Recognizing Textual Entailment for Arabic using semantic similarity and Word Sense Disambiguation

Mabrouka Ben-sghaier[1] Wided Bakari [1, 3] Mahmoud Neji[1, 3]
(1) Faculty of Economics and Management, 3018, Sfax Tunisia
(3) MIR@CL, Sfax, Tunisia
mabrouka.bensghaier@gmail.com, {wided.bakkari, mahmoud.neji}@fsegs.rnu.tn

Abstract. In this paper, we propose a semantic method for recognizing textual entailment in Arabic. The proposed method serves to detect the directional semantic entailment relationship between text/hypothesis pairs. More specifically, we work at the sentence level, conducting semantic similarity measure and word sense disambiguation process in order to detect entailment relationship in the context of Arabic question/answering system. The results obtained are encouraging. Our method has reached an accuracy of 70%.

Keywords: Recognizing Textual entailment, Semantic similarity, Word sense disambiguation, Arabic language

1 Introduction

In recent years, there has been large interest in Arabic Natural Language Processing (NLP) applications; due to the importance of Arabic that is the sixth most spoken language in the world. However, most of the existents NLP applications have concentrated on English. Particularly, for the Recognizing Textual Entailment (RTE) task, Arabic has relatively fewer studies and even many existing approaches may even be inapplicable. One of the reasons is that the Arabic language is one of the most morphologically complex languages since it is an inflected and derivational Semitic language. The lack of the voyellation in Arabic texts is also a big source of ambiguity. However, the majority of written texts are not voyelled. On the other hand, Arabic lacks semantic and world knowledge resources.

Recently, the RTE task has attracted considerable attention. Given two text fragments, the task of RTE enables to determine whether the meaning of one text could be reasonably inferred, or textually entailed, from the meaning of the other one. The task of the RTE becomes fundamental to many applications in NLP; it ensures better performance in multiple NLP applications, such as, machine translation, information retrieval, information extraction, automatic summary, question/answering, etc. Thus, RTE helps to consolidate and promote research on the semantic processing of the natural language and to lay a generic basis for developing these applications [11].

In particular, the relation between an asked question and its answers can be transformed in terms of textual entailment. As it is mentioned in [21], systems developed for the RTE task can provide Q/A systems with valuable semantic information in order to identify exact answers from a list of candidate answers.

In this paper, we propose a semantic method for the RTE in Arabic which determines the relation of entailment between a question and its candidate answers with the use of semantic similarity between sentences and the resolution of word sense disambiguation (WSD). More particularly, the proposed system employs the Simplified Lesk algorithm for the WSD process, employs a dictionary in Arabic to select the best word senses of ambiguous words, and the Arabic WordNet (AWN)[1] thesaurus [27] in order to calculate similarity measures.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 gives an overview of the proposed semantic method for the RTE problem in Arabic. While section 4 details the experiment setup and the obtained results. Finally, section 5 summarizes our conclusion.

2 Related works

RTE is an important problem in NLP, used to determine the entailment relationship (true or false entailment) between a text T and a hypothesis H [22]. The RTE includes the task of determining the semantic entailment between a pair of sentences. A fragment of text entails another if the meaning of the latter can be deduced from that of the first fragment. The main input to an RTE system is a pair of text fragments, possibly in a particular context. The desired output is a judgment that indicates whether these sentences are a pair of textual entailment or not [23].

Since 2005, several challenges have been proposed for English text. There have been eight challenges [8] to the RTE task organized between 2005 and 2013. Also, SemEval 2014 [15], and more recently, RepEval 2017 [17] meant to evaluate the understanding of natural language models on the RTE task. These challenges allowed researchers to compare their work and learn as a research community. They provide common test collections, a common assessment procedure, and means of sharing and discussing the work of researchers. They were responsible for stimulating the research community to work on these research lines.

In Arabic language, the RTE task has few studies. First, authors of [1] developed the ArbTE system in order to evaluate existing RTE techniques when applied to Arabic. Then, authors of [2] described a semi-automatic technique for creating a first dataset for RTE systems in Arabic. Subsequent work proposed the use of extended tree modification distance with sub-trees [3]. Others have closely examined negation and polarity as additional characteristics [4]. In addition, authors of [5] presented a method based on using a semantic and lexical combination. More recently, authors of [6] used features based mainly on distributional representations with the use of word2vec model.

It appears that no work has addressed the RTE task for factual Q/A systems. Besides, there is no research addressing the issue of RTE in Arabic by using semantic similari-

[1] http://globalwordnet.org/arabic-wordnet/

ty measure and with resolving the WSD problem. In fact, there are various measures developed previously in order to quantify how two words are semantically related. In this context, we are going to focus on semantic measures using AWN because regarding the RTE task we observed the common use of WordNet since it is one of the primarily semantic resources used. Besides, we perform the WSD problem by Simplified Lesk algorithm using an Arabic Dictionary, since the AWN does not contain definitions expressed in Arabic language that is needed to calculate the similarity measure based on arabic words. WSD is a well-studied problem, where many approaches have been applied. The main cause of ambiguity of words is the lack of diacritics in the most widely digital documentation available in Arabic, so that the same word can appear with different meanings [10].

We propose in this paper a method for the RTE issue using semantic similarity and WSD resolution, based respectively on AWN and an electronic Dictionary in Arabic for the purpose of establishing the impact of the proposed method on the RTE task.

3 Proposed method

In this section, we describe the proposed method to address the RTE task for Arabic sentence pairs. This method is described below in figure 1; it consists of four main steps. The first step applies a preprocessing of the text T and the hypothesis H. The next step consists on measuring the local similarity between each word of H comparing to all words of T using the Wu and Palmer (Wup) similarity measure via AWN and resolving WSD with the Simplified Lesk algorithm by employing an Arabic dictionary. The third step is dedicated to detect the global similarity and moving from the step of similarity between words to the similarity between the pair T/H (text/hypothesis). Finally, the last step is to determine the entailment relation based on semantic similarity between the text and the hypothesis, with using a machine learning algorithm.

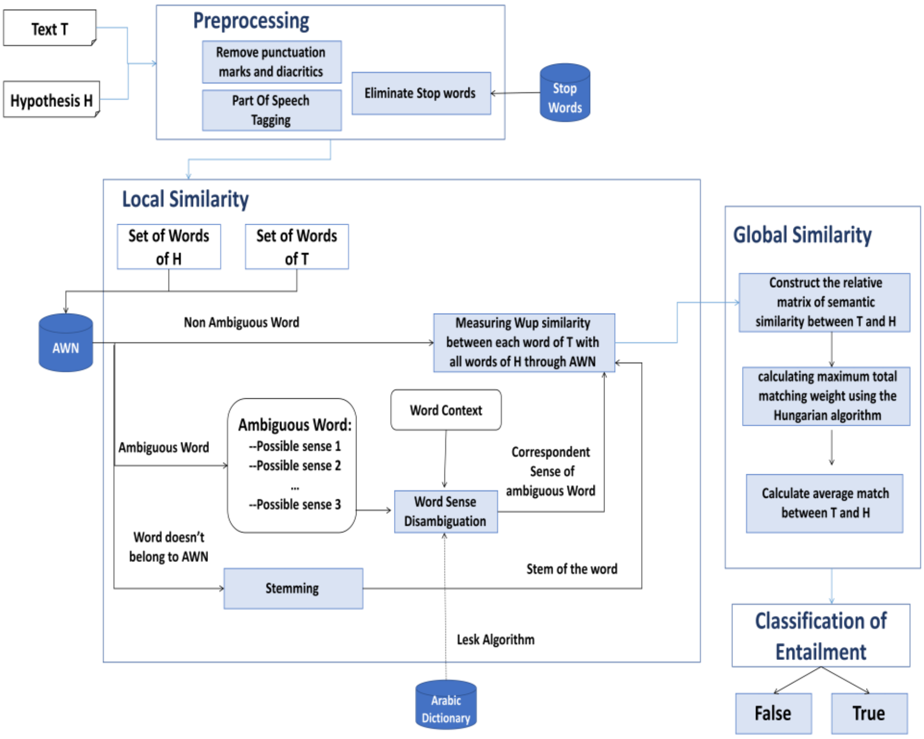
Text T
Hypothesis H
Preprocessing
Remove punctuation marks and diacritics
Part Of Speech Tagging
Eliminate Stop words
Stop Words
Local Similarity
Set of Words of H
Set of Words of T
AWN
Non Ambiguous Word
Measuring Wup similarity between each word of T with all words of H through AWN
Ambiguous Word
Ambiguous Word:
--Possible sense 1
--Possible sense 2
...
--Possible sense 3
Word Context
Correspondent Sense of ambiguous Word
Word Sense Disambiguation
Word doesn't belong to AWN
Stemming
Stem of the word
Lesk Algorithm
Arabic Dictionary
Global Similarity
Construct the relative matrix of semantic similarity between T and H
calculating maximum total matching weight using the Hungarian algorithm
Calculate average match between T and H
Classification of Entailment
False
True

Fig. 1. Process of the proposed method

3.1 Preprocessing

As we are focusing in this work on a particular type of question, namely factual questions, we transform the question into a declarative form by mentioning the type of expected answer. Then, the preprocessing of the text and hypothesis consists of three steps: The first is a removal of punctuation marks and diacritics (if exist). Second step is Part-Of-Speech (PoS) Tagging which assigns a syntactic role for each token in the sentence. In fact, PoS tagging allows the categorization of words in verbs and names allowing consequently the study of names and verbs in a separate way. Third step is eliminating the stop words, which will have a positive impact in the obtained results by focusing the attention on words that may point to entailment relations.

3.2 Local similarity

Studying the semantic similarity has been a part of computational linguistics for many years and the measures of semantic similarity have been employed previously in many NLP applications [14]. In this work, we look for words similarity measures from the AWN thesaurus which is a linguistic resource for Modern Standard Arabic (MSA). It groups Arabic words into sets of synonyms called synsets, where each word can be a part of one or more synsets. In addition, the AWN records the different semantic synsets relations. It is therefore a lexical network where the nodes are the synsets and the relations between the synsets are the edges.

The semantic similarity measures calculate how much two concepts are similar, based on information obtained from hierarchical taxonomy. For example, an automobile may be considered more like a boat than a tree, if the car and boat share a vehicle as a common ancestor in the taxonomy [18]. This characteristic is based on semantic distance and provides a score illustrating the similarity between T/H pair. In recent years, similarity methods based on WordNet have shown their talents and raised great concerns [13]. There are many measures that use a lexical database, such as, WordNet to calculate similarity between English concepts. However, few studies have studied semantic similarity measures using AWN. Experimental results from the application of traditional semantic similarity measurements on AWN revealed that the Wup measure has the highest correlation value with human ratings [16]. This method calculates the similarity by considering the depths of two concepts in the WordNet hierarchy (similarly AWN), as well as the depth of the smaller sub-segment in common. This semantic similarity is calculated as follows:

$$\mathbf{sim}_{\mathbf{WP}}\left(\mathbf{C_i}, \mathbf{C_j}\right) = \frac{\mathbf{2*depth}\left(\mathbf{lCS}\left(\mathbf{C_i},\mathbf{C_j}\right)\right)}{\mathbf{depth}\left(\mathbf{C_i}\right)+\ \mathbf{depth}\left(\mathbf{C_j}\right)} \tag{1}$$

Where depth (C) is the depth of the synset C using the counting of edges in the taxonomy, LCS (C1, C2) is the smallest common sub-segment of C1 and C2. The depth (LCS (C1, C2)) is the length between LCS of C1 and C2 and the root of the taxonomy. So, we first take into account the measure of similarity between each word of H with all the words of T via AWN using Wup measure [20].

So, we determine the wup measure between each word of the text and all words of the hypothesis (For example, the wup similarity through AWN between “قرية” and “ريف” is equal to 0,18) . Besides, we consider the synonyms in the same synset as the same concept (e.g “مستعمل”and “مستخدم“belong to the same synset). In the other side, the stemming process will be performed for the words that do not belong to the thesaurus AWN, since in some cases, a word doesn’t belong to AWN but its stem does.

However, when we intend to look for the measure of semantic similarity between two concepts, we find that a single word can have multiple meanings. In this case it is called an ambiguous word, so it is indispensable to determine the appropriate meaning of each word. Thus, disambiguation becomes an important task in order to remove the ambiguity of the words in question.

3.3 Word Sense Disambiguation

Each word from T or H can belong to one or more senses. This will lead to ambiguity in the analysis of its content. Humans implicitly disambiguate words by matching the word in context with meanings and experiences stored in memory. But, this is not a pretty easy task for the machine. The task of WSD makes it possible to identify the correct meaning of an ambiguous word in a given context. It is a fundamental task in NLP which aims at automatically identifying the correct sense of a given ambiguous word from a set of predefined senses. In WSD, the goal is to tag each ambiguous word in a text with one of the senses known a priori. The main cause of the ambiguity of Arabic words is the lack of diacritics of the most digital documents so the same word can occur with different senses [10].

We adopt in our work for the task of WSD a knowledge based approach. It is an approach based on different knowledge sources as dictionaries, thesauruses and lexicons. This technique is applied to make use of one or more sources of knowledge to associate the most appropriate senses with words in context.

The numerical equivalent for a priori knowledge most used for English is WordNet, where the fundamental construction is not a word but an abstract semantic concept. Each concept (or synset) in WordNet can be expressed by different words and, conversely, the same word can represent different concepts. Nevertheless, Arabic presents several challenges for WSD, due mainly to the particularity of this language and also to the lack of resources needed for the disambiguation process [7]. For example, the AWN does not provide word definitions as does WordNet. So, we solved this problem by applying the simplified Lesk algorithm [12] based on the knowledge given by the Arabic dictionary "Intermediate Lexicon (المعجم الوسيط)" available in SAFAR platform[2]. The Intermediate Lexicon contains the different definitions of Arabic words indicated as ambiguous words, and represents, in our work, a disambiguation resource. The simplified Lesk algorithm is a well-known method of disambiguation that consists in counting the number of common words between the definitions of a word and the definitions of the words of its context. Our WSD process has been accomplished by performing the following steps (Figure 2):

- Step 1: Determine all the candidate senses of the word to disambiguate from AWN and order them in descending order according to their frequencies.
- Step 2: Extract all the definitions from the electronic dictionary for each meaning of the word.
- Step 3: Apply the simplified Lesk algorithm and compare the definitions of each sense with those of the words of its context in order to extract the appropriate meaning.

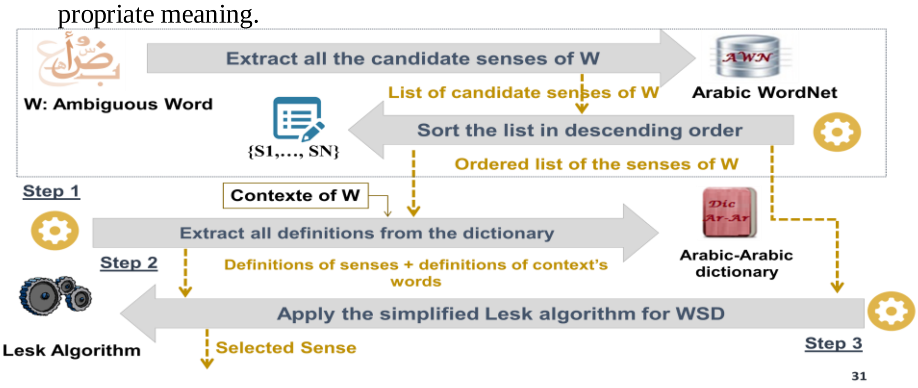
Extract all the candidate senses of W
AWN
Arabic WordNet
W: Ambiguous Word
List of candidate senses of W
Sort the list in descending order
{S1,…, SN}
Ordered list of the senses of W
Step 1
Contexte of W
Dic Ar-Ar
Extract all definitions from the dictionary
Arabic-Arabic dictionary
Step 2
Definitions of senses + definitions of context's words
Apply the simplified Lesk algorithm for WSD
Lesk Algorithm
Selected Sense
Step 3
31

Fig. 2. Word Sense Disambiguation Process

[2] http://arabic.emi.ac.ma/safar/

After disambiguating every ambiguous word, we determine the local similarity between their correspondent synsets.

3.4 Global similarity

In this section, we determine the overall similarity between T and H. In order to look for global similarity, we will move from the level of similarity between words to the similarity between H and T. More precisely, we will decide whether the two sentences are semantically related or not and thus deduce the entailment between them based on the semantic similarity between the meanings of the words. Indeed, there are many strategies to acquire a global similarity of two sets. The overall strategy that we used is the corresponding average strategy. So, we denote m for the length of H, n for length of T. The main steps can be described as follows:

- Construct a relative matrix of semantic similarity R [m, n] of each pair of meanings of extracted words, where R [i, j] is the semantic similarity between the word at position i of H and the word at position j of T.
- The similarity between H and T is reduced to the problem of calculating a total maximum matching weight of a bipartite graph, performed using the Hungarian algorithm on this bipartite graph where X and Y are H and T, and the Graph nodes are related words [9].
- Matching results from the previous step are combined into a single value of similarity calculated as follows [26]:

$$\mathbf{Global}_{\mathbf{similarity}(\mathbf{H},\mathbf{T})} = \frac{\mathbf{2*SimCorresp(H,T)}}{|\mathbf{H}|+|\mathbf{T}|} \tag{2}$$

Where SimCorresp (H, T) is the value of the word matches of H and T. This similarity is calculated by dividing the sum of the similarity values of all the corresponding words candidates of T/H pair by the total number of words. An important point is that this score is based on each of the individual similarity values, so the overall similarity always reflects the influence of these ones.

3.5 Entailment classification

The entailment classification step consists in attributing for each T/H couple the appropriate entailment decision. State-of-the-art systems for RTE in natural language text typically follow a supervised machine learning approach [19]. So, the problem of entailment can be simply considered as a classification problem for classifying a given pair of sentences as a true or false entailment.

4 Experiments and Results

The stemming is performed using Khoja Stemmer [24] which is one of the known and widely used Arabic stemmers. In addition, the Arabic tagger model of Stanford[3] is chosen for the PoS tagging due to its availability and the availability of its documentation. Besides, the Wup similarity between words is calculated using the AWN

[3] https://nlp.stanford.edu/software/tagger.shtml

through the “Java AWN API[4]” tool. The entailment classification is based on a machine learning approach for RTE. Thus, a dataset is necessary for the training and the test of our proposed method. Previous Arabic RTE works used for this purpose the dataset proposed by [2]. However, in our work we search for entailment relation between pairs of factual question and its candidate answers. Thus, we have to make our own experimental data using a different set of T/H pairs. Therefore, we have used a dataset consisting of 200 T/H pairs of factual questions and corresponding answers to different domains, recovered from the AQA-WebCorp corpus presented in [25]. To build a classifier, the system must be trained using a development set, and to validate system performance, it must be tested using a set of tests. Therefore, we have partitioned this dataset into a learning set that includes 70% of the dataset (200 T/H pair) and a test set including 30% (50 T/H pair). Then, we annotated each T/H of the development set by hand according to their entailment, either “True” or “False”. In table 1 below, we present two T/H pairs from the test set and we report the extracted sentence similarity and the entailment decision between T and H.

Table 1 Examples of entailment results

Sentence Pair	Sentence similarity	Entailment Decision

H= “كم يبلغ عدد مستخدمي الإنترنت في تونس سنة 2016”
T= “ يبلغ عدد مستعملي الإنترنت في تونس خلال سنة 2016 حوالي 5472618 بنسبة نفاذ قدرها 48,1 %.”
	0,35	True

H= “كم يبلغ عدد مستخدمي الإنترنت في تونس سنة 2016”
T= “ ذكر تقرير جديد صادر عن لجنة النطاق العريض التابعة للأمم المتحدة أن عدد مستخدمي الإنترنت عالمياً سيصل إلى 3,5 مليار شخص يُمثل 47% من إجمالي سكان بحلول نهاية عام 2016 الحالي، وهو ما العالم”
	0,04	False

From the 50 T/H pair of the test set, 35 entailments relationships are correctly recognized. The performance of an RTE system can be measured by calculating the accuracy of the number of correct textual entailments on the number of tested entailments. Therefore, the accuracy would be defined as follows:

$$\textbf{Accuracy} = \frac{\textbf{Number of correctly recognized entailments}}{\textbf{Total number of tested entailments}} = \frac{35}{50} = 0.7 \quad (3)$$

5 Conclusion

In this paper, we proposed a semantic method for the task of RTE in Arabic for factual Question/Answering system. The proposed method consists of finding the semantic distance between each word of the text and all words of the hypothesis, by employing

[4] https://sourceforge.net/projects/javasourcecodeapiarabicwordnet/

the Wup measure as it has the best performance on AWN compared to other measurements. Besides, we employed WSD process in order to find the appropriate sense of the ambiguous words. For this purpose, we have employed the Simplified Lesk algorithm and extracted senses definitions from an Arabic dictionary. For a sample of 50 test set of questions and answers on different areas, experiments have shown a precision of 70%. In order to increase the accuracy of classification, we plan to develop our data set, extract more features from the T/H pair and take into consideration the type of the searched named entity.

References

1. Alabbas Maytham. ArbTE: Arabic textual entailment. In : Proceedings of the Second Student Research Workshop associated with RANLP 2011. 2011. p. 48-53.
2. Alabbas, Maytham. "A Dataset for Arabic Textual Entailment." Proceedings of the Student Research Workshop associated with RANLP 2013. 2013.
3. Alabbas Maytham et Ramsay Allan. Natural language inference for Arabic using extended tree edit distance with subtrees. Journal of Artificial Intelligence Research, 2013, vol. 48, p. 1-22
4. Al-Khawaldeh Fatima T. A Study of the Effect of Resolving Negation and Sentiment Analysis in Recognizing Text Entailment for Arabic. World of Computer Science & Information Technology Journal, 2015, vol. 5, no 7.
5. Khader Mariam, Awajan Arafat, et Alkouz, Akram. Textual Entailment for Arabic Language based on Lexical and Semantic Matching. International Journal of Computing & Information Sciences, 2016, vol. 12, no 1, p. 67.
6. Almarwani Nada et Diab Mona. Arabic Textual Entailment with Word Embeddings. In : Proceedings of the Third Arabic Natural Language Processing Workshop. 2017. p. 185-190.
7. Bouhriz Nadia et Benabbou Faouzia. Word sense disambiguation approach for Arabic text. Int. J. Adv. Compt. Sci. Appl, 2016, vol. 1, no 7, p. 381-385
8. Bentivogli, L.; Dagan, I.; Dang,H.T.; Giampiccolo, D.;Magnini, B. Fifth PASCAL Recognizing Textual Entailment Challenge. In Proceedings of the Text Analysis Conference, Gaithersburg,MD, USA, 16–17 November 2009
9. Dao Thanh Ngoc et Simpson Troy. Measuring similarity between sentences. The Code Project, 2005
10. Elayeb, Bilel. "Arabic word sense disambiguation: a review." Artificial Intelligence Review (2018): 1-58
11. Korman Daniel Z., et al. "Defining textual entailment." Journal of the Association for Information Science and Technology 69.6, 2018.
12. Kilgarriff, Adam et Rosenzweig, Joseph. Framework and results for English Senseval. Computers and the Humanities, 2000, vol. 34, no 1-2, p. 15-48.
13. Lingling Meng et Junzhong GU. A New Method for Calculating Word Sense Similarity in WordNet1. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2012, vol. 5, no 3, p. 197-206.
14. Majumder Goutam, Pakray Partha, Gelbukh Alexander, et al. Semantic textual similarity methods, tools, and applications: A survey. Computación y Sistemas, 2016, vol. 20, no 4, p. 647-665.
15. Marelli, M.; Bentivogli, L.; Baroni, M.; Bernardi, R.; Menini, S.; Zamparelli, R. SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sen-

tences through Semantic Relatedness and Textual Entailment. In Proceedings of the 8th International Workshop on Semantic Evaluation, COLING, Dublin, reland, 23–24 August 2014; Nakov, P., Zesch, T., Eds.; ACL: Vancouver, BC, Canada, 2014; pp. 1–8.

16. Nababteh Mohammed et Deri Mohammed. Experimental Study of Semantic Similarity Measures on Arabic WordNet. International Journal of Computer Science and Network Security (IJCSNS), 2017, vol. 17, no 2, p. 131.
17. Nangia, N.; Williams, A.; Lazaridou, A.; Bowman, S.R. The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations. arXiv 2017, arXiv:1707.08172.
18. Pedersen, Ted, Patwardhan, Siddharth, et Michelizzi, Jason. WordNet:: Similarity: measuring the relatedness of concepts. In : Demonstration papers at HLT-NAACL 2004. Association for Computational Linguistics, 2004. p. 38-41.
19. Rocha, Gil, and Henrique Lopes Cardoso. "Recognizing Textual Entailment: Challenges in the Portuguese Language." Information 9.4. 2018.
20. WU Zhibiao et Palmer Martha. Verbs semantics and lexical selection. In : Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994. p. 133-138.
21. Ben-Sghaier, Mabrouka, Wided Bakari, and Mahmoud Neji. "An Arabic Question-Answering System Combining a Semantic and Logical Representation of Texts." International Conference on Intelligent Systems Design and Applications. Springer, Cham, 2017.
22. CHEN, Qian, ZHU, Xiaodan, LING, Zhen-Hua, et al.Neural natural language inference models enhanced with external knowledge. In : Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018. p. 2406-2417.
23. ROCHA, Gil et LOPES CARDOSO, Henrique. Recognizing Textual Entailment: Challenges in the Portuguese Language. Information, 2018, vol. 9, no 4, p. 76.
24. LARKEY, Leah S. et CONNELL, Margaret E. Arabic information retrieval at UMass in TREC-10. In : TREC. 2001.
25. Bakari Wided, Bellot Patrice, et Neji Mahmoud. AQA-WebCorp: Web-based factual questions for Arabic. Procedia Computer Science, 2016, vol. 96, p. 275-284.
26. Castillo, Julio J. Recognizing textual entailment: experiments with machine learning algorithms and RTE corpora. Special issue: Natural Language Processing and its Applications, 2010, p. 155.
27. BLACK, William, ELKATEB, Sabri, RODRIGUEZ, Horacio, et al. Introducing the Arabic wordnet project. In : Proceedings of the third international WordNet conference. 2006. p. 295-300.