Vietnamese Sentence Similarity Based on Concepts

Hien Nguyen, Phuc Duong, Vinh Vo

Vietnamese Sentence Similarity Based on Concepts

Hien T. Nguyen, Phuc H. Duong, Vinh T. Vo

Faculty of Information Technology, Ton Duc Thang University, Vietnam
hien@tdt.edu.vn, huuphucduong@gmail.com, vothanhvinh@tdt.edu.vn

Abstract. We propose a novel method for measuring semantic similarity of two sentences. The originality of the method is the way that it explores the similarity of concepts referred to in the sentences using Wikipedia. The method also exploits Wiktionary to measure word-to-word similarity. The overall semantic similarity is a linear combination of word-to-word similarity, word-order similarity, and concept similarity. We build datasets consisting of 45 Vietnamese sentence pairs and then evaluate the method on these datasets. The results show that in the best cases, concept similarity help improving the performance of our method more than 15% point. The proposed method is language-independent and quite easy to employ. Therefore, one can readily adopt our method to measure semantic similarity for sentences written in other languages.

Keywords: Paraphrase Identification; Text Similarity; Semantic Similarity

1 Introduction

We study the task of measuring semantic similarity of short texts, i.e., sentence, text segments or very short text snippets. With the development of natural language applications recently, this task has been playing an increasingly important role in plagiarism detection, question answering, machine translation, text summarization, information retrieval, etc. This is a challenging task. Considering the following two sentence pairs, two sentences in the first pair are more likely similarity in meaning even though not many words they contain are common; while all most all words in two sentences in the second pair are the same but their meanings are different.

- "*She has to pass the exam.*" and "*She must get through the exam*"
- "*To gain admission to UCLA, you need to present an academic profile much stronger than represented by the minimum UC admission requirements below.*"[1] and "*To gain admission to Berkeley, you need to present an academic profile much stronger than that represented by the minimum UC admission requirements*"[2]

Until now, there have been many methods proposed for measuring text similarity. Most of them use either knowledge-based or corpus-based word-to-word semantic

[1] *www.admissions.ucla.edu/prospect/adm_fr/fracadrq.htm*
[2] *admissions.berkeley.edu/sites/default/files/docs/Freshman_Flier.pdf*

similarity measures [1], in combination with syntactic information [4], [5], [13], [14] or string matching algorithm [2]. Some works exploited machine translation metrics [7], discourse information [6], or graph subsumption [12] for paraphrase identification. In summary, the work in literature did not exploit concepts, i.e. named entities and common concepts, including coreference and disambiguation, for measuring similarity of short texts. A concept we mean in this work is a named entity such as a person, a location, or an organization in particular, etc. or a common concept such as computer science or information technology.

We propose a novel method to compute semantic similarity between sentence pairs. We evaluate the method on a dataset consisting of Vietnamese sentences. The method explores concepts to determine how they contribute to the performance of measuring semantic similarity of sentences. It also exploits word-to-word similarity and word-order similarity as proposed by Li *et al*. in [4]. An intuition shows that exploring named entities (as well as concepts) and their features is potentially in improving the performance of semantic similarity of sentences, especially for those with the same meaning containing named entities but few words in common. For example, with these two sentences "*I am currently working at IBM*" and "*I am a developer at International Business Machines*", if we only compute the similarity based on words and word-order, the similarity score may not be high as it would be even though the entities *IBM* and *International Business Machines* are the same in the contexts of the two sentences.

One of the challenging problem when dealing with Vietnamese texts is that Vietnamese is a language with a deficient natural language processing support, such as no Vietnamese WordNet or corpus like Brown Corpus of American English for measuring semantic similarity between Vietnamese words. To overcome that limitation, we exploit Wikipedia[3] to recognize which concepts referred to in compared sentences exist in Wikipedia to expand the context of those sentences by different surface forms of the concepts and exploit Wiktionary[4] to estimate the similarity of words.

Wikipedia is a free online encyclopedia whose content is contributed by a large number of volunteer users. It consists of a large collection of articles, each of which defines and describes a concept. In reality, a concept may have several surface forms and one surface form may be used to refer to different concepts in different contexts. In Wikipedia, many-to-many correspondence between names and entities can be captured by utilizing *redirect pages* and *disambiguation pages*. A redirect page typically contains only a reference to an article. The title of a redirect page is an alternative surface form of the described entity or concept in that article. For example, from redirect pages of the United States, we extract alternative surface forms of the United States such as "US", "USA", "United States of America", etc. A disambiguation page is created for an ambiguous surface form which may use to denote two or more entities in Wikipedia. It consists of links to articles that define the different concepts having the same surface form. Wiktionary is a free-content multilingual dictionary, designed as the lexical companion to Wikipedia and opened for volunteers to edit all the

[3] http://vi.wikipedia.org
[4] http://vi.wiktionary.org

contents. It provides the meaning of vocabulary, not includes the encyclopedic information – an advantage of Wikipedia.

The contributions of this paper is three-fold as follows: (i) we propose a novel method that based on concepts for measuring similarity of Vietnamese sentences, (ii) we build a dataset consisting of 45 Vietnamese sentence pairs, each of which was estimate by human subjects if it is paraphrase or not and evaluate our proposed method on this dataset. The method exploits Wiktionary to measure similarity of words, exploits Wikipedia to identify which concepts referred to in the compared sentences existing in Wikipedia for expanding the contexts of those sentences by different surface forms of the concepts. The originality of this work is the way that our method expands the contexts of compared sentences using Wikipedia.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 presents our proposed approach. Section 4 presents datasets, experiments and results. Finally, we draw conclusion in Section 5.

2 Related work

There is much research on sentence similarity measurement. In [1], the authors proposed a method that estimates the semantic similarity between two short texts using both corpus-based and knowledge-based similarity measures of words. Given two short text segments, the method finds for each word in the first one the most similar matching word in the second one and then the similarity between those word pairs are included in the overall semantic similarity of the two text segments. In [2], the authors introduced a method that computes the text similarity by combining a corpus-based similarity measure of words and a modified version of the Longest Common Subsequence string matching algorithm. The method proposed in [9] uses word-to-word similarity derived from WordNet to identify paraphrase.

The methods proposed in [4], [5], [13], [14] measure text similarity based on semantic and syntactic information contained in the compared texts. In addition to using word-to-word similarity, the method in [4] presents the important role of word-order in improvement of sentence similarity measure, the method in [5] exploited adjectives and adverbs in two sentences, the method in [13] takes nouns and verbs in consideration, and the method in [14] exploited interdependent between word-to-word similarity and sentence similarity and computed both of them simultaneously by an iterative algorithm. In [12], the authors use graph subsumption (originally developed for recognizing entailment) with lexical, syntactic, synonymy and antonymy information to identify paraphrase.

The method proposed in [6] combines machine translation metrics and the ordered similarity between elementary discourse units (EDUs). An EDU are blocks of words playing an important role in sentence similarity. In [3], the authors proposed a method that takes advantage of web search results to extend context of short texts and in [8] the authors uses unfolding recursive auto-encoder method for measuring the similarity. In [11], instead of identifying the similarity, the authors propose a method detecting dissimilarities. The method proposed in [14] combines semantic and statistical

information within short texts to compute the similarity. The method proposed in [13] takes advantage of corpus-based ontology to overcome the problem that evaluates the semantic similarity between irregular sentences.

In this paper, we propose a novel method that explores concepts to measure the similarity of sentences. In contrast with related methods, our method use Wiktionary instead of WordNet to measure the similarity of words. Moreover, it identifies which concepts referred to in the compared sentences existing in Wikipedia for expanding the contexts of those sentences by different surface forms of the concepts.

3 Proposed method

We propose a novel method that computes overall semantic similarity of two sentences, let say S_1 and S_2, as follows:

$$Sim(S_1, S_2) = \alpha \times Sim_{word} + \beta \times Sim_{word-order} + \gamma \times Sim_{concept} \quad (1)$$

where Sim_{word}, $Sim_{word\text{-}order}$, $Sim_{concept}$ are sentence similarity measures based exclusively on word-to-word similarity, word-order similarity, and concept similarity respectively; $\alpha + \beta + \gamma = 1$; and the coefficients $\{\alpha, \beta, \gamma\}$ decide the contribution of word-to-word similarity, word-order similarity and concept similarity to the whole sentences. In following three sub-sections, we respectively present the similarity measures.

3.1 Model

Our method contains four main steps for computing semantic similarity between two sentences. The first step performs tokenizer. The second step calculates the similarity between sentences by using word-to-word similarity based on Wiktionary. The third computes similarity between sentences by using word-order similarity of the two sentences. The fourth step recognizes concepts in the two sentences using Wikipedia and expands the contexts of the sentences to compute the concept similarity; then the similarity between sentences is computed using the concept similarity. Finally, the overall sentence similarity score is derived by combining the word-to-word similarity, the word-order similarity and the concept similarity.

3.2 Sentence similarity based on word similarity

The essence of sentence similarity is word-to-word similarity. Thus, we propose to use the Text Overlap method to compute the word-to-word similarity between two words. To our knowledge, this is the most possible method for computing the word-to-word similarity based on Vietnamese Wiktionary. Text Overlap method was first introduced in 1986 by Michael Lesk [16]. The main idea of the method is based on the level of the intersection of gloss texts. The higher level of intersection is, the more similar two words are and otherwise.

We first have to split two input sentences into tokens, then, we will create semantic vectors base on those tokens. Now, we will go through all steps in word-to-word Sim-

ilarity. Given two sentences S_1 and S_2, a joint word set is defined by $S = S_1 \cup S_2 = \{w_1; w_2; \ldots; w_n\}$. The word set S contains all distinct words from the two sentences and includes all inflectional morphology words. For example, *word* and *words*, *thesis* and *theses* are considered as distinct words and must appear in *S*. For example, given the following sentences:

- S_1 = I am a developer at International Business Machines.
- S_2 = I am currently working at IBM.
- S = {I, am, a, developer, at, International, Business, Machines, currently, working, IBM}.

Because the word set *S* is derived from two sentences, we should use it as a standard semantic vector for comparing, denoted by *V*. The semantic vector of two sentences (V_1 and V_2) must have the same dimension of *V* and each value of an entry of the semantic vector (V_1 and V_2), denoted by s_i *(i = 1, 2, …, m)*, is determined by word-to-word similarity method and should lie between [0,1]. Taking S_1 as an example:

- *Case 1*: if w_i appears in S_1, set s_i to 1.
- *Case 2*: if w_i not appears in S_1, word-to-word similarity method will be used to compute the similarity between w_i and each word in S_1. Thus, the word-to-word similarity score should be the highest number *k*. If *k* exceeds a standard threshold, then $s_i = k$, otherwise, set $s_i = 0$. Suppose that the highest number $k = 0.01$, the value of s_i should be 0, because 0.01 is closer to zero, and that's why we should have a standard threshold.

Unlike other text similarity methods, this approach keeps all function words, since these words contain syntactic information if the sentence is too short. Although they appear in the joint word set *S*, they can't affect the whole meaning of the sentence, as well as the semantic vectors, because we use a threshold to eliminate them. After we have two semantic vectors V_1 and V_2, the semantic similarity between two sentences is computed by cosine coefficient of those vectors:

$$Sim_{word} = \frac{V_1 \cdot V_2}{|V_1| \cdot |V_2|} \tag{2}$$

3.3 Sentence similarity based on word order similarity

Before starting this section, we should look over the two example sentences below, they contain the same words in each sentence but differ from the position of two words *boy* and *girl* as follows:

- S_1 = A boy buys the girl a gift.
- S_2 = A girl buys the boy a gift.

We can see that if we only consider word-to-word similarity on that candidate pair, the similarity score will be maximum because all words in that pair are the same. The dissimilarity between S_1 and S_2 is caused by the different word-order. To resolve this problem, we form a joint word set *S*, each entry value denoted by w_i *(i = 1, 2, …, m)*.

S = {A, boy, buys, the, girl, a, gift}. We assign an index number for each word in S_1 and S_2. To compute the similarity score, a word-order vector R_i is formed for each sentence with the length equals to the size of S and each value of R_i is simply an index number in the sentence. Taking S_1 as an example, for each word w_i in S, we will consider its position or the most similar between it and the others, look the cases as follows:

- *Case 1*: if w_i appears in S_1, entry value of this word in R_1 is the index number from S_1.
- *Case 2*: if w_i doesn't appear in S_1, entry value of this word in R_1 is zero.

For example of S_1 and S_2, two word-order vectors R_1 and R_2 should be:

- R_1 = {1, 2, 3, 4, 5, 6, 7}
- R_2 = {1, 5, 3, 4, 5, 2, 7}

Through this approach, we can see that R_1 and R_2 clearly show the basic structure of the sentence. Eventually, a measure for computing word-order similarity of the two input sentences is:

$$Sim_{word-order} = 1 - \frac{\|R_1 - R_2\|}{\|R_1 + R_2\|} \quad (3)$$

3.4 Sentence similarity based on concept similarity

In order to improve the quality of assessment of similarity between two sentences, we determine the similarity between concepts occurring in sentences base on Wikipedia. To better understand the purpose of applying concept similarity, considering the following example with two sentences S_1 and S_2. If we only compute semantic similarity base on words and word-order, we can't detect two entities *Ho Chi Minh City* and *Sai Gon* are the same. As a result, it may lead to the similarity between two sentences not accurate.

- S_1 = I am living in Sai Gon.
- S_2 = I am living in Ho Chi Minh City.

We adopt the mention recognition of Huy *et al*. [19] to identify surface forms of Wikipedia concepts in a sentence. If an identified surface form has only one candidate concept, our method collects all of its surface forms - extracted from its title and its redirect page titles in Wikipedia. If an identified surface form is ambiguous, our method does not disambiguate it, but collects all surface forms of all of its candidate concepts based on the disambiguation page of the identified surface form in Wikipedia.

Given a pair of sentences, let SF_1 be a set of surface forms identified in the first sentence and SF_2 be a set of surface forms identified in the second sentence. The sentence similarity based on concept similarity is computed as follows:

$$Sim_{concept} = \frac{|SF_1 \cap SF_2|}{Min(|SF_1|, |SF_2|)} \quad (4)$$

4 Dataset and Evaluation

4.1 Dataset

To evaluate the performance of our method, we build a dataset consisting of 45 pairs of sentences. Then the dataset is sent to different persons to estimate if each pair is paraphrase or not. If a pair is paraphrase, result will be 1, otherwise result will be 0. After collecting seven survey results, we analyze and create two datasets to serve the assessment process. The results show that there are 19 pairs getting the same agreement by 7 persons, 6 pairs getting the same agreement by 6 persons, 6 pairs getting the same agreement by 5 persons. In total 31 pairs get the same agreement by at least 5 persons. We give 7 persons a chance to discuss on these 31 pairs and get a dataset consisting of 31 sentence pairs with agreement by 7 persons, namely Dataset-1. We build Dataset-2 as follows: the assessments which have at least four number of 1 (the same agreement by at least 4 persons) will unify to 1, the others will be 0. It means that Dataset-2 consists of 45 pairs of sentences.

4.2 Evaluation

In the fields of science, engineering, industry, and statistics, "the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value". In our case, we use a binary classification to measure of how well a test correctly identifies or excludes a condition. Let's take a look the problem we have: the output of computational process lies in [0;1], thus, which value of threshold τ give the similarity of two sentences is 1 or 0; and with that τ, how well the accuracy of the approach is. To solve the problem, we will initially set $\tau = 0.5$, each loop τ will increase a value $t = 0.01$, if the output of computational process is larger than τ, the similarity will be 1, otherwise, will be 0. The value of $accuracy$ calculated by applying the formula:

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

where *TP* denotes true positive, *TN* denote true negative, *FP* denote false positive, *FN* denote false negative.

We evaluate our method on Dataset-1 and Dataset-2 respectively. Table 1 shows the results after running our method on Dataset-1. The column "τ" shows the τ values, The column "accuracy with concepts" shows overall sentence similarity of our method in the term of accuracy, and the column "accuracy without concepts" shows overall sentence similarity of our method in the case we do not use concept similarity. The results shown in Table 2 are explained as the same as doing for the results shown in Table 1.

As shown in Table 1, for Dataset-1, the accuracy is maximized at threshold $\{0.50 \text{ to } 0.55, 0.62, 0.63\}$ and minimized at $\{0.94, 0.95\}$. As shown in Table 2, for Dataset-2, the accuracy is maximized at threshold $\{0.50 \text{ to } 0.55, 0.62, 0.63\}$ and min-

imized at $\{0.94, 0.95\}$. Fig.1 and Fig 2 respectively show the curves comparing the performance of our method on Dataset-1 and Dataset-2 with and without using concept similarity.

We can see that, the higher threshold is, the smaller accuracy is, because the value of τ at low level is closer to people judgment. All in all, the results show that in the best cases, concept similarity help improving the performance of our method more than 15% point; which prove that concept similarity significantly contribute to the performance of our method.

Table 1. The overall accuray of our method on Datatset-1

τ	Accuracy with Concepts	Accuracy without Concepts	τ	Accuracy with Concepts	Accuracy without Concepts
0.50	93.55	77.42	0.73	64.52	38.71
0.51	93.55	77.42	0.74	64.52	38.71
0.52	93.55	77.42	0.75	64.52	38.71
0.53	93.55	77.42	0.76	64.52	38.71
0.54	93.55	77.42	0.77	64.52	38.71
0.55	93.55	77.42	0.78	61.29	38.71
0.56	90.32	74.19	0.79	54.84	35.48
0.57	90.32	74.19	0.80	48.39	35.48
0.58	90.32	74.19	0.81	41.94	35.48
0.59	90.32	74.19	0.82	41.94	35.48
0.60	90.32	70.97	0.83	38.71	35.48
0.61	90.32	70.97	0.84	38.71	35.48
0.62	93.55	70.97	0.85	38.71	32.26
0.63	93.55	67.74	0.86	38.71	29.03
0.64	90.32	64.52	0.87	35.48	29.03
0.65	87.1	64.52	0.88	32.26	29.03
0.66	83.87	58.06	0.89	32.26	29.03
0.67	83.87	58.06	0.90	32.26	25.81
0.68	80.65	58.06	0.91	29.03	25.81
0.69	80.65	51.61	0.92	29.03	22.58
0.70	77.42	45.16	0.93	29.03	22.58
0.71	74.19	45.16	0.94	25.81	22.58
0.72	67.74	41.94	0.95	25.81	22.58

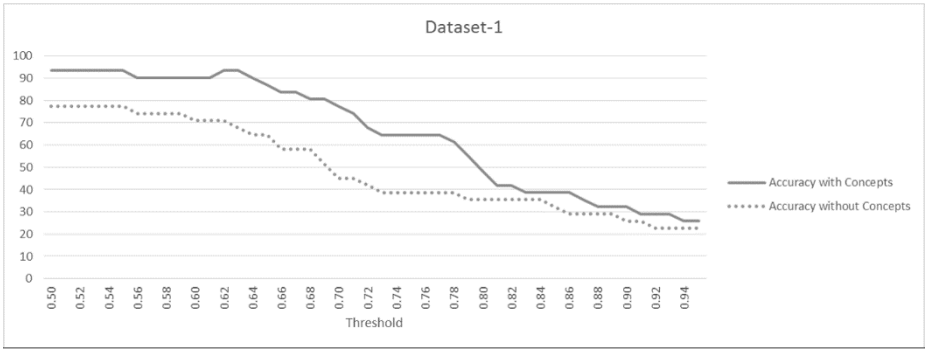

Fig. 1. The curves comparing the performance on Dataset-1 with and without using concept similarity

Table 2. The overall accuray of our method on Datatset-2

τ	Accuracy with Concepts	Accuracy without Concepts	τ	Accuracy with Concepts	Accuracy without Concepts
0.50	86.67	73.33	0.73	64.44	46.67
0.51	86.67	73.33	0.74	64.44	46.67
0.52	86.67	73.33	0.75	64.44	46.67
0.53	86.67	73.33	0.76	66.67	46.67
0.54	86.67	73.33	0.77	66.67	44.44
0.55	86.67	73.33	0.78	62.22	44.44
0.56	84.44	71.11	0.79	57.78	44.44
0.57	84.44	71.11	0.80	55.56	44.44
0.58	84.44	71.11	0.81	51.11	44.44
0.59	84.44	71.11	0.82	48.89	44.44
0.60	84.44	68.89	0.83	46.67	44.44
0.61	84.44	68.89	0.84	46.67	44.44
0.62	86.67	68.89	0.85	44.44	42.22
0.63	86.67	68.89	0.86	46.67	40.00
0.64	82.22	66.67	0.87	44.44	40
0.65	80	64.44	0.88	42.22	40
0.66	77.78	60	0.89	42.22	40
0.67	77.78	60	0.90	42.22	37.78
0.68	75.56	60	0.91	40	37.78
0.69	75.56	55.56	0.92	40	33.33
0.70	73.33	53.33	0.93	40	33.33
0.71	71.11	51.11	0.94	37.78	33.33
0.72	66.67	48.89	0.95	37.78	33.33

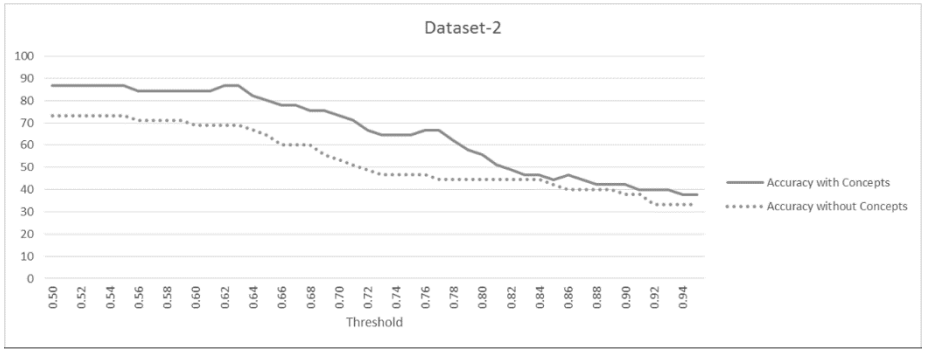

Fig. 2. The curves comparing the performance on Dataset-2 with and without using concept similarity

5 Conclusion

In this paper, we present a novel method for measuring semantic similarity of two sentences. The overall semantic similarity is a linear combination of word-to-word similarity, word-order similarity, and concept similarity. The method exploits Wiktionary to measure similarity of words, exploits Wikipedia to identify which concepts referred to in the compared sentences existing in Wikipedia for expanding the contexts of those sentences by different surface forms of the concepts in order to compute the concept similarity. We evaluate the method on the datasets consisting of Vietnamese sentence pairs. The results show that in the best cases, concept similarity help improving the performance of our method more than 15% point. The proposed method is language-independent and quite easy to employ. Therefore, one can readily adopt our method to measure semantic similarity for sentences written in other languages.

References

1. Mihalcea, R., Corley, C., & Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In AAAI , Vol. 6, pp. 775-780 (2006)
2. Islam, A., & Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. In: ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), Article 10 (2008).
3. Sahami, M., & Heilman, T. D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international conference on World Wide Web, pp. 377-386 (2006)
4. Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. In: IEEE Transactions on Knowledge and Data Engineering, 18(8), 1138-1150 (2006)
5. Oliva, J., Serrano, J. I., del Castillo, M. D., & Iglesias, Á.: SyMSS: A syntax-based measure for short-text semantic similarity. In: Data & Knowledge Engineering, 70(4), pp. 390-405 (2011)

6. Bach, N. X., Minh, N. L., & Shimazu, A.: Exploiting discourse information to identify paraphrases. In: Expert Systems with Applications, 41(6), 2832-2841 (2014).
7. Madnani, N., Tetreault, J., and Chodorow, M.: Re-examining Machine Translation Metrics for Paraphrase Identification. In: Proceedings of 2012 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2012), pp. 182-190 (2012)
8. Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., & Manning, C. D.: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *NIPS* (Vol. 24, pp. 801-809) (2011)
9. Fernando, S., & Stevenson, M.: A semantic similarity approach to paraphrase detection. In: Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics (pp. 45-52) (2008).
10. Das, D., and Smith, N.: Paraphrase identification as probabilistic quasi-synchronous recognition. Ing: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 468-476, (2009)
11. Qiu, L. and Kan, M.Y. and Chua, T.S.: Paraphrase recognition via dissimilarity significance classification. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 18-26 (2006)
12. Rus, V. and McCarthy, P.M. and Lintean, M.C. and McNamara, D.S. and Graesser, A.C.: Paraphrase identification with lexico-syntactic graph subsumption. In: FLAIRS 2008, pp. 201-206.
13. Lee, M. C.: A novel sentence similarity measure for semantic-based expert systems. In: Expert Systems with Applications, 38(5), 6392-6399 (2011)
14. Wenyin, L., Quan, X., Feng, M., & Qiu, B.: A short text modeling method combining semantic and statistical information. In: Information Sciences, 180(20), 4031-4041 (2010)
15. Blacoe, W., & Lapata, M.: A comparison of vector-based representations for semantic composition. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 546-556
16. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on Systems documentation, pp. 24-26 (1986)
17. Tsatsaronis, G., Varlamis, I., & Vazirgiannis, M.: Text relatedness based on a word thesaurus. Journal of Artificial Intelligence Research, 37(1), 1-40 (2010)
18. Rubenstein, H., & Goodenough, J. B.: Contextual correlates of synonymy. Communications of the ACM, 8(10), 627-633 (1965)
19. Huynh, H. M., Nguyen, T. T., & Cao, T. H.: Using coreference and surrounding contexts for entity linking. In 2013 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF 2013), pp. 1-5, (2013)