Semantic Enrichments in Text Supervised Classification: Application to Medical Domain

Shereen Albitar – Bernard Espinasse – Sébastien Fournier

Aix-Marseille University – LSIS CNRS 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 supervised text classification can improve its effectiveness especially in specific domains. Most state of the art works use concepts as an alternative to words in order to transform the classical bag of words (BOW) into a Bag of concepts (BOC). This transformation is done through conceptualization task. Furthermore, the resulting BOC can be enriched using other related concepts from semantic resources. This enrichment may enhance classification effectiveness as well. This paper focuses on two strategies for semantic enrichment of conceptualized text representation. The first one is based on semantic kernel method while the second one is based on enriching vectors method. These two semantic enrichment strategies are evaluated through experiments using Rocchio as the supervised classification method in the medical domain, using UMLS ontology and Ohsumed corpus.

1. Introduction

Supervised text classification is currently a challenging research topic, particularly in areas such as information retrieval, recommendation, personalization, user profiles etc. The most popular text classification methods are Naïve Bayes Classifier (NB), Support Vector Machines (SVMs), Rocchio, and K-Nearest Neighbors (kNN). These methods, using BOW for text representation, suffer the lack of semantics in text representation and in the rest of the classification process; they ignore all semantics included in the original text that can be deployed in text classification. Nevertheless, it's possible to replace, in these methods, the classical BOW by BOC through "conceptualization" that enriches document representation model using semantic resources (Hotho *et al*., 2003; Ferretti *et al*., 2008). Many works argued that the use of semantics in text classification can enhance its effectiveness especially in specific domains (Bloehdorn et Hotho, 2006; Aseervatham and Bennani, 2009).

In this work, according to experiments, we try to estimate the impact of two semantic enrichment strategies on supervised classification of conceptualized text. These experiments are realized in the biomedical domain on the corpus Ohsumed, using the well-known Unified Medical Language System (UMLS) as knowledge base and Rocchio as supervised classification method.

In section 2 we first present a general conceptual framework to integrate semantics in supervised text classification with two strategies for semantic enrichment of the BOC text representation. In section 3 we briefly present semantic resources, corpus, Rocchio and tools used in this research. In section 4 we present the first strategy of semantic enrichment, based on *Semantic Kernel* method, and the results obtained. In section 5 we present the second enrichment strategy, based on *Enriching Vectors* method, and the results obtained. Finally, we conclude with an assessment of our work, followed by different research perspectives.

2. Supervised Text Classification with Semantic Enrichment

According to the literature, many works propose approaches involving semantics in text classification at different levels, arguing the utility of semantics in text representation (Aseervatham and Bennani, 2009; Séaghdha, 2009). Most of these works transformed the classical "bag of words" (BOW) representing the text in the Vector Space Model (VSM) into a "bag of concepts" (BOC) choosing concepts as an alternative feature to words (Bloehdorn and Hotho, 2006; Huang *et al*., 2012). Some works use semantic similarity between concepts as well as concepts in text classification in *representation enrichment* or *prediction* levels. Three major approaches are distinguished for representation enrichment: (i) *Semantic Kernels* - usually deployed with SVM classifiers (Aseervatham and Bennani, 2009; Séaghdha, 2009; Wang and Domeniconi, 2008), (ii) *Generalization* (Bloehdorn

and Hotho, 2006), and (iii) *Enriching Vectors* (Huang *et al*., 2012). Nevertheless, authors in (Bloehdorn and Hotho, 2006) conclude that applying generalization in domain specific tasks causes performance deterioration. Thus, this work is focused on *Semantic Kernels* and *Enriching Vectors* only.

A conceptual framework summarizing different approaches that involve semantics in the process of supervised text classification is illustrated in the figure 1. Semantics may be involved in different steps of the classification process: conceptualization, indexing and before or after training. Conceptualization is the process of finding a match or a relevant concept in a semantic resource that conveys the meaning of a word or multiple words from text. Concepts covering a text document constitute its semantic vector that can represent the document as a BOC. Two semantic enrichments are possible, one before and one after training phase.

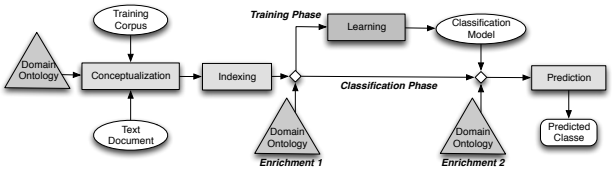

Fig. 1. A conceptual framework to integrate semantics in supervised text classification

In this work, we intend to investigate these two *enrichment* strategies and apply them in the medical domain in order to assess their influence on supervised text classification. Concerning the conceptualization step, we choose concepts as alternative feature to words in the classical BOW, for more details see (Albitar *et al*., 2012). Thus, we involve semantic knowledge in indexing by using concept in text representation.

3. Resources and Tools used

The various resources and tools used in this research are:

Unified Medical Language System (UMLS) was developed in order to model the language of biomedicine and health. It organizes concepts of the various source vocabularies (like MeSH, SNOMED-CT, etc.) according to their senses grouping common concepts together. We choose to conceptualize text with concepts from SNOMED-CT exclusively.

Ohsumed corpus is composed of abstracts of biomedical articles of the year 1991 retrieved from the Medline database indexed using MeSH (Medical Subject Headings). The first 20000 documents of this database were selected and categorized using 23 sub-concepts of the MeSH concept "Disease". The corpus is divided into Training and Test sets for experiments. In this work class centroïds are calculated for each of the most frequent classes listed in Table 1.

MetaMap Tool. In addition to the UMLS semantic resources, many tools are developed and provided in order to facilitate deploying these sources for medical information system developers. In this work we use MetaMap that improves biomedical text retrieval using UMLS Metathesaurus.

UMLS::Similarity Tool is a Perl module that assesses the semantic similarity between concepts in UMLS. The UMLS-Similarity-1.33 version used in this work considers nine different semantic similarity measures.

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

Table 1. Ohsumed Corpus

Three major families of semantic similarity measures are generally distinguished: *Ontology-based measures, Information Theoretic-based measures* and *Feature-based measures*. For this research we have chosen *Ontology-based measures* as they depend only on the structure of the ontology. Their simplicity is the origin of its demonstrated efficiency in different application domains where semantic similarity is used (Han and Karypis, 2000). We choose five ontology structure-based similarity measures to build semantic proximity matrices:

- *cdist*: it counts the number of edges between the compared concepts (Caviedes *et al*., 2004). Its range is between zero and twice the depth of the taxonomy
- *wup*: is twice the depth of the concepts' most specific common abstraction (msca) divided by the product of the depths of the concepts (Wu and Palmer, 1994). Its range is between zero and one.
- *lch*: is the negative log of the shortest path between two concepts divided by twice the total depth of the ontology (Leacock and Chodorow, 1998). Its range is unbounded and depends on the depth of the taxonomy
- *zhong*: is the sum of the difference between the milestone of the msca and each of the concepts (Zhong *et al*., 2002). The millstone is a calculated factor and is related to the specificity of concepts. Its range is between zero and one.
- *nam*: is the log of a formula of the shortest distance between the two concepts and the depth of the taxonomy minus the depth of the concepts msca (Al-Mubaid and Nguyen, 2006). Its range depends on the depth of the taxonomy.

Semantic Proximity matrix is a square matrix in which each cell is a measure of similarity between elements to

which row and column correspond. Semantic similarity is used in involving semantic information through enriching representation. Using ontology, we can assess semantic similarity between the concepts of the vocabulary pair-to-pair. To perform a semantic enrichment, we deploy a *semantic proximity matrix* built using semantic similarities between concepts of the BOC pair-to-pair.

In *Rocchio Method* [10], each class is represented by a centroid vector. The learned centroïds represent a classification model that summarizes the characteristics that occur in training documents. During exploitation phase, each test document is compared to classes' centroïds using similarity measures. 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 classification model. Most of other traditional classification methods, such as SVM and NB, allow the integration of semantics essentially in text representation.

4. Semantic Enrichment before Training using Semantic Kernel Method

In this section, we present a first strategy of semantic enrichment of the conceptualized text vector-based representations before training based on semantic Kernel method. This method is applied on vectors representing the training corpus and the test documents after indexing and before the Rocchio training phase. This enrichment is possible via a *semantic proximity matrix* built using semantic similarities between concepts of the BOC pair-to-pair, resulting from previous conceptualization. First we introduce the *Semantic Kernel* method using the *Semantic Proximity matrix*, and then, we present experimental process and results on Ohsumed using SNOMED-CT.

Semantic kernels Method

In general, semantic kernels are used with SVMs in order to transform the feature space into a BOC in which training examples are linearly separable. Different research works deployed general-purpose semantic resources in building their semantic kernels like WordNet (Séaghdha, 2009), Wikipedia (Wang and Domeniconi, 2008) or domain specific ontologies like UMLS for the biomedical domain (Bloehdorn and Moschitti, 2007).

To apply the semantic kernel method for enriching vectors of a conceptualized document (using a complete conceptualization strategy), first indexing builds the vector representing the text document as a BOC. Then, the system applies the *Semantic Kernel* method using a proximity matrix to the vector in order to enrich it with similar concepts. After applying semantic kernel to the vectors representing documents in the BOC model, resulting vectors are in general less sparse which might help Rocchio learn the classification model and predict classes of new documents.

Experimental Process

In order to assess the effect of *Semantic Kernels* on the process of text classification using Rocchio, we use the experimental platform illustrated in Figure 2. Conceptualization step is realized with MetaMap tool on text before indexing step. Training corpus and each new document are enriched using the *semantic proximity matrix*. Five different *semantic proximity matrices* are built using each of the five previous semantic similarity measures. Vectors resulting of applying *Semantic Kernel* to the training corpus documents are the input to the training step in order to learn the classification model. Enriched index of test documents is the input to prediction step where Rocchio compares the centroïds with the enriched index of each new document in order to predict its class.

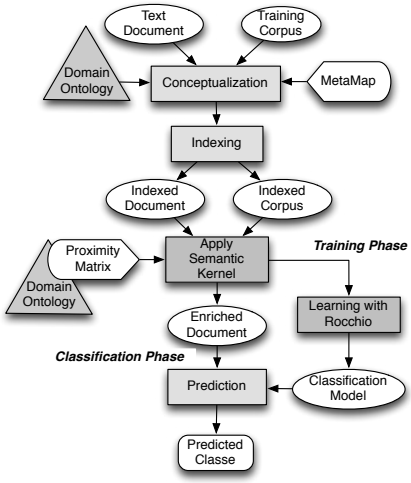

Fig. 2. Semantic Enrichment using Semantic Kernels

In experiments, the number of similar concepts involved in enriching text representation can be limited. We vary this parameter from 1 to 10 in order to evaluate its effect on the process of classification. Learning is performed 50 times: once for each of the proximity matrices and once for each value of the parameter related to the number of the similar concepts used in the enrichment. Rocchio uses each of the preceding models with each of its 5 variants (*Cosine, Jaccard, KullbackLeibler, Levenshtein, Pearson*) resulting in 5*5*10=250 executions.

Experimental Results

The MacroAveraged F1-measure resulting from these executions that are related to each semantic similarity measure are grouped together in the five graphics of Figure 3 in order to analyze the impact of the number of similar concepts used in enrichment on the effectiveness of the five variants of Rocchio.

According to observations, two variants of Rocchio showed very similar behavior: *Cosine* and *Pearson*. In fact,

Pearson is considered as a centered *Cosine* as all vectors are centered before assessing their similarities. As for *Jaccard*, we noticed important decrease in F1-Measure; this is due to the fact that *Jaccard* depends on commonalities, which are generally modified after enrichment. Results using *KullbackLeibler* showed similar behavior to other variants, except for the case that used *nam* as the semantic similarity measure. In experiments using *nam* semantic similarity measure, all variants demonstrated peaks and irregular decrease in the curves. This is due to the particular range of values that the measure *nam* returns and also to the relatively slight differences among similarities of different pairs of concepts. Finally, we report that *zhong* that has the maximum correlation coefficient with expert ratings showed the minimum decrease in F1-Measure as compared with the four other semantic similarity measures.

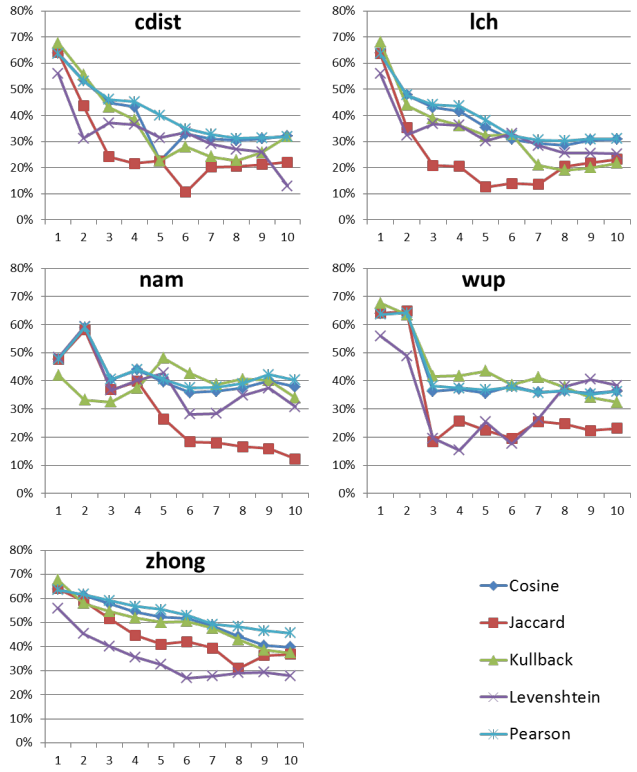

Fig. 3. MacroAveraged F1-measure obtained applying Semantic Kernels

In all experiments, enriching representation with the five to the seven most similar concepts, the effectiveness of all classifiers deteriorates significantly. This is due to the fact that Rocchio is dependent on text statistics and that applying Semantic Kernels introduced noise to the representation model. This had harmful effect on classification results according to our previous observations. Moreover, adding more concepts to the model increased in some cases the MacroAveraged F1-Measure. Taking a closer look at class level, classifier in such cases declined one, two and sometimes four classes in favour of the rest; this justifies the increase at Macro level.

In conclusion, results showed significant deterioration in classification effectiveness after applying *Semantic Kernel*, this means that this approach is not helpful to Rocchio in classifying Ohsumed documents whereas it was reported quite useful using SVMs (Wang and Domeniconi, 2008) This is quite similar to the conclusion of authors in (Bloehdorn and Moschitti, 2007) when applying Adaboost to Ohsumed corpus after enriching text representation through generalization. Enriching domain specific text representation with related concepts needs much more investigation, which leads us to next experiments using another approach for enriching text representation.

5. Semantic Enrichment after Training using Enriching Vectors Method

In this section we present a second strategy of semantic enrichment, after training phase and before prediction step, based on *Enriching Vectors* method. This method enriches the BOC text representation of the classification model and test documents before prediction using proximity matrix as well. In the following sub-sections, first we introduce the *Enriching Vectors* method, then we present experimental process and results with SNOMED-CT, Ohsumed and Rocchio.

Enriching Vectors Method

Authors in (Huang *et al*., 2012) proposed this method and applied it in the context of clustering using K-means and classification using kNN. In order to compare two documents, authors apply this method to the vectors that represent these documents and then apply a classical text-to-text similarity measure like Cosine. This method demonstrated a better correlation with human judgment as compared to applying the classical similarity measure on the original vectors.

Classical similarity measures, like Cosine, depend on lexical matching in comparing text documents represented in the vector space model. In fact, these measures take into consideration the shared features among the compared vectors neglecting any other similarities such as semantic similarity among the unshared features. In other words, if two texts do not share the same words but use synonyms, they are presumed dissimilar. We previously identified this drawback of the classical BOW (Albitar *et al*., 2012).

In order to go beyond lexical matching, we intend to apply *Enriching Vectors* to each pair of vectors before comparison: each of the compared vectors enriches the other vector using its exclusive features. Given two documents A, B represented using a vocabulary of several concepts. We note that a feature is exclusive for B if it is

mapped to B's text only and respectively an exclusive feature for A is mapped to A's text only. The main goal of this approach is to give an appropriate weight for each exclusive feature of A in B and vice versa. These weights are estimated using weights of other features of the document and semantic similarity between these features and the missing feature.

To apply Enriching vectors on two text documents that are conceptualized using a complete conceptualization strategy, we have to follow different steps. First, indexing step extracts conceptual features from the documents and transforms them into vectors as BOCs. Then, by means of a semantic proximity matrix (using a particular semantic similarity measure), both vectors are mutually enriched as a second step. Finally, we compare the enriched vectors using a classical similarity measure. The resulting similarity takes into consideration similar concepts as well as common concepts.

Experimental Process

In order to assess the effect of *Enriching Vectors* on the process of text classification using Rocchio, we use the experimental platform illustrated in Figure 4. Similar to the previous platform, this platform uses Rocchio for training and prediction as the classification technique. For conceptualization, same configurations are used in this platform. For enriching step, the test document vector is compared to each of the centroïds learned during training.

Before applying one of the classical similarity measures, the vector of the document and the vector of the centroïd are mutually enriched using the semantic proximity matrix of one of the five semantic similarity measures. After this enrichment, vectors are less sparse and share more common features (concepts). Finally, prediction step applies one of the classical similarity measures of the VSM and evaluates the results.

In these experiments, the platform executes learning once. This means that Rocchio learns one classification model or an ensemble of centroïds. As for classification, Rocchio uses this model with each of its 5 variants and each of the 5 semantic proximity matrices.

Experimental Results

The detailed results from the executions that are related to each similarity measure are grouped together to analyze the impact of Enriching Vectors on the effectiveness of the five variants of Rocchio. We perform an experimental study on the effects of *Enriching Vectors* on Rocchio's performance using five semantic similarity measures (*cdist, lch, nam, wup, zhong*) on concepts of SNOMED-CT pair-to-pair. Tests were realized on completely conceptualized Ohsumed corpus using Ids of the best mappings (Albitar *et al*., 2012). The results of the 5 variants are illustrated in the figure 5. Experimental results lead us to the following points.

First of all, in all cases, using the semantic similarities *lch* and *wup* caused deterioration in Rocchio's performance while other similarity measures showed some improvements. Note that the only aspect that *cdist*, *nam*, and *zhong* share is the relatively low values of semantic similarity they return as compared to both *lch* and *wup* which justifies their different influence of text representation. Best overall performance was obtained using with *Cosine* and *zhong* Rocchio variants with a MacroAveraged F1-Measure of (64.33%). This value is higher than the one reported in (Huang and al., 2012) where authors tested Enriching Vectors on a small corpora retrieved from Ohsumed using kNN classifier.

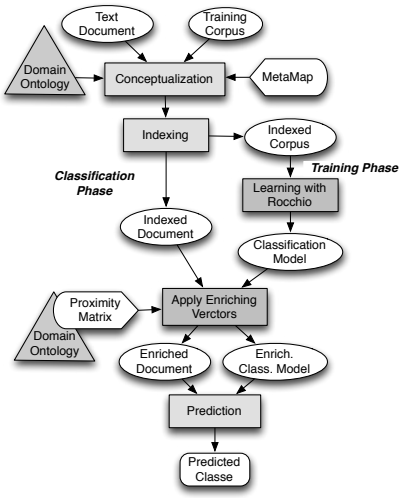

Fig. 4. Semantic Enrichment using Enriching Vectors

Second, we distinguish two groups of Rocchio variants according to their performance after applying *Enriching Vectors*: first group contains *Cosine, Jaccard and Pearson* and the second one contains *KullbackLeibler and Levenshtein*. The main difference between these groups is that the first one assesses similarity among vectors using their commonalities whereas the second one depends on their differences in order to assess their similarities. In general, *Enriching Vectors* aims to reduce the sparseness of text representation; this seems to help the first group in assessing similarities. On the contrary, this enrichment seems to be harmful to assessing similarities as the differences between vectors are modified after enrichment.

Third, when the system performance using a specific method has a low F1-measure value, as it is the case for the class (C23), *Enriching Vectors* can improve this value with a maximum gain reaching (9.45%) in the case of *Jaccard* Rocchio variant. Similar to our observations after applying conceptualization, the class "C23" is very large compared to others and thus enriching class representation with similar concepts might result in a better identification of this class, which led to better results.

Finally, it seems beneficial to Rocchio-based classification to apply *Enriching Vectors* before prediction as it modifies the behavior of the classifier and can improve its effectiveness. However, resulting performance is dependent on the semantic similarity measure used in enrichment and also on the similarity measure used for prediction. Consequently, it is necessary to verify experimentally in order to check whether *Enriching Vectors* is useful in a particular context.

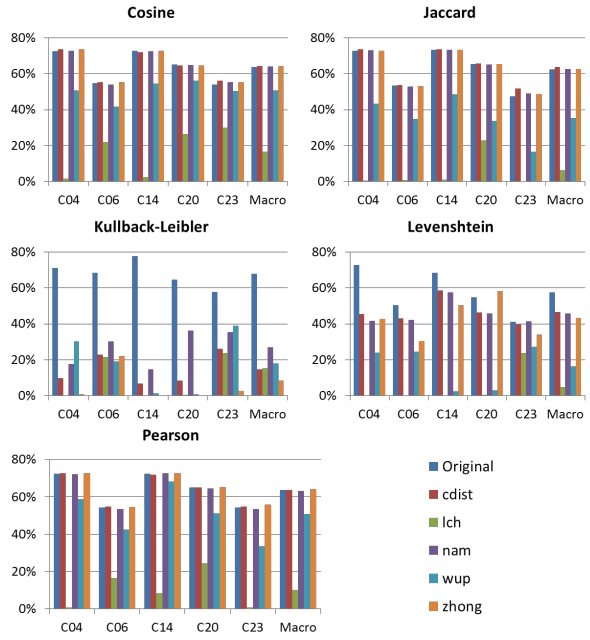

Fig. 5. Number of improved classes after applying Enriching Vectors

5. Conclusion

According to experiments in the biomedical domain on the corpus Ohsumed using UMLS, and Rocchio supervised classification method; we evaluated the impact of two semantic enrichment strategies on supervised text classification in the concept space.

The first enrichment strategy, before training, is based on *Semantic Kernels* method. Obtained results showed deterioration in the performance of Rocchio and its variants after applying Semantic Kernels on vectors that represent corpus documents. Thus, Semantic Kernels seem to introduce noise to text representation and weakens its capability to distinguish classes.

The second enrichment strategy, after training and before prediction, is based on *Enriching Vectors* method. We reported better results than those obtained without enrichment as well as those obtained after applying Semantic Kernels. Nevertheless, this improvement depends on both the semantic similarity measure used in enrichment and the similarity measure used in prediction.

In future works, we intend to test other families of semantic similarity measures like IC-based or feature based measures on Ohsumed and other medical corpora like TREC genomics or i2b2.

References

Hotho, A., Staab, S., and Stumme, G. 2003. Text clustering based on background knowledge. Institute AIFB, Universität Karlsruhe.

Ferretti, E., Errecalde, M., and Rosso, P. 2008. Does Semantic Information Help in the Text Categorization Task? Journal of Intelligent Systems 17, 91–107.

Bloehdorn, S. and Hotho, A. 2006. Boosting for text classification with semantic features. 6th intern. conference on Knowledge Discovery on the Web, Seattle, WA.

Aseervatham, S., and Bennani, Y. 2009. Semi-structured document categorization with a semantic kernel. Pattern Recogn., 42(9), 2067-2076.

Bloehdorn, S. and Moschitti, A. 2007. Combined syntactic and semantic Kernels for text classification. 29th European conference on IR research, Rome, Italy.

Séaghdha, D. Ó. 2009. Semantic classification with WordNet kernels. Human Language Technologies, Companion Volume: Short Papers, Boulder, Colorado.

Huang, L., Milne, D., Frank, E., and Witten, I. H. 2012. Learning a concept-based document similarity measure. J. Am. Soc. Inf. Sci. Technol., 63(8), 1593-1608.

Wang, P. and Domeniconi, 2008. C. Building semantic kernels for text classification using wikipedia. 14th ACM SIGKDD, Las Vegas, Nevada, USA.

Han E.-H. and Karypis G. 2000. Centroid-Based Document Classification: Analysis and Experimental Results. Proc. 4th European Conference on Principles of Data Mining and Knowledge Discovery.

Sanchez, D., Batet, M., Isern, D., and Valls, A. 2012. Ontology-based semantic similarity: A new feature-based approach. Expert Syst. Appl., 39(9), 7718-7728. doi: 10.1016/j.eswa.2012.01.082.

Caviedes, J. E. and Cimino, J. J. 2004. Towards the development of a conceptual distance metric for the UMLS. J. of Biomedical Informatics, 37(2), 77-85.

Wu, Z. and Palmer, M. 1994. Verbs semantics and lexical selection. Proc. of the 32nd annual meeting on Association for Computational Linguistics, Las Cruces, New Mexico.

Leacock, C. and Chodorow, M. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. In C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, pp. 265-283, The MIT Press.

Zhong, J., Zhu, H., Li, J., and Yu, Y. 2002. Conceptual Graph Matching for Semantic Search. Proc. of the 10th International Conference on Conceptual Structures: Integration and Interfaces.

Al-Mubaid, H. and Nguyen, H. A 2006. Cluster-Based Approach for Semantic Similarity in the Biomedical Domain. Proc. of 28th IEEE-EMBS'06.

Albitar, S., Fournier, S., and Espinasse, B. 2012. The impact of conceptualization on text classification. Proc. of the 13th International conference on Web Information Systems Engineering (WISE 2012), Paphos, Cyprus.