View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Institute of Advanced Engineering and Science

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 3, June 2020, pp. 2934~2943
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i3.pp2934-2943

Towards optimize-ESA for text semantic similarity: A case study of biomedical text

Khaoula Mrhar[1], Mounia Abik[2]
[1]IPSS Research Team, FSR, Mohammed V University, Morocco
[2]IPSS Research Team, ENSIAS, Mohammed V University, Morocco

Article Info

Article history:

Received Jun 4, 2019
Revised Jan 3, 2020
Accepted Jan 10, 2020

Keywords:

Explicit semantic analysis ESA
Natural language processing
NLP
Semantic relatedness
Semantic similarity

ABSTRACT

Explicit Semantic Analysis (ESA) is an approach to measure the semantic relatedness between terms or documents based on similarities to documents of a references corpus usually Wikipedia. ESA usage has received tremendous attention in the field of natural language processing NLP and information retrieval. However, ESA utilizes a huge Wikipedia index matrix in its interpretation by multiplying a large matrix by a term vector to produce a high-dimensional vector. Consequently, the ESA process is too expensive in interpretation and similarity steps. Therefore, the efficiency of ESA will slow down because we lose a lot of time in unnecessary operations. This paper propose enhancements to ESA called optimize-ESA that reduce the dimension at the interpretation stage by computing the semantic similarity in a specific domain. The experimental results show clearly that our method correlates much better with human judgement than the full version ESA approach.

Copyright © 2020 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Khaoula Mrhar,
IPSS Research Team,
FSR, Mohammed V University,
Ibn Batouta avenue, Rabat, Morocco
Email: Khaoula_mrhar@um5.ac.ma

1. INTRODUCTION

Semantic relatedness measures quantify the degree in which two words or concepts are related in a taxonomy by using all relations between them, such as synonymy, hyponymy. Semantic similarity is a special case of relatedness and it is limited to hyponymy (i.e. is-a) relations. Measures of relatedness or similarity are used in many Natural Language Processing (NLP) applications, such as word sense disambiguation, Information retrieval , automatic detection and spelling correction, semantic annotation, text clustering and classification, topic detection [1, 2]. Measuring the semantic similarity between texts is a challenging task. The traditional lexical approach based on Bag of Word (BOW) [3] and vector space model [4] which convert each text into a word vector, has a notorious disadvantage that is ignore the semantic relationship among words and treat words independent of each other [3]. One solution to resolve this problem is to enrich text representation with an external source of knowledge. Some technique use large corpora such as the statistical corpus based similarity approach, which measures the semantic similarity metric between two text and word based on the information gained from corpora. A Corpus refers to a large collection of written or spoken texts that is used to study and describe a language. The most relevant technique of this approach is HAL [4], LSA [4], ESA [5]. However , the corpora techniques are unstructured and imprecise. Morever, other techniques use a lexical structures such as taxonomies specially wordnet [6], but wordnet is limited in scope and coverage and does not include the information about named entities and specialized concept, and doesn't give a good results in text similarity [7].

In contrast, to solve these shortcomings, Wikipedia is an outstanding resource for text semantic similarity problem. It's a large-scale collaborative open encyclopedia that has evolved into a comprehensive resource with very good coverage on diverse topics, important entities, events, it widely covers named entities, domain specific entities, and new entities. The English Wikipedia currently contains over 4 million articles (including redirection articles). Furthermore, WikiRelate [7] was the first work which compute the measures of semantic relatedness using Wikipedia, this approach applied the familiar technique used in semantic relatedness based on wordnet and modified it to be used in Wikipedia, such as path-length measure [8], but in general the results are similar. However, Gabrilovich and Markovitch (2007) [5] propose a new approach with Explicit Semantic Analysis (ESA) that achieve highly accurate results, this method has been extensively studied in many applications [9]. ESA use Wikipedia as a semantic interpreter and builds a weighted inverted vector that maps each term into a list of Wikipedia articles in which it appears, and computes the similarity between vectors generated from two terms or texts. It means that the inverted vector may contain a millions of columns with many 0 value considering the sheer size of Wikipedia articles (more than 4Mconcepts). Accordingly, interpreting text based on all Wikipedia concepts can be expensive and computing semantic relatedness after between this huge vectors using Cosine similarity, the efficiency of ESA will slow down.

Several related paper are interested to this problem. [10] Propose Economy-ESA which is an economic schema of explicit semantic analysis ESA, by reduce the ESA index matrix dimension using random selection, k-means and norm-based clustering approaches. The authors in [11] propose a novel graph-based relatedness assessment method using Wikipedia features to avoid the drawbacks. It propose Naive-ESA algorithm to return the top *k* most relevant Wikipedia in order to reduce the dimensional space of Explicit Semantic Analysis (ESA). An efficient and effective algorithm was proposed in [12], it's represent the meaning of a text by using the concepts that best match it. This approach first computes the approximate top-k Wikipedia concepts that are most relevant to the given text and then leverage these concepts for representing the meaning of the given text. Following the above-mentioned studies, in this paper we present a new method that optimize ESA approach and resolve some of its limitation and drawbacks. Optimize-ESA reduce the dimension at the interpretation stage by computing the semantic similarity in a specific domain.

Thus, based on several works [13], using a domain knowledge base is more beneficial and performant in sematic similarity computation process [14]. This result has pushed many researchers to use domain knowledge base when the text input domain is already known.The based majority of work in semantic similarity in a specific domain are in a biomedical domain because of the proliferation of textual resources and the importance of the terminology. In this context, the state-of-the-art methods for calculating semantic relatedness in a specific domain can be roughly divided into two main groups. Those that are concentrated on ontology based methods [15] And distributional methods that use the domain specific corpus [16]. Many attempts to use Wikipedia to compute semantic similarity in a specific domain. [17] assesses the suitability of Wikipedia in the biomedical domain as a potential knowledge resource for semantic relatedness computation by comparing it with other methods (ontology based, distributional methods). However, Jaiswal [18] propose a method for calculating the semantic relatedness of text related to diseases, conditions, and wellness issues that uses ESA with MedlinePlus as its knowledge base instead of Wikipedia.

In this paper, we propose an approach optimize-ESA that perform the ESA approach and provides significant gains in execution time and space consuming without causing significant reduction in precision. In our approach we limit the K concept based on the category Wikipedia tree and the domain input. After that, we leverage these concepts vector to map a text from the keyword-space into the concept-space optimized. All evaluations are performed on datasets containing pairs of terms from biomedical domain and a gold standard semantic similarity value for each pair. The results are compared with the results of the ESA approach and the other state of art semantic similarity approach. The remainder of this paper is organized as follows. Section 3 present our method optimize-ESA and it architecture, Section 4 details the experiments that evaluate the effectiveness of our method and reports the analysis of results in the biomedical domain. Finally, we remark our conclusion and present some perspectives for future research in Section 5.

2. PROPOSED APPROACH: OPTIMIZE-ESA FOR SEMANTIC SIMILARITY MEASURES

2.1. The Wikipedia features

Wikipedia is a large online encyclopedia founded in 2001 and it is a free, editable by users, web-based, collaborative, multilingual encyclopedia. While it underwent a tremendous growth and currently comprises more than 2,382,000 articles in about 250 languages. And become one of the most important information resources in the web.

Wikipedia content is presented on pages:

- Articles: Are the normal page in Wikipedia that contain encyclopedic information, Each article describes a single concept or topic with a concise title that can be used in a ontologies and a brief overview of the topic. There is only one article for each concept or topic.
- Redirects: Redirects is a Wikipedia page which automatically redirects users to another page (connect articles to articles or section of an article). It is possible to redirect to just a specific section of the target page.
- Disambiguation pages: disambiguation is the process of resolving conflicts when article title is ambiguous, it contain a list of articles corresponding to different meaning of the same word. For example, the word "Java" can refer to an island of Indonesia, a programming language, a French band, and many other things.
- Categories: categories are nodes for hierarchical organization of articles, it intend to group pages on similar subjects, almost all Wikipedia articles are within one or more categories.Wikipedia category is organized as a network that we present briefly in section 3.3.1.

2.2. ESA Approach

Explicit Semantic Analysis created by Gabrilovich and Markovitch [19]. This approach consist to represent texts as weighted mixture of a set concepts and using Wikipedia concept which each concept is a title of Wikipedia page. The main advantage of this approach is the use of a vast amount of highly human knowledge. The first step of this approach is to construct the semantic interpreter that maps fragments of natural language text into a weighted sequence of Wikipedia concepts ordered by their relevance to the input. Given a input text Fragment T compose of I words T={wi}, we first represent it as an interpretation vectors using TFIDF Schema Vi , where Vi is the weight of the word wi. Then, we use Wikipedia articles as index documents, each Wikipedia concept is represented as a vector of words that occur in the corresponding article. Entries of these vectors are assigned weights using TFIDF scheme. Hence, these weights quantify the strength of association between words and concepts. We build an inverted index which maps each word into a list of concept in which it appears. Let Kj be an inverted index entry for word Wi , which Kj quantifies the strength of association of word Wi with Wikipedia concept cj , {cj, c1, . . . , cN} (where N denotes the total number of Wikipedia concepts). Then, the semantic interpretation vector V for text T is a vector of length N, in which the weight of each concept Cj is defined as ∑wi€T vi . kj Entries of this vector reflect the relevance of the corresponding concepts to text T . After That ESA uses Cosine metric to compute semantic relatedness of a pair of text fragments by comparing their vectors. The Figure 1 below present the whole ESA process.

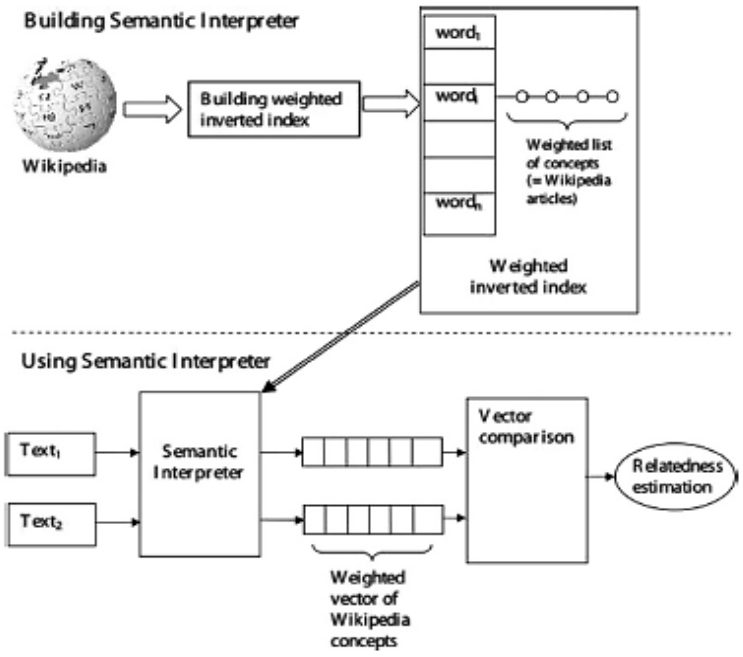

Figure 1. Explicit semantic analysis ESA system

The ESA approach is simple and efficient, however the process is too expensive for many reasons. Firstly, the dimension of concept vector for a given word is too large because it length equals to all concepts in Wikipedia considering the sheer size of Wikipedia article (more than 4 M concept). Secondly, to produce this concept vector, the overall index matrix must be multiplied by a term vector and give a large index matrix that requires numerous multiplications. Thirdly, the space vector of a word is a matrix in which most of the elements are zero because the word will appears just in a few Wikipedia articles. The reinterpretation of text based on Wikipedia concept can be very expensive and slow, because we lose a lot of time in unnecessary operations because the zero value in high-dimensional sparse vectors can impact efficiency and performance of ESA approach. Finally the computations of similarity or relatedness between two vectors with numerous dimensions are very costly. Thus, because of this problems, we propose in this paper an approach which optimize the ESA approach and allowed us to not return the vector space for the whole concepts in Wikipedia but only the top k concepts most relevant. Indeed, given a domain specific, we select the most relevant Wikipedia articles related to domain Di based on Wikipedia category network. Furthermore, we create a domain index Ui that save the inverted index of Wikipedia articles of each domain calculated after a domain Di entered. And for each text T in a specific domain Di, we semantically reinterpret it based on k concept saved in domain index Uj. We process an update for this domain index according to Wikipedia update frequency. We present briefly the optimize ESA approach in the section below.

2.3. Optimize-ESA approach

In this paper, we propose an approach to compute a semantic similarity in a specific domain called the Optimize-ESA approach. This approach resolve some of the shortcomings of ESA approach and optimize it in term of space consuming and time similarity computation. The architecture of our approach presented in Figure 2, it consist of two layers: filter k concept for domain Di and build a domain inverted index.

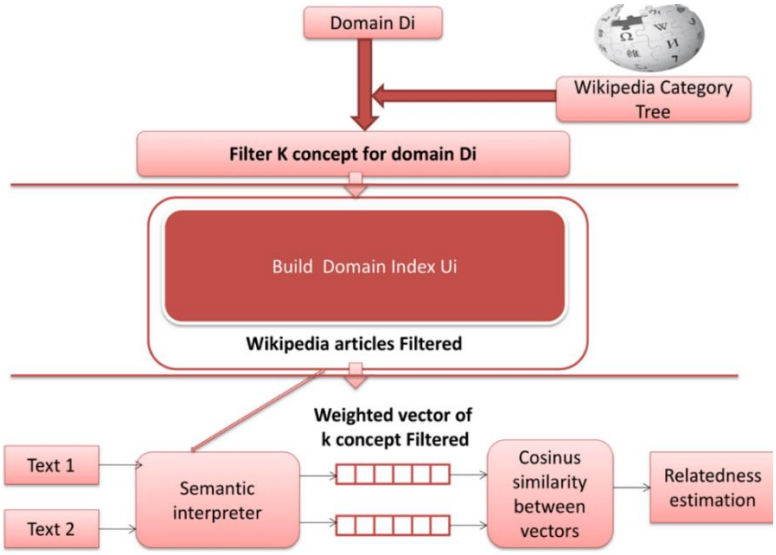

Figure 2. Optimize-ESA architecture

2.3.1. First Layer: filter K concept for domain Di

The relationship between concept or article and category in Wikipedia is expressed by a link called category link (the English version contain 49.98 million inter links in September 2006 [20]). Indeed, the Wikipedia category system is socially created and edited and any user can create an article and classify it into category. This leads to a tremendous growth of articles and categories in Wikipedia (more than 500000 categories in English Wikipedia article [20]). Consequently, Wikipedia editors try to better organize Wikipedia category structure by purifying certain concepts and split category into multiple fine-grained categories (the number of categories in wiki-14 was increased 25% than wiki-12). Furthermore, the category system in Wikipedia is represented as a directed graph where nodes represent pages or categories, and edges represent the oriented relationship "is assigned to". Every category has a multiple parents and children categories. And each category is connected to a number of articles (coverage all Wikipedia articles by a category). Besides, the category system in Wikipedia has a taxonomy structure which is a hierarchy of

topics and subtopics as shown in Figure 3. It enables us to search articles by narrowing from broader categories to the down categories. Indeed, Wikipedia offer a category tree system [21] which enable users to browse categories but not all concepts belonging to a specific category because it's not a tree structure. Nevertheless, starting by a category we can traverse the descendant categories and detect all articles connected.

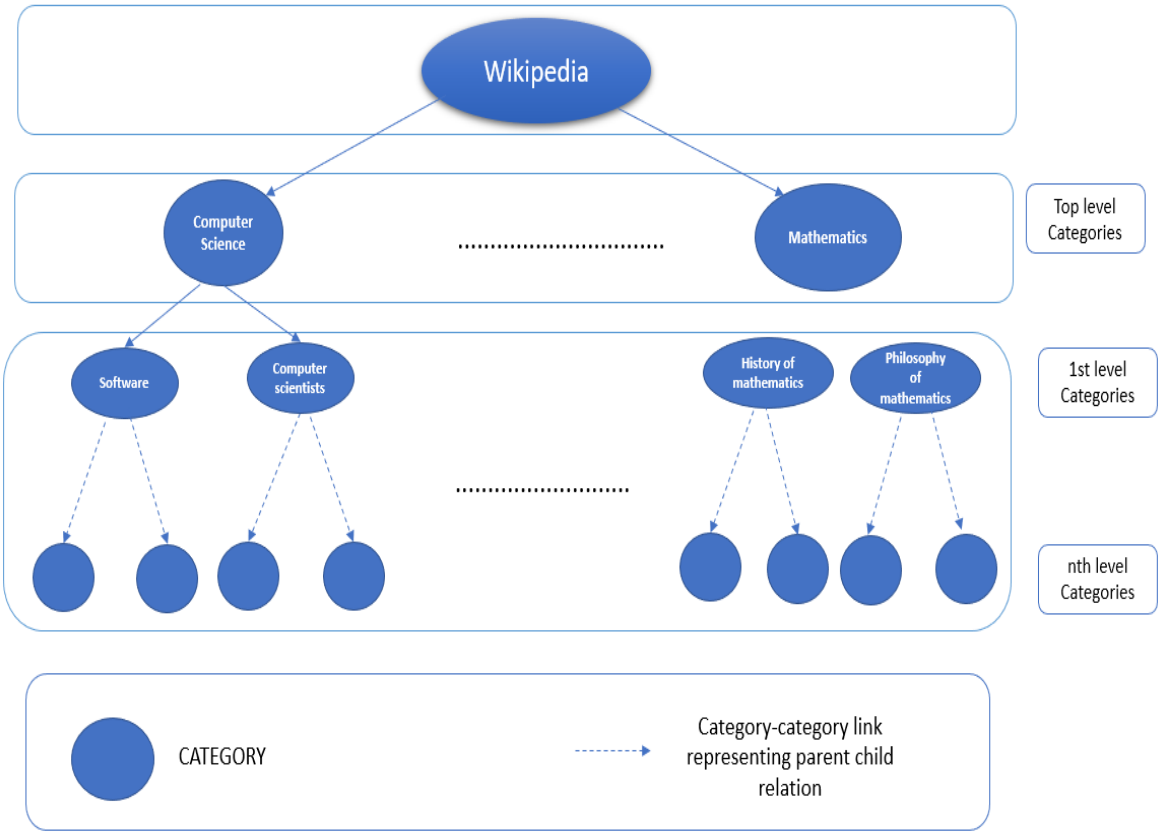

Figure 3. Category tree wikipedia

In this part, we use the Wikipedia category system to extract the articles or concept related to an input domain D. using this category system, we can consider our input domain as a category in Wikipedia and try to search all category belonging, as well as by traversing the descendant categories extract all articles connected. However, as the level increases, we can note that the articles covered are augmented more and more almost all the articles in the Wikipedia are covered. That means, all the articles belong to all the broad categories, which is incorrect. So our issue is how to define which level of the breadth first traversal we need to stop, in other words, in which level in Wikipedia tree structure the categories are effectively related to the category input. Therefore, we propose to compute the semantic similarity between category input and all categories in each level, and deciding after experimentation in which level we need to stop. The Table 1 below present the result of our experimentation.

Based on several experimentation and observation, we find that the categories level that are effectively related to the domain input changes from one domain to another and is not always correct to stop in a specific category level (computer science at 8 level and bioinformatics at 7 level). because it is according to the number of down categories of this domain existing in Wikipedia category system. Therefore, the categories extracted must be based on a semantic similarity measure between domain input and the categories in each level. Consequently, after experimentation, we decided to stop the extraction of sub categories related to domain input after a similarity value of 0.4. The Figure 4 presents the whole process of detecting the Wikipedia articles related to a specific domain input.

Table 1. Correlation semantic similarity between wikipedia category tree levels

Category input	Correlation Semantic Similarity											
	Category Tree Levels											
	1	2	3	4	5	6	7	8	9	10	11	12
Computer science	0.8389	0.6858	0.6127	0,587	0,557	0,517	0,489	0,435	0,387	0,345	0,287	0,198
Bioinformatics	0.629	0.5058	0,502	0,497	0,476	0,456	0,436	0,427	0,357	0,245	0,227	0,175
Biology	0.7205	0.6063	0,598	0,576	0,554	0,518	0,486	0,423	0,397	0,297	0,267	0,109
Medicine	0.7379	0.64498	0.5870	0.5630	0.5563	0.536	0,501	0,456	0,345	0,329	0,234	0,206

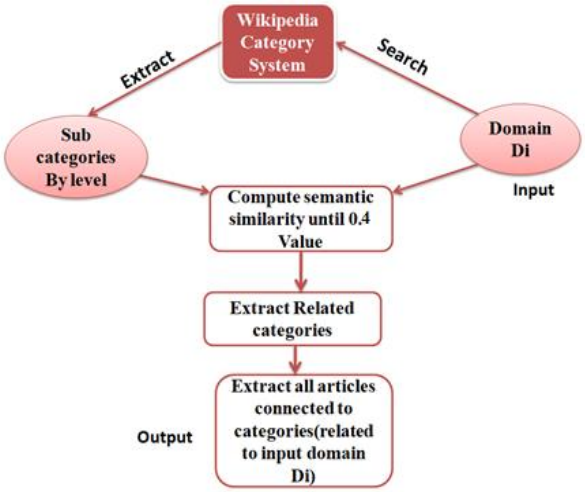

Figure 4. The process of detecting wikipedia articles related to domain input

2.3.2. Second layer: Build domain index Ui

After the filtering of the Wikipedia articles related to a specific domain Di, we build an inverted index domain Di which maps each word into a list of concept in which it appears as presented in section 3.2.1. Let kj be an inverted index entry for word wi, where kj quantifies the strength of association of word Wi with Wikipedia concept cj , {cj ∋ C1,…..,Cn}, where n denotes the number of Wikipedia concept filtered for domain Di as appear in table 2.

Table 2. Wikipedia articles filtered for domain Di

	WA1	…..	…..	WAj
Term 1	T[0,0]			
…..				
…..				
Term k	…..	….	….	T[i,j]

Terms in wikipedia articles filtered for domain Di

After building the weighted inverted index for domain Di, We store it in a database as Ui to use it for any future interpretation to optimize the computation of semantic similarity. Our database must be updated for selecting new articles added to Wikipedia, the algorithm of our method is presented below:

//the algorithm create the inverted index wikipedia for a specific domain that can be used in the similarity semantic measures between text based on ESA method
// Input : domain Di
// output : domain index Ui
step 1
//extract k concept related to domain input Di
for domain Di
if Di exist in U
return U[Di]
Else

//Search Di in node category tree wikipedia
Ck =search [Di,CT]
//extract K concept Ck [C1….Ck] belongs to Di
Return Ck
Step 2
// build inverted index for domain Di, WDi
For C1 to Ck
WDi [C1…..Ck]
store WDi in Ui
return Ui
stop

Furthermore, To compute the semantic similarity between two text T1 and T2 , we consider it as a bag of words T1= {t1,t2,…tn} with n words. And we semantically reinterpret it based on k concept saved in domain index Ui. And finally we compute the sematic similarity between the two text vectors based on a cosines similarity metric.

3. RESULTS AND DISCUSSION

3.1. Case study: biomedical domain

In the last years, the amount of information available in textual format is rapidly increasing in the biomedical domain such as patient health records and medical documents. Therefore, Measures of semantic relatedness between concepts and texts is widely used in this domain, discovering similar diseases [22], and redundancy detection in clinical records [23], comparing gene products [24], identifying direct and indirect protein interactions within human regulatory pathways using gene ontology [25], coding medical diagnoses and adverse drug reactions using semantic distance [26]. Furthermore, the classical semantic similarity computation measures have been adapted to be used in several domain. However, these measures are less efficient due to the limited coverage of specialized domains. That is why, the need to use a specialized knowledge base such as in the biomedical domain, by exploiting the medical ontologies, knowledge repositories and biomedical structured vocabularies. For this reason, we propose in this paper a domain specialized method that optimize ESA semantic similarity approach. We choose to test the performance of our method on three biomedical dataset because of the availability and proliferation of the resources. We present in the section below the dataset used in our experimentation and the interpretation of our result.

3.2. Experimentation

Humans have an innate ability to judge semantic relatedness of texts. Accordingly, to evaluate the performance of machine measurement of semantic similarity between texts, we compare them with human rating on the same setting by compare the correlation between human judgement and machine calculations. In this work, because of the no suitability of dataset of biomedical pairs sentences as appear in Table 3. We use BIOSSES Dataset [27], which is a benchmark dataset for biomedical sentences similarity estimation. It contain 100 sentences pair selected from the TAC (Text Analysis Conference) biomedical summarization track training containing articles from the biomedical domain. The sentences pairs were evaluated by five different human expert that give a scores ranging from 0 (no relation) to 4 (equivalent). Which averaged for each pair to produce a single relatedness score. We test our method also on two French Web corpora [28]. The first corpus is about “epidemics” and the second one is about “space conquest.” Each corpus contains reference sentences and each of them was associated with six sentences chosen with similarities score ranging from 0 (the sentences are unrelated) and 4 (the sentences are completely equivalent).

Table 3. The datasets used in semantic similarity task

Dataset	Pairs	Scale	References
BIOSSES	100	1-5	[27]
Epidemics	60	0-4	[28]
Space conquest	60	0-4	[28]

Following the literature on semantic relatedness, we evaluate the performance by measuring a pair correlation scores between the score assigned by the proposed method and human judgement score for each dataset we report the correlation computed on all pairs with the metric Pearson’s correlation coefficients.

The Pearson's correlation metric denoted as P reflects the linear correlation between measuring result with human judgments, where 0 means uncorrelated and 1 means perfect correlated. The corresponding formula is defined as:

$$P = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2}},$$

where xi refers to the value of the ith in the dataset given by human judgments, yi to the corresponding value returned by an Optimize-ESA method, and n to the length of the target dataset.

Table 4 show the correlation coefficient Pearson by the ESA algorithm and our methods Optimize-ESA for the three datasets BIOSSES, Epidemics and Space Conquest. Our method optimize ESA gets a correlation of 0.612 compared to 0.595 for ESA method for sentences dataset BIOSSES. On the Epidemics dataset, our method gets a correlation of 0.544 compared to 0.525 for the full version ESA. And ESA approach with Wikipedia knowledge base get a correlation of 0.558 for Space conquest dataset compared to 0.571 for our method. This clearly show that our method correlates much better with human judgement than the full version ESA approach. A comparison of our method Optimize-ESA and some state-of-art for computing semantic relatedness in the biomedical domain is shown in Table 5. We compare it with Resink and Lin which is the most popular information content measures in knowledge based methods. In addition, Levenshtein which is a string based measure. Besides comparing our optimize ESA with the traditional ESA approach with wikipedia as a knowledge graph.

Table 4. The comparison of Pearson's correlation coefficient on BIOSSES, Epidemics, Space conquest Datasets

Dataset	ESA Algorithm (Gabrilovich & Markovitch, 2007)	Optimize-ESA
	Pearson's (P)	Pearson's (P)
BIOSSES	0.595	0.612
Epidemics	0.525	0.544
Space conquest	0.558	0.571

Table 5. Correlation coefficients pearson (P) between related studies

Related studies	Dataset			References
	BIOSSES	Epidemics	Space Conquest	
		IC-based measures		
Resink	0.473	0.396	0.412	P.Resnik [29]
Lin	0.645	0.591	0.611	D.Lin [30]
		String similarity measures		
Levenshtein	0.592	0.601	0.591	Finkelstein et al., [31]
		ESA similarity measures		
ESA-wiki	0.595	0.525	0.558	Gabrilovich and Markovitch [19]
Optimize-ESA	0.612	0.544	0.571	

As the above results in Table 5 indicate that the optimize-ESA can obtain competitive results for Pearson correlation especially for the small dataset. In contrast, in the big size dataset, the use of the full version ESA including all concepts in Wikipedia or optimize-ESA in a domain specific is more performant compared to string similarity measure and IC based measures. Furthermore, we noticed that our method optimize-ESA is faster than ESA with full Wikipedia after an experimentation presented in Figure 5. We measured the cosines similarity processing cost of six pairs from each test collection and we compute the running time comparison between ESA and Optimize-ESA.

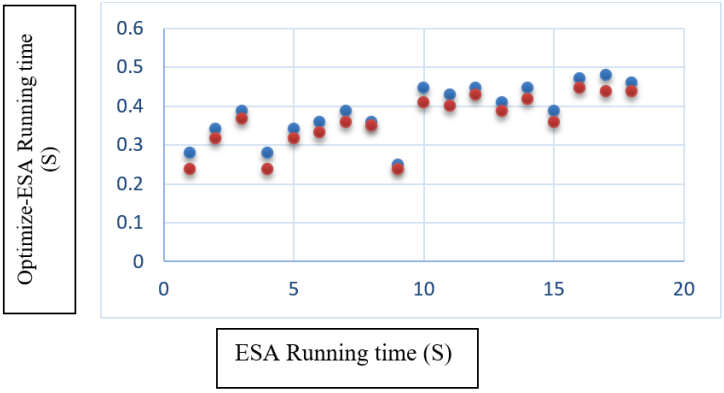

Figure 5: ESA & Optimize-ESA Running Time

4. CONCLUSION AND FUTURE WORK

The study of semantic similarity between words has long been an integral part of information retrieval and natural language processing. Based on the theoretical principles and the way in which ontologies are investigated to compute similarity, different kinds of methods can be identified according to type, size and domain of dataset. Among these methods, we can cite the Explicit Semantic Analysis ESA approach with Wikipedia knowledge base which perform very well the task of computing the sematic relatedness of word and text fragment. However, The ESA process is too expensive due to the large length dimension of concept vector for a given word which equals all Wikipedia concept (4 M). And the efficiency of ESA will slow down because we lose a lot of time in unnecessary operations.

We propose in this paper a new method called optimize-ESA which reduce the dimension at the interpretation stage by computing the semantic similarity in a specific domain. To evaluate the performance of our method, we give a comparison between different algorithms for Semantic Relatedness in the biomedical domain. We choose the biomedical domain because of the availability of different ontologies and methods, which is significantly higher than any other domain. We conclude that our method outperforms the current state-of-the-art methods for calculating the semantic relatedness of biomedical texts as it correlates much better with human judgements. There are two other interesting lines of future research related to the method presented in this work. Firstly, we plan to more optimize our method by filtering the Wikipedia concept using the domain specific knowledge based leveraged with Wikipedia category tree. Secondly, we plan to more perform the result of ESA by adding to the weighted inverted index a category index. Finally, a wider evaluation will be desirable, considering larger sets of text pairs as benchmark data in other domain.

REFERENCES

[1] S. Tongphu, "Toward Semantic Similarity Measure Between Concepts in An Ontology," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 3, pp. 1356-1372, 2019.
[2] Bazi and N. Laachfoubi, "Arabic Named Entity Recognition using Deep Learning Approach," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 3, pp. 2025-2032, 2019.
[3] Y. Zhang, R. Jin and Z. Zhou, "Understanding Bag-of-Words Model: A Statistical Framework", *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43-52, 2010.
[4] T. Landauer, P. Foltz and D. Laham, "An Introduction to Latent Semantic Analysis," Discourse Processes, vol. 25, no. 2-3, pp. 259-284, 1998.
[5] Evgeniy Gabrilovich and Shaul Markovitch. "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," In AAAI' 06, pp. 1301–1306, 2006.
[6] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13-47, 2006.
[7] M. Strube, and S. P. Ponzetto. WikiRelate! Computing Semantic Relatedness using Wikipedia," Proceedings of the *21st National Conference on Artificial intelligence*, Boston, Massachusetts, pp.1419-1424, 2006.
[8] Nurifan, R. Sarno and C. Wahyuni, "Developing Corpora using Wikipedia and Word2vec for Word Sense Disambiguation," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 3, pp. 1239-1246, 2018.
[9] K. Dramé, F. Mougin and G. Diallo, "Large Scale Biomedical Texts Classification: A kNN and An ESA-Based Approaches," *Journal of Biomedical Semantics*, vol. 7, no. 1, 2016.

[10] F. Rahutomo, Y. Manabe, T. Kitasuka and M. Aritsugi, "Econo-ESA Reduction Scheme and the Impact of its Index Matrix Density," *Procedia Computer Science*, vol. 35, pp. 474-483, 2014.
[11] P. Li, B. Xiao, W. Ma, Y. Jiang and Z. Zhang, "A Graph-Based Semantic Relatedness Assessment Method Combining Wikipedia Features," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 268-281, 2017.
[12] Kim, J. W. Kashyap, A and Bhamidipati, S. Wikipedia-Based Semantic Interpreter using approximate Top-K Processing and Its Application, vol. 18, pp. 650-675, 2012.
[13] X. Song, "Ontology-based Domain-specific Semantic Similarity Analysis and Applications," *Computer Science*, All Dissertations, 2018.
[14] T. Gottron, M. Anderka and B. Stein, "Insights into Explicit Semantic Analysis," In Proceedings of the *20th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 1961-1964, 2011.
[15] V. Garla and C. Brandt, "Semantic Similarity in the Biomedical Domain: An Evaluation Across Knowledge Sources," *BMC Bioinformatics*, vol. 13, no. 1, 2012.
[16] D. Sánchez, M. Batet, and A. Valls, "Web-Based Semantic Similarity: an evaluation in the Biomedical Domain," *Int J Softw Inform*, vol. 4, pp. 39-52, 2010.
[17] E. Costa, H. Tjandrasa and S. Djanali, "Text Mining for Pest and Disease Identification on Rice Farming with Interactive Text Messaging," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 3, pp. 1671, 2018.
[18] A. Jaiswal, and A. Bhargava, "Explicit Semantic Analysis for Computing Semantic Relatedness of Biomedical Text," In Proceedings of *Confluence The Next Generation Information Technology Summit (IEEE)*, 2014
[19] E. Gabrilovich and S. Markovitch. "Computing Semantic Relatedness using Wikipedia-based Explicit Se mantic Analysis," In Proceedings of the *20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1606-1611, 2007.
[20] R. B. Bairi, M. Carman, and G. Ramakrishnan, "On the Evolution of Wikipedia: Dynamics of Categories and Articles," *Ninth International AAAI Conference on Web and Social Media*, pp. 6-10, 2015.
[21] "PetScan", Petscan.wmflabs.org, 2019. [Online]. Available: https://petscan.wmflabs.org/. [Accessed: 31- May- 2019].
[22] S. Mathur and D. Dinakarpandian, "Finding Disease Similarity Based on Implicit Semantic Similarity," *Journal of Biomedical Informatics*, vol. 45, no. 2, pp. 363-371, 2012.
[23] Zhang R, Pakhomov S, McInnes B. T, and Melton G. B, "Evaluating Measures of redundancy in clinical texts," AMIA Annu Symp Proc. pp. 1612-1620, 2011.
[24] C. Pesquita, D. Faria, A. Falcão, P. Lord and F. Couto, "Semantic Similarity in Biomedical Ontologies," *PLoS Computational Biology*, vol. 5, no. 7, 2009.
[25] X. Guo, R. Liu, C. Shriver, H. Hu and M. Liebman, "Assessing Semantic Similarity Measures for the Characterization of Human Regulatory Pathways," *Bioinformatics*, vol. 22, no. 8, pp. 967-973, 2006.
[26] C. Bousquet, *et al.*, "Appraisal of the MedDRA Conceptual Structure for Describing and Grouping Adverse Drug Reactions," *Drug Safety*, vol. 28, no. 1, pp. 19-34, 2005.
[27] G. Soğancıoğlu, H. Öztürk and A. Özgür, "BIOSSES: A Semantic Sentence Similarity Estimation System for the Biomedical Domain," *Bioinformatics*, vol. 33, no. 14, pp. i49-i58, 2017.
[28] Vu, H. H., Villaneau, J., Saïd, F., and Marteau, P. F. "Sentence Similarity by combining Explicit Semantic Analysis and Overlapping N-Grams," In Proceedings of the *17th International Conference on Text, Speech and Dialogue (TSD 2014)*, Brno, Czech Republic, Springer International Publishing, pp. 201–208, 2014.
[29] Resnik, P., "Using Information Content to Evaluate Semantic Similarity in A Taxonomy. In Proceedings of *IJCAI*, pp. 448–453. 1995.
[30] Dekang Lin, "An Information-Theoretic Definition of Word Similarity," In ICML'98, 1998
[31] L. Finkelstein, *et al.*, "Placing Search in Context: The Concept Revisited," *ACM Transactions on Information Systems*, vol. 20, pp. 116–131, 2002.

BIOGRAPHIES OF AUTHORS

Khaoula Mrhar is currently Ph.D student at IPSS research team in Mohammed V university Rabat, Morocco. Her background includes a degree in mathematics and computer science. Her research interest contains Formal and Non Formal learning, artificial intelligence, data integration, text mining and recommender system.

Mounia Abik I received a PhD from the National High School for Computer Science and Systems Analysis (ENSIAS) in 2009 and an Habilitation to Drive Research (HDR) from Mohammed V University of Rabat in 2014. My main research interests focus on e-Learning, Knowledge Extraction from Social Networks, Semantic Web and Cyber-violence.