval using UMLS Metathesaurus. Indeed, MetaMap can discover links between medical text and the knowledge in the Metathesaurus. We apply complete conceptualization using MetaMap results as described in our earlier work [3] using UCI's of concepts for conceptualization implies using concepts as features during indexing, documents are thus represented as bags of concepts (BOCs).

Semantic Similarity Engine. As shown in Fig. 1, the semantic similarity Engine computes, using the vocabulary and UMLS ontology, *the semantic proximity matrix.* We chose to use the following ontology-based semantic similarity measures: (i) ***cdist***

[20]; (ii) ***wup*** [21]; (iii) ***lch*** [22]; (iv) ***zhong*** [23] and (v) ***nam*** [12]. Our choice of ontology-based measures is for their efficiency as compared to other families.

Semantic Proximity matrix is a square matrix in which each cell represents the similarity between elements to which row and column correspond. We deploy the semantic engine to assess semantic similarity between concepts of the BOCs pair-to-pair in SNOMED-CT using a semantic similarities measure.

Rocchio Classification Method. Rocchio or centroïd-based method is widely used in Information Retrieval tasks, in particular for relevance feedback and was investigated for the first time by J.J.Rocchio [2]. Afterwards it was adapted for text classification. For centroïd-based classification, each class is represented by a vector at the center of the sphere (centroïd) delimited by training documents related to this class. The class of a new document is the one represented by the most similar centroïd.

In this work, we consider Rocchio an adequate baseline text classifier for its efficiency and simplicity in addition to its extendibility with semantic resources at both levels: text representation and similarity calculation (see section 2). Training is realized on the corpus and so five class centroïds are calculated for each of the classes. As for prediction step, the test document vector is compared to each of the centroïds learned during training. The platform uses two Text-To-Text Semantic Similarity Measures along with Cosine as a baseline to assess the similarity between the vector of the document and the vector of the centroïd.

5 Experimental Results

In these experiments, the platform executes classification five times once for each of the semantic proximity matrices and once for each aggregation function. Rocchio learns once a unique classification model as a set of centroïds. As for classification, Rocchio uses each of (*semIDF* and *semTFIDF*) in prediction using one of the five proximity matrices resulting in $5 * 2 = 10$ executions. The detailed results from these executions that are related to each semantic similarity measure (between concepts pair-to-pair) are grouped together to analyze the impact of Text-To-Text Semantic Similarity measures on the effectiveness of Rocchio.

In this work we used three evaluation measures: *Precision*, *Recall*, and *F_β-Measure*. Most classification techniques emphasize on either Precision or Recall, thus we use their harmonic mean in Fβ-Measure which is more significant [24].

In next subsections, we use as a baseline of comparison Rocchio with *Cosine* classical similarity measure applied on conceptualized Ohsumed using the CUIs of the best mapped concepts. First we present experiments using a text-to-text semantic similarity measure based on IDF proposed in the literature [15, 16] (cf. 3.1 section), and then, we present experiments using our text-to-text semantic similarity measure based on TF/IDF (cf. 3.2 section).

Results using SemIDF Measure . Results of these experiments are detailed in Fig. 2a. We notice that using this semantic similarity measure for prediction in Rocchio did not improve its performance at MacroAveraged level. Nevertheless, local significant improvements occurred when treating documents related to (C06) that is one of the least populated classes in the training corpus. This improvement varied from

(5.44%) using *wup* to (15.16%) using *lch* resulting in F1-Measure ranging between (57.65%) and (62.97%). These improvements are statistically significant according to McNemar test. Other improvements occurred as well: the first is significant using *lch* on (C04) and the second using *cdist* on (C14). Note that the class (C06) is the least populated class among the five considered classes.

Results using SemTFIDF Measure. Using our TF/IDF-based semantic similarity measure (SemTFIDF) for prediction improved the classification results of (C06). Detailed results are in Fig. 2.b. This improvement is high using all of the five semantic similarity measures ranging between (16.46%) and (18.13%) for *nam* and *wup* respectively. These improvements led to a better F1-Measure in the range [63.68%, 64.60%] as compared with results using Cosine as similarity measure on the same class (54.68%). Using all measures, except for *nam*, improved the F1-Measure of classes (C04) and (C14), these improvements are lower if compared to those on (C06). As for (C04), the improvements ranged from (2.75%) to (4.96%) using *zhong* and *lch* respectively resulting in F1-Measure in [74.65%, 76.25%]. On the other hand, improvements treating (C14) ranged from (0.18%) to (3.67%) using *wup* and *cdist* respectively resulting in F1-Measure in [73.01%, 75.55%]. Only three similarity measures *cdist*, *lch* and *zhong* increased Rocchio's Macro F1-measure.

According to previous observations, the maximum increase in F1-Measure occurred when treating the class (C06) and is of a percentage of (18.13%) using *lch* for Semantic Text-To-Test Similarity measure. In fact, this class is the least populated class in the corpus and Rocchio with Cosine obtained on the completely conceptualized corpus a relatively low value of F1-Measure for this class. These improvements at class level influenced the MacroAveraged F1-Measure with a gain ranging from (0.20%) to (2.27%) using semantic similarities *lch* and *cdist* respectively. In fact, the overall performance of Rocchio using Cosine on the conceptualized corpus is significantly different from its performance on the corpus after applying our measure according to McNemar test and using two semantic similarity measures *zhong* and *dist*.

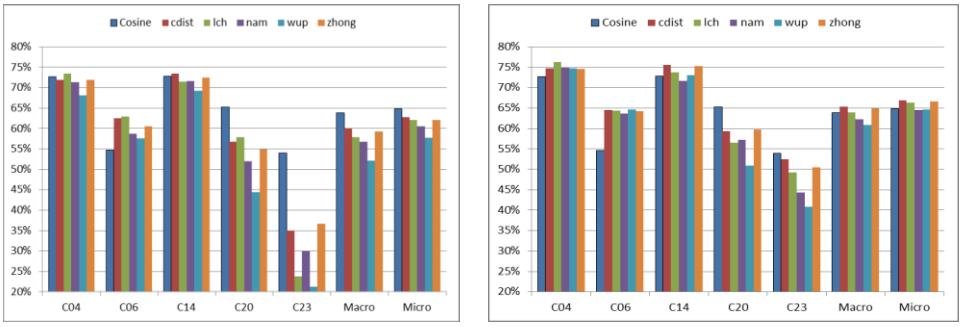

Fig. 2a. (left)Results of applying Rocchio (Cosine) and SemIDF measure- **Fig. 2b.** (right) Results of applying Rocchio (Cosine) and SemTFIDF measure on Ohsumed using F1-measure

Using *cdist*, *lch* or *zhong*, the increase in F1-Measure at class level increased the MacroAveraged F1-Measure. This approach has no impact on the weighting scheme which makes it less sensitive than others of different ranges of values retuned by these measures. Rocchio with this measure gave best results by using *cdist* as a semantic similarity measure; this resulted in a MacroAveraged F1-Measure of (65.32%) (see fig. 2.b). Note that *cdist* returns low values in the range [0, 1].

Discussion. Table 2 illustrates the difference between both text-to-text semantic similarity measures and Cosine. Both measures use cdist with which we obtained best classification results at macro level. SemTFIDF measure outperforms SemIDF measure and Cosine at the macro level. Moreover, it outperforms SemIDF measure for all the class. In fact, the SemTFIDF measure takes into account the TF/IDF weighting model in assessing similarities between a document and a centroid. Thus, it is essential to an aggregation function to take into account language and text statistics in assessing similarities. More precisely, first all semantic similarity measures improved Rocchio's performance for the class C06. Nevertheless, only three cases using our SemTFIDF measure improved results at MacroAveraged level. Best overall performance occurred with Rocchio and cdist similarity measure with a MacroAveraged F1-Measure of (65.32%). Both similarity measures: wup and lch, improved the performance of Rocchio at class level.

Second, we distinguish two important points for developing Semantic Text-To-Text Similarity Measures. The first point is that these measures worked with the five similarity measures and especially with *cdist*, *lch* and *zhong*. This means that they are less sensitive to differences between the ranges of the values returned by these measures.

Category	***C04***	***C06***	***C14***	***C20***	***C23***	***Macro***	***Micro***
Cosine	72,65	54,68	72,88	65,20	53,96	63,87	64,81
SemTFIDF	**74,75 ***	**64,56 ***	**75,55 ***	59,31	52,45	**65,32 ***	**66,91 ***
SemIDF	71,90	**62,56 ***	**73,46**	56,74	35,07	59,95	62,74

Table 2. Comparison, using F1-measure, between Cosine using TFIDF, SemIDF and SemTFIDF measures (* for significant increases according to McNemar test)

Third, least populated classes like (C06) are challenging for classification technique as compared to other classes for which the classification model is much easier to learn. However, Text-To-Text Semantic Similarity Measures helped the classifier distinguish this class with a maximum gain reaching (18.13%) in the case of our measure using lch. Similar to our observations after applying conceptualization, the class "C06" is among the least populated classes as compared to others and so using Text-To-Text Semantic Similarity Measures might result in a better identification of this class which led to better results.

6 Conclusion

In this work, we proposed a new text-to-text semantic similarity measure based on TF/IDF and we evaluated it as a prediction criterion for supervised text classification using Rocchio. We tested this new measure and compared it with another text semantic similarity measure based on IDF proposed in the literature, along the Cosine classical similarity measure that are usually used with BOW representation model. We tested these measures in the biomedical domain on the Ohsumed corpus, using domain specific knowledge base UMLS®.

According to our experimental results, it appears relevant to use text-to-text semantic similarity measures for prediction in centroïd-based classification as it modifies the behavior of the classifier and can improve its effectiveness, particularly with the new TF/IDF based text-to-text similarity measure that we propose. However, resulting performance is dependent on the semantic similarity measure used in assessing similarities between concepts and the aggregation function used in prediction. Consequently, it necessary to develop text-to-text semantic similarity measures, those are adapted to the application context.

Finally, we assume that semantic similarities are more adequate than classical similarities like Cosine in comparing texts represented as BOCs. In other words, we recommend using semantic similarities when concepts are used as features in the vector space model. As for future work, we intend to evaluate other factors that may influence the performance of our measure. In addition, we intend to evaluate its influence on other tasks related to information retrieval such as question answering and centroïd-based clustering.

7 References

1. Bloehdorn, S. and A. Hotho, Boosting for text classification with semantic features, in Proceedings of the 6th international conference on Knowledge Discovery on the Web: advances in Web Mining and Web Usage Analysis2006, Springer-Verlag: Seattle, WA. p. 149-166.
2. Salton, G., The SMART Retrieval System-Experiments in Automatic Document Processing1971: Prentice-Hall, Inc.
3. Albitar, S., S. Fournier, and B. Espinasse, The Impact of Conceptualization on Text Classification, in Proceedings of the 13th international conference on Web Information Systems Engineering2012, Springer-Verlag: Paphos, Cyprus. p. 326-339.
4. Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. Journal of Machine Learning Research, 2003. **3**: p. 993-1022.
5. Bloehdorn, S. and A. Moschitti, Combined syntactic and semantic Kernels for text classification, in Proceedings of the 29th European conference on IR research2007, Springer-Verlag: Rome, Italy. p. 307-318.
6. Albitar, S., S. Fournier, and B. Espinasse, Conceptualization Effects on MEDLINE Documents Classification Using Rocchio Method, in Web Intelligence2012. p. 462-466.
7. Hotho, A., S. Staab, and G. Stumme, Text clustering based on background knowledge, 2003.
8. Guisse, A., K. Khelif, and M. Collard. PatClust : une plateforme pour la classification sémantique des brevets. in Conférence d'Ingénierie des connaissances. 2009. Hammamet, Tunisie.
9. Huang, L., et al., Learning a concept-based document similarity measure. J. Am. Soc. Inf. Sci. Technol., 2012. **63**(8): p. 1593-1608.
10. Peng, X. and B. Choi, Document classifications based on word semantic hierarchies, in International Conference on Artificial Intelligence and Applications (AIA'05}2005. p. 362--367.

11. Wang, P., et al., Improving Text Classification by Using Encyclopedia Knowledge, in Proceedings of the 2007 Seventh IEEE International Conference on Data Mining2007, IEEE Computer Society. p. 332-341.
12. Al-Mubaid, H. 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. 2006.
13. Rada, R., et al., Development and application of a metric on semantic nets. Systems, Man and Cybernetics, IEEE Transactions on, 1989. **19**(1): p. 17-30.
14. Azuaje, F., H. Wang, and O. Bodenreider. Ontology-driven similarity approaches to supporting gene functional assessment. in Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies. 2005.
15. Mihalcea, R., 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 12006, AAAI Press: Boston, Massachusetts. p. 775-780.
16. Mohler, M. and R. Mihalcea. Text-to-text semantic similarity for automatic short answer grading. in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. 2009. Athens, Greece: Association for Computational Linguistics.
17. Dolan, B., C. Quirk, and C. Brockett, Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources, in Proceedings of the 20th international conference on Computational Linguistics2004, Association for Computational Linguistics: Geneva, Switzerland. p. 350.
18. Hersh, W., et al. OHSUMED: an interactive retrieval evaluation and new large test collection for research. in 17th annual international ACM SIGIR conference on Research and development in information retrieval. 1994. Dublin, Ireland: Springer-Verlag New York, Inc.
19. Aronson, A.R. and F.M. Lang, An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc, 2010. **17**(3): p. 229-236.
20. Caviedes, J.E. and J.J. Cimino, Towards the development of a conceptual distance metric for the UMLS. J. of Biomedical Informatics, 2004. **37**(2): p. 77-85.
21. Wu, Z. and M. Palmer, Verbs semantics and lexical selection, in Proceedings of the 32nd annual meeting on Association for Computational Linguistics1994, Association for Computational Linguistics: Las Cruces, New Mexico. p. 133-138.
22. Leacock, C. and M. Chodorow, Combining Local Context and WordNet Similarity for Word Sense Identification, in WordNet: An Electronic Lexical Database (Language, Speech, and Communication), C. Fellbaum, Editor 1998, The MIT Press. p. 265-283.
23. Zhong, J., et al., Conceptual Graph Matching for Semantic Search, in Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces2002, Springer-Verlag. p. 92-196.

Sebastiani, F., Text Categorization, in Encyclopedia of Database Technologies and Applications2005, Idea Group. p. 683-687.