International Journal of Healthcare Information Systems and Informatics
Volume 16 • Issue 4 • October-December 2021

An Efficient Method for Biomedical Word Sense Disambiguation Based on Web-Kernel Similarity

Mohammed Rais, LISA, Department of Electrical and Computer Engineering, ENSA, Sidi Mohamed Ben Abdellah University, Morocco

Mohammed Bekkali, LISA, Department of Electrical and Computer Engineering, ENSA, Sidi Mohamed Ben Abdellah University, Morocco

Abdelmonaime Lachkar, Systems and Data Engineering Team, Department of Information System and Communication, *ENSA,* Abdelmalek Essaâdi University, Morocco

ABSTRACT

Searching for the best sense for a polysemous word remains one of the greatest challenges in the representation of biomedical text. To this end, word sense disambiguation (WSD) algorithms mostly rely on an external source of knowledge, like a thesaurus or ontology, for automatically selecting the proper concept of an ambiguous term in a given window of context using semantic similarity and relatedness measures. In this paper, the authors propose a web-based kernel function for measuring the semantic relatedness between concepts to disambiguate an expression versus multiple possible concepts. This measure uses the large volume of documents returned by PubMed search engine to determine the greater context for a biomedical short text through a new term weighting scheme based on rough set theory (RST). To illustrate the efficiency of our proposed method, they evaluate a WSD algorithm based on this measure on a biomedical dataset (MSH-WSD) that contains 203 ambiguous terms and acronyms. The obtained results demonstrate promising improvements.

KEYWORDS

Biomedical Word Sense Disambiguation, Conceptualization, Context Concept, MSH-WSD, Rough Set Theory, Short Text Similarity

1. INTRODUCTION

Recently, information amount in the medical field has grown exponentially with more than 23 million published citations listed in the MedLine database and available via PubMed. As a result, automatically finding high-precision and relevant information has become a challenging task. Many biomedical text-mining applications such as information retrieval, text categorization and machine translation aim to provide suitable solutions for this purpose. Note that, one of the major problems in these applications is the document's representation where we still limited only by the terms or words that occur in the document. The usual way of representing a text is the Bag of Words (BoW) representation, which simply look at the surface word forms and ignore all semantic or conceptual information in the text. Biomedical documents make these issues even more serious, due to their sparseness and lexical ambiguity.

DOI: 10.4018/IJHISI.20211001.oa9

To overcome the previous limitations of BoW representation, external semantic resources such as thesaurus or ontologies, may be used to swap BoW representation by a Bag of Concepts (BoC) one, which transforms the data representation into a shorter, more compact, and more predictive one. Concepts are unambiguous: each concept represents a unique meaning. Synonyms are mapped to the same concept while terms with multiple meanings are mapped to different concepts corresponding to their intended meanings. However, while searching polysemic term corresponding concepts in semantic resources, multiple matches are detected and introduce some ambiguities in final document representation. Therefore, assigning a term to a corresponding concept is a crucial step in the conceptualization process. Therefore, Word Sense Disambiguation (WSD) algorithms generally use the context information in order to assign a unique sense or concept to an ambiguous term.

WSD play a vital role in many biomedical text-mining application, the NLM (National Library of Medicine) concluded that lexical ambiguity in the Unified Medical Language System (UMLS) was the biggest challenge in indexing biomedical journals with concepts from the UMLS Metathesaurus (Laura Plaza et al., 2011). There are three main approaches for WSD: knowledge based approach which define explicit sense distinctions for assigning the correct sense of a word in context, supervised learning approach where Machine Learning techniques is used on a sense-annotated data set to classify the senses of the words, and unsupervised approach where the sense of a word will depend on those of neighboring words. In the latter approach, semantic similarity or relatedness measures are performed to quantify the semantic proximity between two concepts. Previous studies compare and evaluate the efficiency of semantic similarity and relatedness measure on Biomedical WSD (McInnes and Pedersen, 2013). Because of the short nature of the concept definition, we believe that exploiting short text similarity measure can improve the identification process of which the meaning of a word is used in a context.

In this paper, we propose a Web-based Kernel function for measuring the semantic relatedness between concepts to disambiguate an expression versus multiple possible concepts. In short, we compute the similarity between the context of the word to be mapped and the different corresponding concepts; the concept with the greatest similarity is the one to be chosen. This measure uses the large volume of documents returned by PubMed search engine to determine the greater context for a biomedical short text through a new term weighting scheme based on Rough Set Theory (RST) which is a mathematical tool to deal with vagueness and uncertainty (Pawlak, 1991).

To test the effectiveness of our proposed method, we built a WSD algorithm based context using the proposed similarity measure on the biomedical dataset (MSH-WSD) that contains 203 ambiguous terms and acronyms. The obtained results illustrate the efficiency of our proposition.

The remaining parts of this paper are organized as follows: we begin with a brief survey on related work about WSD algorithms and semantic similarity measure. Section 3 introduces our proposed WSD method, starting with the mathematical background of Rough Set Theory, our enhanced similarity measure named Web-RST, and how this measure can be used in the WSD process. Dataset and the evaluation system are presented in section 4. The obtained results are discussed in section 5. Finally, section 6 concludes this paper and presents some perspectives.

2. RELATED WORKS

There are many proposed approaches which attempt to address the problem of WSD. Existing methods can be classified into three categories: Supervised (Zhong and Ng, 2010), Unsupervised (Brody and Lapata, 2009) and Knowledge-Based methods (Navigli, 2011).

For supervised methods, training data must be created for each ambiguous word manually or automatically, which is infeasible on a large scale. For unsupervised methods do not use training data, but use the distributional characteristics of an outside corpus. For Knowledge-based methods, we use external semantic resources and possibly information from corpus. In this work, we focus on knowledge-based methods.

2.1 Knowledge-Based Methods for WSD

State of the art knowledge-based approaches rely on context information using similarity measures, Garla and Brandt (2012) have developed and evaluated a knowledge-based WSD method that uses semantic similarity measures derived from the Unified Medical Language System (UMLS), and integrated with MetaMap the clinical Text Analysis and Knowledge Extraction System on standard biomedical WSD datasets.

To exploit the effectiveness of similarity and relatedness measures extracted from biomedical resources, McInnes and Pedersen (2013) have developed a method named **SenseRealte**, which use a fixed window size and taking into consideration the distance weight on how far the terms in the context are from the target word. In the same context RAIS and LACHKAR (2016) have proposed a simple modified version of ***SenseRealte*** named ***NoDistanceSenserealte*** which simply ignores the distance, that is the terms in the context will have the same distance weight.

Because of the short definition of a concept, we believe that short text similarity can be useful in Biomedical WSD process. In the following sub-section, we present a brief state of the art concerning short text similarity measures.

2.2 Short Text Similarity Measure

Measuring short text similarity has become a crucial stage in the process of many text mining task because of its shortness and sparseness. As a result, it has been intensively studied. We consider three categories of methods: surface matching, corpus-based methods, and query-log methods.

2.2.1 Surface-Matching Methods

Given two text segments (x, y), the idea of surface matching methods is based on the number of words that occur in both text segments. This technique relies on the assumption that more similar texts share more of the same words, which is effective only if both texts are of sufficient lengths. Let's say X and Y are the sets of words in texts x and y. Common similarity measures presented in (Lund et al., 1995) are listed as follows:

$$\text{Mat}(x, y) = | X \cap Y | \tag{1}$$

$$\text{Jacc}(x, y) = | X \cap Y | / | X \cup Y | \tag{2}$$

$$\text{Dice}(x, y) = | X \cap Y | / (| X | \cap | Y |) \tag{3}$$

$$\text{Over}(x, y) = | X \cap Y | / \min (| X|, |Y|) \tag{4}$$

$$\text{Cos}(X, Y) = | X \cap Y | / \sqrt{} | X | * | Y | \tag{5}$$

2.2.2 Query Log Methods

Search engines like Google or PubMed process millions of search queries per day. The produced search query logs have become a great resource for measuring the similarity between short texts. This information is then used for search query suggestion generation, which is becoming more and more accurate and useful (Wen-tau and Christopher 2007).

2.2.3 Corpus-Based Methods

The corpus-based method depends on a large corpus. Once the method was proposed for an application domain, it can hardly be used in another domain. The Latent Semantic Analysis (LSA) (Landauer et al., 1998) and the Hyperspace Analog to Language (HAL) model (Burgess et al., 1998) are two well-known methods in corpus-based similarity. Sahami and Heilman (2006) have introduced a function for measuring the similarity between short text snippets by leveraging Web search results to provide greater context for the short texts. This measure can be improved, for example, by considering other weighting schemes.

In the following section of this paper, we will introduce an enhanced Web-Kernel similarity measure based on RST which is a mathematical tool to deal with vagueness and uncertainty (Pawlak 1991), and how this measure can be useful to improve of the WSD process.

3. PROPOSED METHOD FOR WSD

In this section, we present in detail our proposed method for WSD. First, we present the RST mathematical background; after that, we describe our enhanced similarity measure for short text based on RST. Then, we will explain how this measure can be used in the WSD process.

3.1 Rough Set Theory

In this section, we present the Rough Set Theory as a mathematical tool for imprecise and vague data, and its tolerance model to deal with textual information.

3.1.1 Generalized Approximation Spaces

Rough Set Theory, has been originally developed as a tool for data analysis and classification (Pawlak, 1991) (Komorowski et al., 1998). It has been successfully applied in various tasks, such as features selection/extraction, rule synthesis, and classification. The central point of RST is the notion of set approximation: any set in U (a non-empty set of object called the Universe) can be approximated by its lower and upper approximation. In order to define lower and upper approximation, we need to introduce an indiscernibility relation that could be any equivalence relation R (reflexive, symmetric, transitive). For two objects x, y ϵ U, if xRy then we say that x and y are indiscernible from each other. The indiscernibility relation R induces a complete partition of universe U into equivalent classes $[x]_R$, x ϵ U (Ngo Chi Lang 2003).

$$LR(X) = \{x \epsilon U: [x]_R \subseteq X\} \tag{6}$$

$$UR(X) = \{x \epsilon U: [x]_R \cap X \neq \Phi\} \tag{7}$$

Approximations can also be defined by mean of rough membership function. Given rough membership function $\mu X: U \rightarrow [0, 1]$ of a set $X \subseteq U$, the Rough approximation is defined as:

$$LR(X) = \{x \epsilon U: \mu X(x, X) = 1\} \tag{8}$$

$$UR(X) = \{x \epsilon U: \mu X(x, X) > 0\} \tag{9}$$

Note that, given rough membership function as:

$$\mu X(x, X) = \| [x]R \cap X \| / \| [x]R \| \tag{10}$$

RST is dedicated to any data type but when it comes with documents representation, we use its Tolerance Model described in the next section.

3.1.2 Tolerance Rough Set Model.

Let D = {d1, d2…, dn} be a set of documents and T= {t1, t2…, tm} set of index terms for D. with the adoption of the vector space model, each document di is represented by a weight vector {wi1, wi2…, wim} where wij denotes the weight of index term j in document i. The tolerance space is defined over a Universe of all index terms U = T = {t1, t2…, tm} (Zhang and Shuxuan, 2013).

Let fdi(ti) denotes the number of index terms ti in document di; fD(ti, tj) denotes the number of documents in D in which both index terms ti an tj occurs. The uncertainty function I with regards to threshold θ is defined as:

$$I\theta = \{tj \mid fD(ti, tj) \geq \theta\} \cup \{ti\} \tag{11}$$

Clearly, the above function satisfies conditions of being reflexive and symmetric. So Iθ(Ii) is the tolerance class of index term ti. Thus we can define the membership function μ for Ii ϵ T, X ⊆ T as (Jimeno-Yepes et al., 2011):

$$\mu X(ti, X) = v(I\theta (ti), X) = | I\theta (ti) \cap X | / | I\theta (ti) | \tag{12}$$

Finally, the lower and the upper approximation of any document di⊆ T can be determined as:

$$LR(di) = \{ti \in T: v(I\theta(ti), di) = 1\} \tag{13}$$

$$UR(di) = \{ti \in T: v(I\theta(ti), di) > 0\} \tag{14}$$

Each document is represented by a weight's vector where the original TF*IDF formula which combines the definitions of term frequency and inverse document frequency is replaced by the following formula:

$$W_{ij} = \begin{cases} (1 + \ln(f_{di}(t_i))) * \ln \dfrac{N}{f_D(tj)}, t_j \in d_i \\ \min t_k \in w_{ij} * \dfrac{\ln\left(\dfrac{N}{f_D(tj)}\right)}{1 + \ln\left(\dfrac{N}{f_D(tj)}\right)}; t_j \in U_R(d_i) / d_i \\ 0; t_j \notin U_R(d_i) \end{cases} \tag{15}$$

Where wij is the weight of the term j in document di. This formula ensures that each term occurring in the upper approximation of di but not in di has a weight smaller than the weight of any terms in di. Normalization by vector's length is applied to all weights of document vectors wij (Ngo Chi Lang 2003).

$$w_{ij} = \frac{w_{ij}}{\sqrt{\sum_{t_k \epsilon U_R(d_i)} \left(w_{ij}\right)^2}} \tag{16}$$

In the next section, we present our proposed similarity measure for short text based on Web-Kernel function combined with the Rough Set Theory.

3.2 WEB-RST Similarity Measure

Inspired by the Web-based kernel similarity measure introduced by (Sahami and Heilman, 2006) and by making some crucial changes; in this section we present our proposed method named Web-RST (Figure 1) for computing the similarity between a pair of short text down to the smallest detail.

Let x represent a short text, we calculate the query expansion denoted QERST(x). The detailed steps of the new query expansion method are described below:

1. Let Dn(x) be the set of top n documents returned by a search engine when using x as the query term (we use the services offered by PubMed e-utilities (Sayers, 2009)).
2. Each document di ϵ Dn(x) will be cleaned by removing stop words and special characters like (/, #, $, etc…) and tokenized. Then we generate the upper approximation U(di) using the formula (14).
3. For each document U(di), we construct the term vector vi, where each element is the weighted score wi,j of term tj, defined as with the formula (15).
4. Let C(x) be the centroid of the L2 normalized vectors vi: C(x) = vi / ||vi||
5. Let QERST(x) be the L2 normalization of the centroid C(x): QERST(x) = C(x) / ||C(x)||
6. Given a pair of short text segments x and y, the Web-Kernel similarity measure is QERST(x). QERST(y).

The Web-RST similarity process algorithm can be summarized and presented as bellow (Algorithm1).

In this section, we have been presented an efficient method to compute the similarity between Biomedical short text pair. This method can be used as a sense similarity measure between concepts as presented in the following section.

3.3 WSD Process

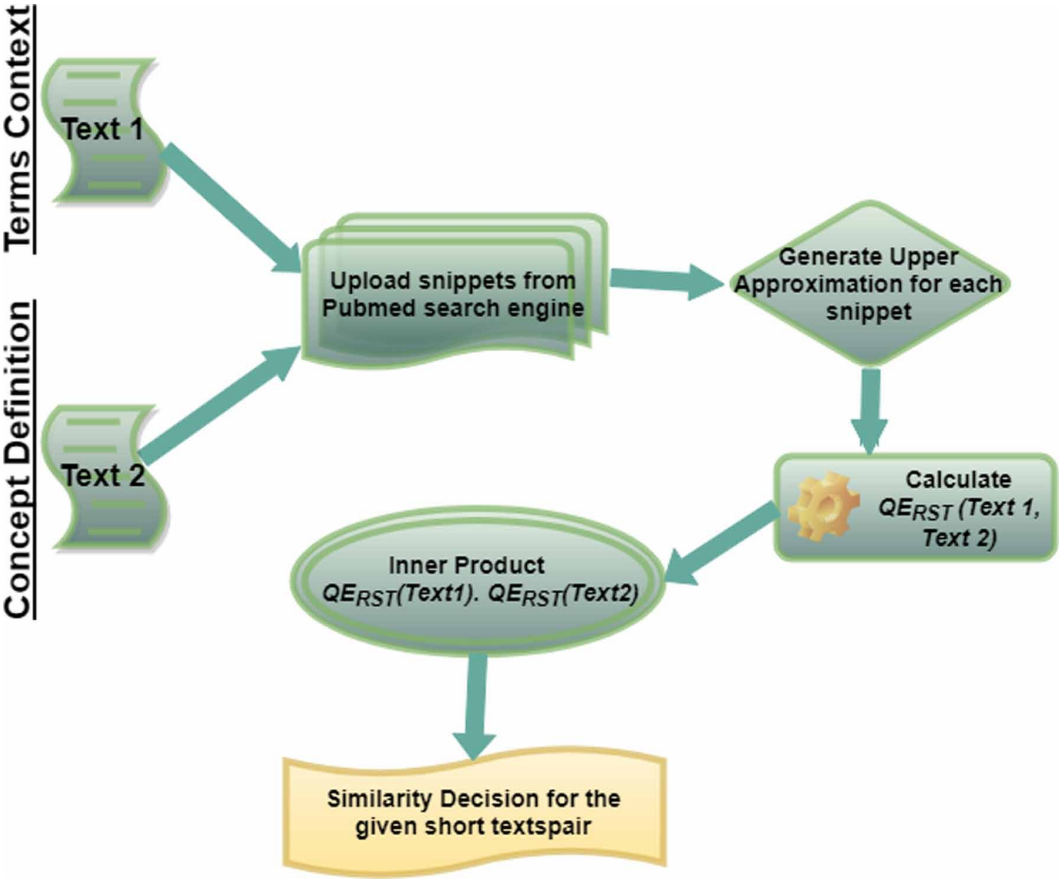
Previously, we have been presented an efficient method to compute the similarity between Biomedical short text pair, we believe that exploiting short text similarity measure can improve the identification process of which sense of a word is used in a context. This method can be used as a sense similarity measure between concepts.

The purpose of WSD step is to identify the concept that corresponds to word. To this end, we apply our Web-RST

similarity measure between a pair of short texts where the two short texts will be formed as follows: the first text will be formed by the term to be mapped with the neighboring words to consider its context and the second one will be the concept definition. We repeat this treatment for all candidate's concepts and the concept with the greatest similarity will be the chosen on.

The disambiguation process algorithm can be summarized and presented as bellow (Algorithm 2).

Figure 1. Flowchart of Similarity Computation Algorithm



4. EVALUATION SYSTEM

In order to evaluate our enhanced similarity measure, we WSD system was developed based on this measure. In this section, we provide the dataset and flowchart of the evaluation system.

DataSet

In our experiments, we have used the NLM's MSH-WSD dataset developed by Jimeno-Yepes et al. (2011). The dataset contains 106 ambiguous term, 88 ambiguous acronyms, and 9 that are a combination of both, for a total of 203 ambiguous words. Each instance containing the ambiguous word was assigned a CUI from the 2009AB version of the UMLS. For each ambiguous term/abbreviation, the data set contains a maximum of 100 instances per sense obtained from MEDLINE; totaling 37,888 ambiguity cases in 37,090 MEDLINE citations.

Flowchart System

As mentioned above, we have used MSH-WSD as corpora to evaluate web-kernel similarity. Figure 2, presents the flowchart of our evaluation system.

In this system, we proceed as follows:

Algorithm 1. double WebRSTSim(String $query_1$, String $query_2$)

```
Begin
  list<Snippet> sl1 = NTopResults(query1)
  list<Snippet> sl2 = NTopResults(query2)
  preProcess(sl1)
  preProcess(sl2)
  generateUpperApproximation(sl1)
  generateUpperApproximation(sl2)
  Compute the term weight vector vi for each document di in sl1
  Compute the term weight vector wi for each document di in sl2
  //Let C1 be the centroid of the L2 normalized vectors vi
  for each vi do
     C1 = vi / ||v1||2
  end for
  C1 = C1 / N
  //Let QE1 the query expansion of query1
  QE1 = C1 / ||C1||2
  for each wi do
     C2 = wi / ||w1||2
  end for
  C2 = C2 / N
  QE2 = C2 / ||C2||2
  double sim = QE1 . QE2 // . is the inner product between QE1 and QE2
  return sim
End
void generateUpperApproximation(list<document> dl)
Begin
  Initialise the TermSet s
  Initialise the co-occurrence matrix
  for each term ti in s do
     for each term tj in s do
        if occurTogether(ti, tj) > θ then
           addToToleranceClass(ti, tj)
        end if
     end for
  end for
  for each document d in dl do
     for each term tj in s do
        tct = toleranceClass(tj)
        cof = |tct Ç dl / |tct|
        if cof > 0 then
           addTermToupperApprox(d, tj)
        end if
     end for
  end for
end
```

- From the corpus, we begin by extracting the ambiguous term and Executing the Term Extractor module for the biomedical domain which attempts to extract biomedical terms in the context of the ambiguous word. Considering the ambiguous term: Malaria from MSH-WSD: "Antimicrobials in children admitted to hospital in <e>malaria</e> endemic areas", the output of extracting terms return 5 terms: Antimicrobials, Children, admitted hospital, malaria, areas.
- Creating an object Map containing the ambiguous term as key and the list of terms on the context as value, for example, the preceding sentence will be presented by Entry<malaria, List<Antimicrobials, Children, admitted hospital, areas>>.

Algorithm 2. WSDProcess(String term)

```
// term is a term to be mapped
Begin
List<Concept> conceptList = PubeEd(term)
String termQuery = term + the Neighboring Terms
Concept c = c0 //c0 is the first concept
String conceptDef = c0.getDefition()
double maxSim = WebRSTSim(termQuery, conceptDef)
for each Concept ci in conceptList do
  conceptDef = ci.getDefinition()
  double sim = webRSTSim(termQuery, conceptDef)
  if sim > maxSim then
    maxSim = sim c = ci
  end if
end for
return c
end
```

- Mapping Ambiguous term to their concepts. As result, the new map contains terms (ambiguous term and terms on the context) with theirs concepts: Entry<malaria [C0206255, C0024530], List<Antimicrobials, Children, admitted hospital, areas >>.
- Creating query text for each concept of the ambiguous term using their definitions derived from the UMLS Terminology Services (UTS), provide Web Services to search and retrieve UMLS data. Creating the second query text using terms in the context.
- Carrying out Web Kernel similarity (described in section 3) using queries from the last step. Witch return the similarity measure for each concept of the ambiguous term, from theses returned measures; we select the proper concept of the ambiguous term.
- Finally, executing the evaluation step to compare selected concepts using Web Kernel similarity method and concepts assigned by experts from MSH-WSD corpus.

Figure 2. Flowchart Web-Kernel similarity evaluation system

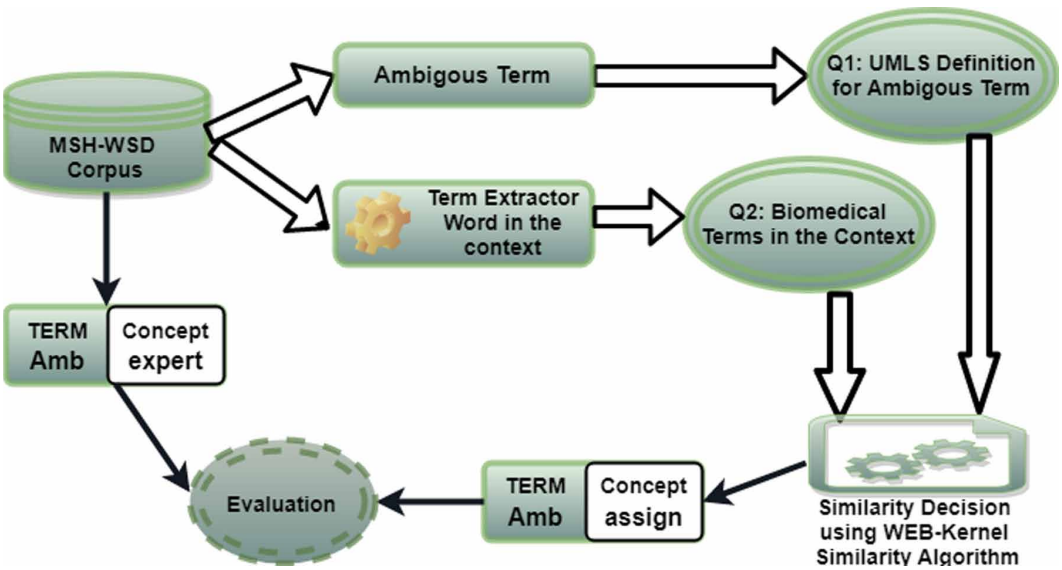
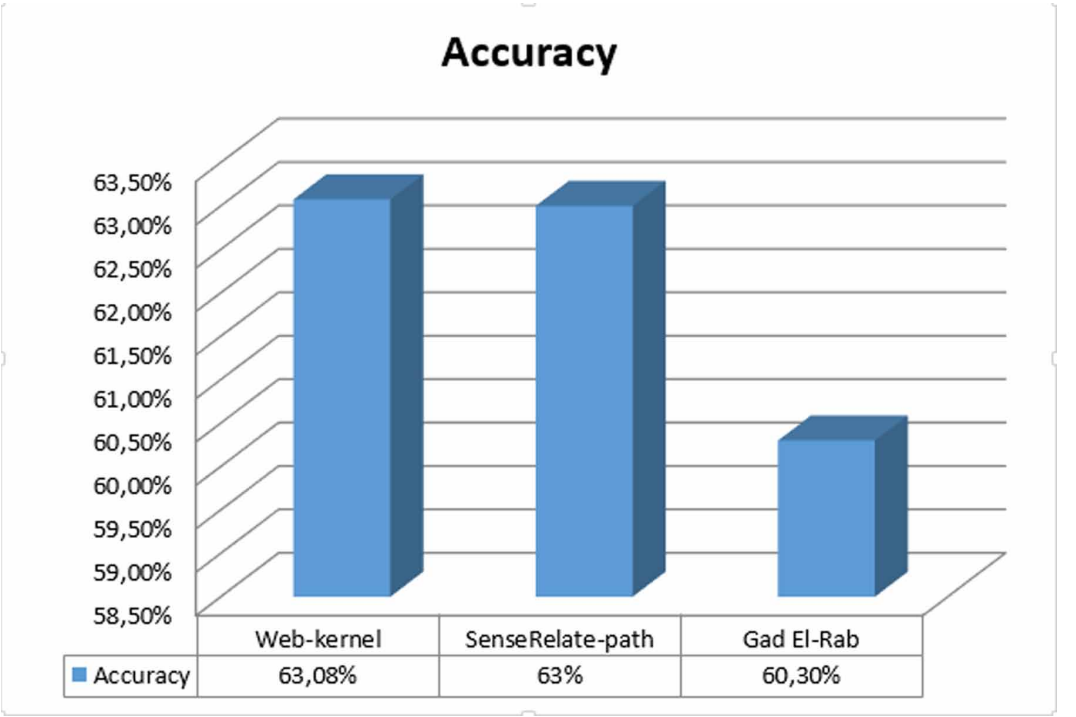


Figure 3. The result of Accuracy on MSH-WSD



5. RESULT AND DISCUSSION

Given the previously presented system, we have performed experiments in order to evaluate our WSD method based on WEB-RST similarity measure on MSH-WSD Dataset. Concerning the context information, we use a window size of 2 that means two terms surrounding the ambiguous term to the right and two terms to the left to determinate the appropriate concept.

We compare our WSD method with two other methods among the best known in the literature: The Bridget method named SenseRelate (McInnes and Pedersen, 2013) and the Gad El-Rab method (Gad El Rab et al., 2013).

Figure 3 presents the accuracy of our proposed method, SenseRelate using the path measure (path), and Gad El-Rab method.

As illustrated in figure 3, we observe that the accuracy of our WSD method and SenseRelate-path outperforms the Gad El-Rab one, achieving an accuracy of respectively 63.03% and 63%; with a signification improvement of 3% over the Gad El-Rab method.

To better evaluate the performance of our proposed method, table 1 shows the highest 10 accuracies and table 2 shows the lowest 10 accuracies grouped by Ambiguous Term. Our WSD method achieved a high accuracy value of 98.58% while executing our method on CCD ambiguous term from MSH-WSD corpus, and a low accuracy value of 20.16% on PCA ambiguous term. In this context, figure 4 presents the 203 ambiguous term accuracy, and we can observe that most of the term accuracy are between 45% and 98%. Which present interesting values regarding Gad El-Rab method, which top value term Accuracy is 97% and the low accuracy is 0%. That can demonstrate the interest of our proposed method.

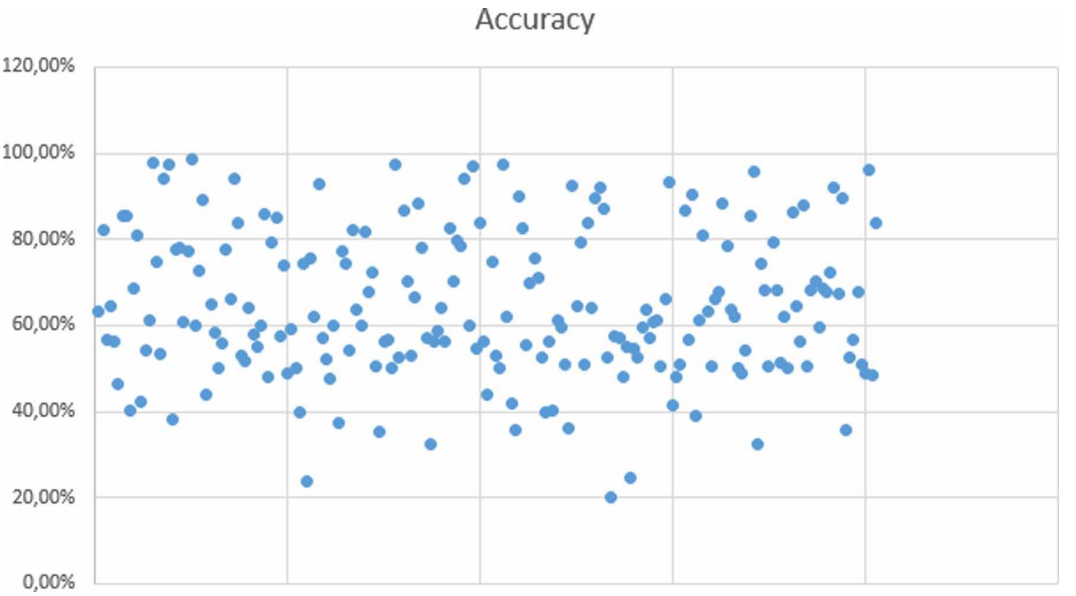
Table 1. HIGHEST 10 ACCURACIES

Ambiguous Term	Total Example	Total Correct	Accuracy
CCD	141	139	**98.58%**
BPD	198	194	**97.98%**
BSE	198	193	**97.47%**
GAG	198	193	**97.47%**
Lawsonia	115	112	**97.39%**
IP	196	190	**96.94%**
WBS	128	123	**96.09%**
SLS	164	157	**95.73%**
INDO	122	115	**94.26%**
CLS	34	32	**94.12%**

Table 2. LOWEST 10 ACCURACIES

Ambigous Term	Total Example	Total Correct	Accuracy
EM	129	48	**37. 21%**
NM	122	44	**36.07%**
TSF	53	19	**35.85%**
Lupus	297	106	**35.69%**
Fish	198	70	**35.35%**
Sodium	197	64	**32.49%**
HGF	192	62	**32.29%**
PHA	110	27	**24.55%**
DE	126	30	**23.81%**
PCA	491	99	**20.16%**

Figure 4. Accuracies for all Ambiguous terms from MSH-WSD



6. CONCLUSION AND PERSPECTIVES

In the biomedical domain, many algorithms had been proposed to overcome the WSD problem, and mostly they rely on an external source of knowledge, like a thesaurus or ontology, for automatically selecting the proper concept for an ambiguous term in a given context using semantic similarity and relatedness measures.

Previous studies compare and evaluate the efficiency of semantic similarity and relatedness measure on Biomedical WSD. In this paper, we have proposed a Web-based Kernel function for measuring the semantic relatedness between concepts to disambiguate an expression versus multiple possible concepts. This measure uses the large volume of documents returned by PubMed search engine to determine the greater context for a biomedical short text through a new term weighting scheme based on Rough Set Theory (RST).

To illustrate the efficiency of our proposition, we have performed experiments evaluating our WSD method using MSH-WSD Dataset. Indeed the results of our experiments show, considerable accuracy improvement for tops and lows accuracies while regrouping performance for each term.

In our future work, we try to integrate and explore the web similarity method in some tasks that implies natural language processing such as biomedical text categorization and information retrieval.

REFERENCES

Bridget, T. M., & Ted, P. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, *46*(6), 1116–1124. doi:10.1016/j.jbi.2013.08.008 PMID:24012881

Brody, S., & Lapata, M. (2009). Bayesian word sense induction. *Proceedings of the 12th conference of the European chapter of the association for computational linguistics*, 103–11. doi:10.3115/1609067.1609078

Garla, V., & Brandt, C. (2012). Knowledge-based biomedical word sense disambiguation: An evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association: JAMIA*, *0*(5), 1–5. PMID:23077130

Navigli, R., Faralli, S., Soroa, A., Lopez de Lacalle, O., & Agirre, E. (2011). Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2317–20. doi:10.1145/2063576.2063955

Plaza, L., Jimeno-Yepes, A., Díaz, A., & Aronson, A. (2011). Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC Bioinformatics*, *12*(1), 355. doi:10.1186/1471-2105-12-355 PMID:21871110

Rais, M., & Lachkar, A. (2016). *Evaluation of Disambiguation Strategies on Biomedical Text Categorization. LNBI – IWBBIO'16*. Springer Verlag.

Rais, M., & Lachkar, A. (2016). Biomedical Word Sense Disambiguation context-based: Improvement of SenseRelate method. In *Proceeding of International Conference on Information Technology for Organizations Development (IT4OD 2016)*. IEEE. doi:10.1109/IT4OD.2016.7479309

Zhong, Z., & Ng, H. (2010). It makes sense: a wide-coverage word sense disambiguation system for free text. *Proceedings of the ACL 2010 system demonstrations, the association for computational linguistics*, 78–83.

Sahami, M., & Heilman, T. (2006). A web-based kernel function for measuring the similarity of short text snippets. *Proceeding of WWW '06*.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Kluwer. doi:10.1145/1135777.1135834

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*(2-3), 211–257.

Dinh, D., & Tamine, L. (2010). Sense-based biomedical indexing, and retrieval. In *Proceeding of International Conference on Applications of Natural Language to Information Systems*. Springer-Verlag.

Eric, S. (2009). Entrez Programming Utilities Help. In *NCBI Help Manual. Bethesda, MD*. National Center for Biotechnology Information.

Gad, E. R. W., Zaïane, O. R., & El-Hajj, M. (2013). Biomedical text disambiguation using UMLS. In *Proceeding of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. IEEE/ACM.

Jimeno-Yepes, A., McInnes, B., & Aronson, A. (2011). An unsupervised vector approach to biomedical term disambiguation: Integrating umls and medline. *BMC Bioinformatics*, *12*(1), 223.

Komorowski, J., Pawlak, Z., Polkowski, L., & Skowron, A. (1998). *Rough Sets: A Tutorial*. Academic Press.

Landauer, T., Foltz, P., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. *Proceeding Cognitive Science (COGSCI-95)*, 660-665.

Ngo, C. L., & Nguyen, S. H. (2003). *A tolerance rough set approach to clustering web search results*. Warsaw University.

Wen-tau, Y., & Christopher, M. (2007). Improving Similarity Measures for Short Segments of Text. *Proceeding of the 22nd national conference on Artificial intelligence*, *(AAAI'07), 2*, 1489-1494.

Zhang & Chen. (2013). A study on clustering algorithm of Web search results based on rough set. In *Proceeding 4th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE.

Mohammed Rais is a Ph.D. student at National school of applied sciences in Fez, Morocco, working on Biomedical text mining subject.

Mohammed Bekkali, a PhD candidate at Laboratory Engineering Systems and Applications, USMBA, Fez, Morocco. Their research area is about Arabic Natural Language Processing and more specifically on Social Networks Content Anaylsis. These social networks allow users to publish short text and it is of course different from traditional document. The main idea of our work is how can we enrich the short text representation from dataset itself or using external resources such as ontologies or thesaurus.

Abdelmonaime Lachkar is working as a Full Professor of Computer Science Engineering in the Department of Information and Communication System E.N.S.A-Abdelmalek Essaadi University, Tanger Morocco. He is a member of Systems and Data Engineering Team. His current research interests include Biomedical Text Mining Applications, Biomedical Information Retrieval Systems, Text Mining Applications for Arabic Language, Sentiment Analysis and Opinion Mining from Social Media, Arabic Web Document Clustering and Categorization. Arabic Information Retrieval Systems, Arabic Text Summarization, Image Indexing and Retrieval, 3D Shape Indexing and Retrieval in large 3D Objects Databases.