Vietnam J Comput Sci (2017) 4:51–60
DOI 10.1007/s40595-016-0080-2

REGULAR PAPER

Enhancing the sentence similarity measure by semantic and syntactico-semantic knowledge

Wafa Wali[1] · Bilel Gargouri[1] · Abdelmajid Ben Hamadou[2]

Received: 27 April 2016 / Accepted: 29 August 2016 / Published online: 16 September 2016

Abstract The measure of sentence similarity is useful in various research fields, such as artificial intelligence, knowledge management, and information retrieval. Several methods have been proposed to measure the sentence similarity based on syntactic and/or semantic knowledge. Most proposals are evaluated on English sentences where the accuracy can decrease when these proposals are applied to other languages. Moreover, the results of these methods are unsatisfactory, as much relevant semantic knowledge, such as semantic class, thematic role and syntactico-semantic knowledge like the semantic predicates, are not taken into account. We must acknowledge that this kind of knowledge is rare in most of the lexical resources. Recently, the International Organization for Standardization (ISO) has published the Lexical Markup Framework (LMF) ISO-24613 norm for the development of lexical resources. This norm provides, for each meaning of a lexical entry, all the semantic and syntactico-semantic knowledge in a fine structure. Profiting from the availability of LMF-standardized dictionaries, we propose, in this paper, a generic method that enhances the measure of sentence similarity by applying semantic and syntactico-semantic knowledge. An experiment was carried out on Arabic, as this language is processed within our research team and an LMF-standardized Arabic dictionary is at hand where the semantic and the syntactico-semantic knowledge are accessible and well structured. Moreover, the experiments yielded better results, showing a high correlation with human ratings.

Keywords Sentence similarity · Lexical semantic knowledge · Syntactico-semantic knowledge · LMF-ISO 24613 · Standardized dictionaries

1 Introduction

The issue of measuring similarity between sentences is crucial in some research fields, such as knowledge management, information retrieval and artificial intelligence. Computing sentence similarity is not a trivial task due to the variability of natural language expressions. In the last few years, sentence similarity measure has been increasingly in demand from a variety of applications and numerous achievements which have been carried out recently in this area and classified into three categories: statistical-based methods, semantic-based methods and hybrid methods. Initially, researchers started with statistical-based methods such as [2] and [14]. These methods compute the sentence similarity by calculating the co-occurring words in a string sequence. However, these methods may not fit to sentences as they may be very similar while co-occurring words are infrequent. To overcome this drawback, other authors proposed the semantic-based methods, such as [12] and [14]. These approaches used the semantic nets, like the WordNet, the vector space model and the statistical corpus to compute the semantic similarity between words using different known measures, such as Leacock and Chodorow [11], Wu and Palmer [22] and Jiang and Conrath [8]. Nevertheless, these semantic-based methods are limited to computing the sentence similarity based only on semantic similarity between words, whereas the syntactic

✉ Wafa Wali
wafa.wali@fsegs.rnu.tn

Bilel Gargouri
bilel.gargouri@fsegs.rnu.tn

Abdelmajid Ben Hamadou
abdelmajid.benhamadou@isimsf.rnu.tn

[1] MIR@CL Laboratory FSEGS, Sfax, Tunisia

[2] MIR@CL Laboratory ISIMS, Sfax, Tunisia

information and other semantic knowledge, such as semantic class and thematic roles, are missing. To address this weakness, other researchers proposed hybrid methods to compute sentence similarity taking into account both semantic and syntactic knowledge, such as [6,13] and [19]. However, these hybrid methods may have some disadvantages, such that the semantic measurement is isolated from the syntactic measurement in which the semantic similarity is calculated based on word semantic similarity, while string matching, word order and word co-occurring are counted to compute the syntactic similarity. Moreover, the results of these proposals are far away from the aims of a human expert and are evaluated on English databases where accuracy can decrease if they are applied to other languages. Furthermore, some knowledge is not considered in measuring sentence similarities, such as the semantic class, the thematic role and the relationship between syntactic and semantic levels through the semantic predicates. Indeed, when two sentences have the same syntactic structure (subject verb object) but the semantic classes of these objects are dissimilar (i.e., the first is human and the second is vegetal), the pair of these sentences is syntactically similar according to these hybrid methods, whereas in reality both sentences are totally different according to an expert. The semantic class and the thematic role for each argument of a sentence provide knowledge about the relationships between words and perform a role in conveying the meaning of sentences. Besides, the syntactico-semantic knowledge supplied a mechanism for the interaction between the syntactic processor, the discourse model and the real-world knowledge. Furthermore, the semantic predicates favor the creation of coherence in the local discourse structure. Thus, incorporating the semantic and syntactico-semantic knowledge in computation sentence similarity improves the quality of the sentence similarity measure. Unfortunately, there is a lack of linguistic resources that provide such relevant knowledge. Few years ago, the technical committee ISO TC37/SC4 of the International Organization for Standardization (ISO) published the Lexical Markup Framework (LMF) ISO-24613 norm [3]. This norm promotes the construction of large lexical resources in a fine and modular structure. In particular, it provides for each meaning of a lexical entry, the whole semantic and syntactico-semantic knowledge. In this paper, we propose profiting from the available knowledge in LMF standardized dictionaries, notably the semantic and syntactico-semantic knowledge, to enhance the measure of sentence similarity. Our proposal consists of a hybrid method that can be applied to all natural languages. The proposed method is an extension of the existing ones. It measures the semantic similarity via the synonymy relations between words in sentences. Besides, the syntactico-semantic similarity is measured based on the common semantic arguments that are associated with semantic predicates in terms of the thematic role and the semantic class. An experiment was carried out on Arabic, because this language is dealt with within our research team and the existence of an LMF standardized Arabic dictionary in which semantic and syntactico-semantic knowledge can be accessed and well structured. Due to the lack of Arabic suitable benchmarks for the evaluation of sentence similarity, we assess the outcome of our proposal using 690 pairs of Arabic sentences extracted from various definitions and examples of Arabic dictionaries of human use, such as Alwasit, AlMuhit, Lissan Al Arab and Tj-Al-Arous. The results demonstrate that our proposal presents a good performance that approximates to human intuitions.

This paper is organized as follows. First, we present an overview of the existing methods of similarity measures. Section 3 presents the main features of the LMF standard. Our proposed method is described in Sect. 4. Section 5 reports on the experiments and the obtained results. The final section presents the conclusion and recommendations for future works.

2 State of the art

2.1 Overview on the sentences similarity methods

There is extensive literature on measuring the similarity between sentences, which can be grouped into three categories: syntactic-based methods, semantic-based methods, and hybrid methods. In this section, we report only hybrid methods to explore their advantages and limitations.

Li et al. [13] defined a sentence similarity measure as a linear combination of semantic vector similarity and word order similarity. Their proposed method dynamically forms a joint word set only by using all the distinct words in the pairs of sentences. For each sentence, a raw semantic vector is derived with the assistance of the WordNet lexical database [15]. Moreover, a word order vector is formed for each sentence. Since each word in a sentence contributes differently to the meaning of the whole sentence, the significance of a word is weighted by using information content derived from a corpus. By combining the raw semantic vector with information content from the corpus, a semantic vector is obtained for each of the two sentences. Semantic similarity is computed based on the two semantic vectors. An order of similarity is calculated using the two order vectors. Finally, the sentence similarity is derived by combining semantic similarity and order similarity. The relative contribution of semantic and syntactic measures is controlled by an alpha coefficient. It has been empirically proved that a sentence similarity measure performs better when the semantic measure is weighted more than the syntactic one.

Islam and Inkpen [6] determined the similarity between two sentences from semantic and syntactic information (in terms of common word order) that they contain. Indeed, the string similarity is computed using a normalized and modified version of the longest common subsequence (LCS) string matching algorithm. The authors used the longest common subsequence (LCS) measure with some normalization and small modifications for their string similarity measure. They used three different modified versions of LCS. Finally, the sentence similarity is derived by combining the string similarity, the semantic similarity and the common word order. This measure has several inadequacies, particularly, the time complexity of the string matching.

Lee et al. [12] introduced an algorithm to compute the similarity between sentences using semantic and syntactic relationships derived from natural languages. The algorithm proposed a semantic model using the word similarity based on the WordNet and the grammatical rules taking advantage of the Stanford parser.

In addition, [19] proposed the method for computing sentence semantic similarity by exploiting a set of its characteristics, namely features-based measure of sentences semantic similarity (FM3S). The proposed method aggregates in a non-linear function between three components: the noun-based semantic similarity, including compound nouns, the verb-based semantic similarity using the tense information, and the common word order similarity. It measures the semantic similarity between concepts that play the same syntactic role. Concerning the word-based semantic similarity, an information content-based measure is used to estimate the semantic similarity degree between words by exploiting the WordNet is-a taxonomy. The proposed method yielded competitive results compared to the previously proposed measures with regard to Li's benchmark [13], showing a high correlation with human ratings.

However, the hybrid methods presented previously are evaluated on the English databases where accuracy can decrease if these methods are applied to other languages. Besides, in these methods, some semantic knowledge, such as semantic class, thematic role and semantic predicates, are not taken into account in measuring sentence similarity.

In the following section, we will detail the proposed sentence similarity method that takes into account semantic and syntactic–semantic knowledge extracted from LMF-standardized dictionaries [3].

2.2 LMF-ISO 24613 standard

LMF [3] was developed by the technical committee TC 37/SC of the ISO. It was conceived as a generic platform for the specification of lexical structures at any level of linguistic description covering monolingual and multilingual lexicons. The specification of LMF follows the Unified Modeling Language (UML)[1] modeling principles defined by the Object Management Group (OMG).[2] It is composed of core meta-model and lexical extensions. The modeling principles allow a lexical database designer to combine any component of the LMF meta-model with data categories to create an appropriate model. These data categories function as UML attribute–value pairs in the diagrams. The core model covers the backbone of a lexical entry. It specifies the basic concepts of vocabulary, word, form and sense. The LMF core model is a hierarchical structure consisting of several components. The lexical entry is one of the components that represents the basic resource in the lexicon.

Figure 1 shows the principle classes of LMF standardized dictionaries and their appropriate attributes. It focalizes on the lexical entry and its related meaning knowledge and associated syntactical knowledge. We can show in this figure that a lexical entry might contain many meanings (or sense). Each meaning is explained by definitions, examples, a subject field and has some relations (i.e., synonymy, antonymy). Each meaning has other specific semantic knowledge such as the semantic class. Each meaning is attached to the possible syntactic behaviours and semantic predicates. Moreover, several researchers elaborated the LMF dictionaries for many languages. Elmadar[3] is an Arabic lexical resource that conforms to the LMF standard ISO-24613. The model of this dictionary [10] covers all lexical levels: morphological, syntactic, semantic and syntactico-semantic. This dictionary contains about 37,000 lexical entries, among which 10,800 are verbs and 3800 roots.

3 The proposed method

In this section, we are proposing a hybrid method to measure sentence similarity. It consists of an extension of previous methods by considering the most relevant semantic knowledge and the syntactico-semantic knowledge, taking advantage of the LMF standardized dictionaries [21]. The proposed method is composed of three phases, as indicated in Fig. 2, such as preprocessing, similarity score attribution and supervised learning.

3.1 Preprocessing

Most of the content-based detection methods include a preprocessing phase in which stop words are removed and words are reduced to their root forms. In our context, we will not remove the stop words because they can be bearer data.

[1] https://www.labri.fr/perso/guibert/DocumentsEnseignement/UML.pdf.

[2] www.omg.org.

[3] http://elmadar.miracl-apps.com/.

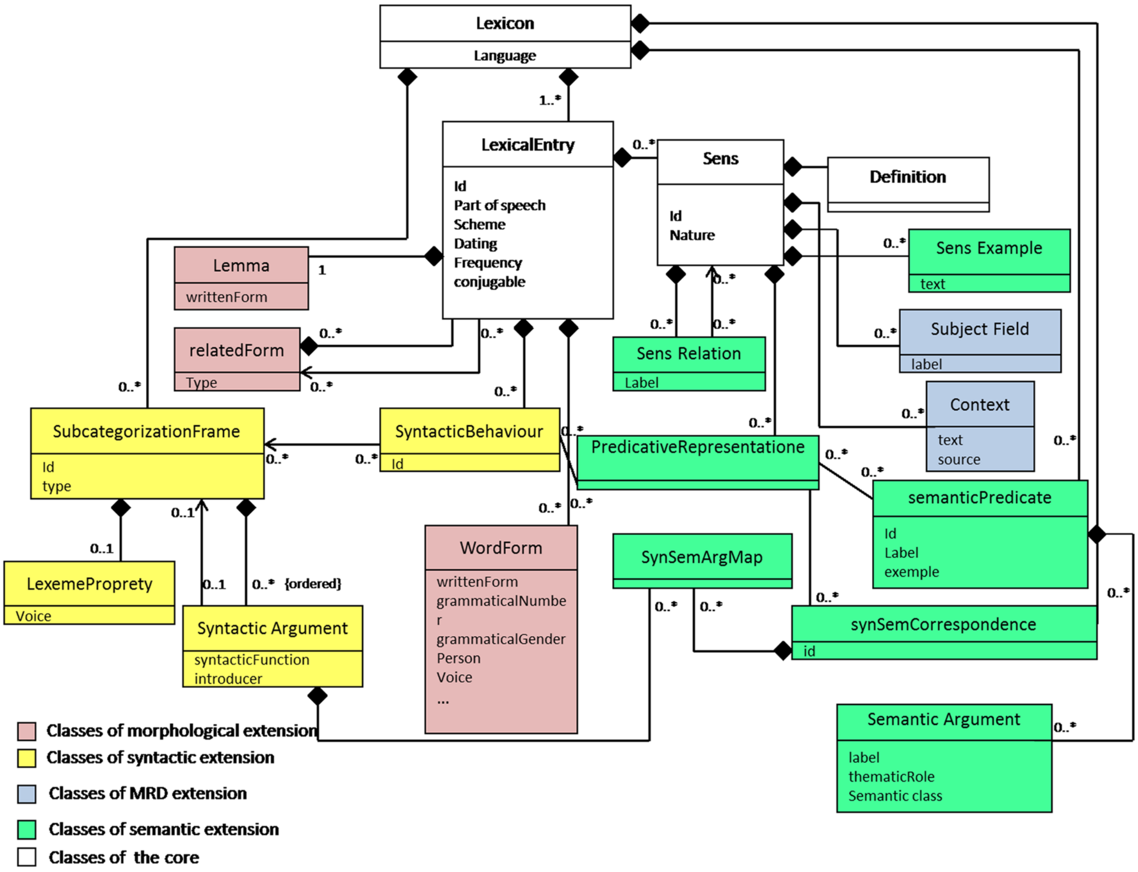
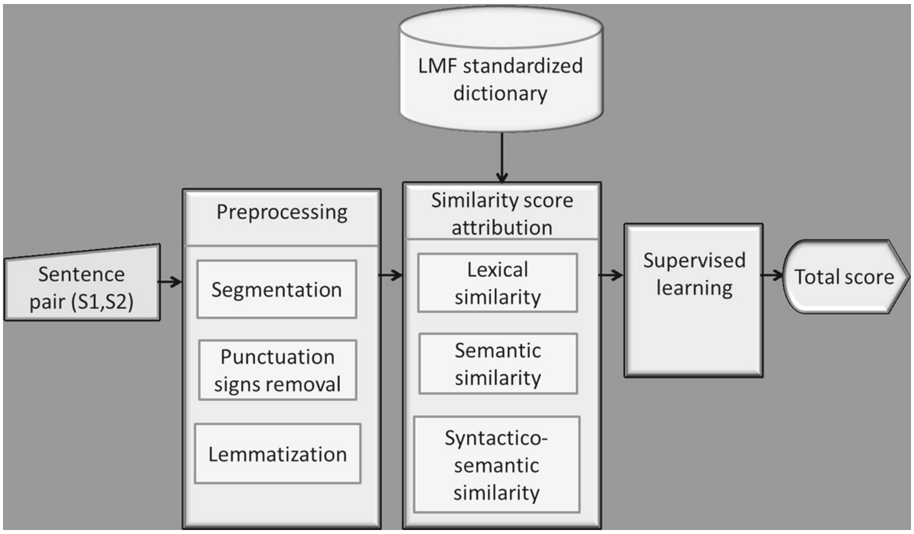


Fig. 1 An extract of an LMF standardized model

Fig. 2 The proposed phases for measuring the similarity between sentences



For example, in the sentences S1: “the boy goes to school” and S2: “the boy does not go to school”, if we remove the word “not” as a stop word, both sentences become similar, whereas they have contradictory meanings. Besides, we will not reduce the word to its root form, but to its stem form. Indeed, the meaning of the word can be different from that of its root, like in Arabic the word “كتيب -booklet” does not have the same meaning as its root “كتب -write”.

The following steps are performed to transform a sentence into a structured and formatted representation, which will be more convenient for the similarity computation process.

- Tokenization: input sentences are broken up into tokens (words).
- Punctuation sign removal: punctuation signs are used in any text. They are considered as unimportant information between sentences. They are removed to get more significant results.
- Lemmatization: morphological variants are reduced to their stem form.

3.2 Similarity scores attribution

We measure three similarity scores as lexical, semantic and syntactico-semantic based on the content of the LMF standardized dictionaries [3].

The score of lexical similarity is computed based on the lexical unit constituting the sentences to extract the lexically similar words.

The lexical similarity score is based on the number of common terms between the sentences. To calculate the score SL(S1,S2), we used Jaccard coefficient [7] that is a fairly quite useful and easy standard to automate the measurement. Thus, the following formula describes how to compute the lexical similarity between sentences.

$$\mathrm{SL}(\mathrm{S1},\ \mathrm{S2}) = \frac{\mathrm{MC}}{\mathrm{MS1} + \mathrm{MS2} - \mathrm{MC}}, \qquad (1)$$

where:

MC is the number of common words between the sentences S1 and S2,
MS1 is the number of words contained in sentence S1 and
MS2 is the number of words contained in sentence S2.

The score of the semantic similarity is computed by the use of LMF standardized dictionaries [3]. The procedure to compute the semantic similarity consists, firstly, in forming a joint word set using only the distinct stems in the pair of sentences. For each sentence, a raw semantic vector is derived and enriched using the LMF standardized dictionary [3]. Indeed, each sentence is readily represented by the use of the joint word set as follows: The vector derived from the joint word set is denoted *T*. Each entry of the semantic vector corresponds to a stem in the joint word set, so the dimension equals the number of stems in the joint word set. The value of an entry of the lexical semantic vector, Ti (i = 1, 2, m), is determined by the semantic similarity of the word corresponding to a word in the sentence. Given that Wi is the word of the joint word set,

Case 1: if Wi appears in the sentence, then Ti is set to 1.
Case 2: if Wi is not contained in the sentence, then a semantic similarity score is computed between Wi and each word in the sentence using the synonymy relations of LMF standardized dictionary (extracted from Sense Relation class). Thus, the most similar word to Wi in sentence is the one with the highest similarity score θ, then Ti is set to θ.

The process of semantic similarity detection is presented in Fig. 3.

In fact, the LMF normalized dictionary model defines many types of semantic relationships (e.g., synonymy, antonymy, etc.) between the meanings of two or several lexical entries by means of the Sense Relation class. Given two words W1 and W2, we need to find the semantic similarity Sim(W1,W2). We can do this by analyzing the synonymy relations between the senses of words as follows: words are linked by a semantic relationship in the LMF standardized dictionary and with relation pointers to other synsets. One direct method for word similarity calculation is to find the synonymy set of each word so as to detect the common synonyms between the two words. For example, the common synonyms between the words “stable” and “constant” are “steady” and “firm”, as the synonyms of “stable” are steady, constant, enduring, firm, stabile, while the synonyms of ”constant” are steady, abiding, firm, perpetual, hourly.

Once the two sets of synonyms for each word are collected, we calculate the degree of similarity between them using the Jaccard coefficient [7]:

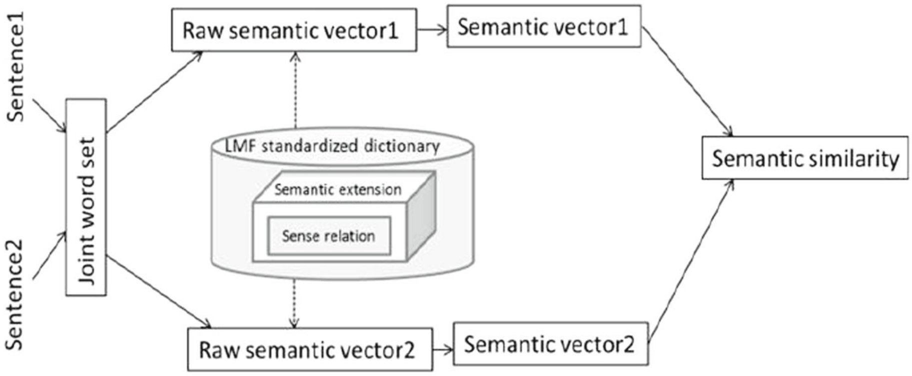


Fig. 3 Semantic similarity computation diagram

$$\mathrm{Sim(W1, W2)} = \frac{\mathrm{MC}}{\mathrm{MW1}} + \mathrm{MW2} - \mathrm{MC}, \quad (2)$$

where:

MC is the number of common words between the two synonym sets,
MW1 is the number of words contained in the w1 synonym set and
MW2 is the number of words contained in the w2 synonym set. From the generated semantic vectors, as described above, we compute the semantic similarity score, which we call SM(S1, S2), between them, using the Cosine similarity [18].

$$\mathrm{SM(S1, S2)} = \frac{V1.V2}{||V1||.||V2||}, \quad (3)$$

where:

V1 is the semantic vector of sentence S1 and
V2 is the semantic vector of sentence S2. Semantic knowledge and especially semantic arguments, which aim at characterizing the meanings of lexical units in sentences, have attracted considerable interest in both linguistic and computational linguistic domains. Such semantic arguments can be defined as a semantic linguistic property that can be used as a valuable means of comprehending the specific meaning of a sentence. Moreover, the semantic argument is characterized by the semantic class and the thematic role that provides information about the relationships between words and provides a mechanism of interaction among the syntactic processors. The thematic role refers to a semantic relationship between a predicate and its arguments. For example, the thematic role, "the broom-handle" is different in a sentence S1: "He banged the broom-handle on the ceiling", and S2: "He banged the ceiling with the broom-handle", because it presents an object in S1 and an instrument in S2. Likewise, the semantic argument "the ceiling" plays the role of a location in S1 and an object in S2.

In our method, these bits of knowledge are extracted from the semantic argument class of LMF normalized dictionary that are associated with a semantic predicate and linked to an appropriate syntactic behaviour.

To calculate the syntactico-semantic similarity, we have first proceeded to extract the proprieties of the semantic arguments of each sentence from LMF dictionary [3]. Therefore, we have used, on the one hand, a syntactic parser to determine the syntactic behaviour of the sentences, and on the other hand, with the help of an expert, the semantic predicates are determined. Then, in the LMF-standardized dictionary, we have looked for the meanings of the lexical entry (verb of the sentence), the predicative representation that combines the syntactic behaviour and the semantic predicate predefined in the first step. Once the predicative representation is found, we extract the semantic arguments. The pairs of the semantic arguments are considered similar if they have the same attributes like the thematic role and the semantic class. The process that describes the determining of semantic class and the thematic role of each sentence argument is presented in Fig. 4.

Afterward, we calculate the degree of syntactico-semantic similarity between the two sentences, S1 and S2, from the common semantic arguments between the pair of sentences, which we call SSM(S1,S2), using the Jaccard coefficient [7]:

$$\mathrm{SSM(S1, S2)} = \frac{\mathrm{ASC}}{\mathrm{ASS1} + \mathrm{ASS2} - \mathrm{ASC}}, \quad (4)$$

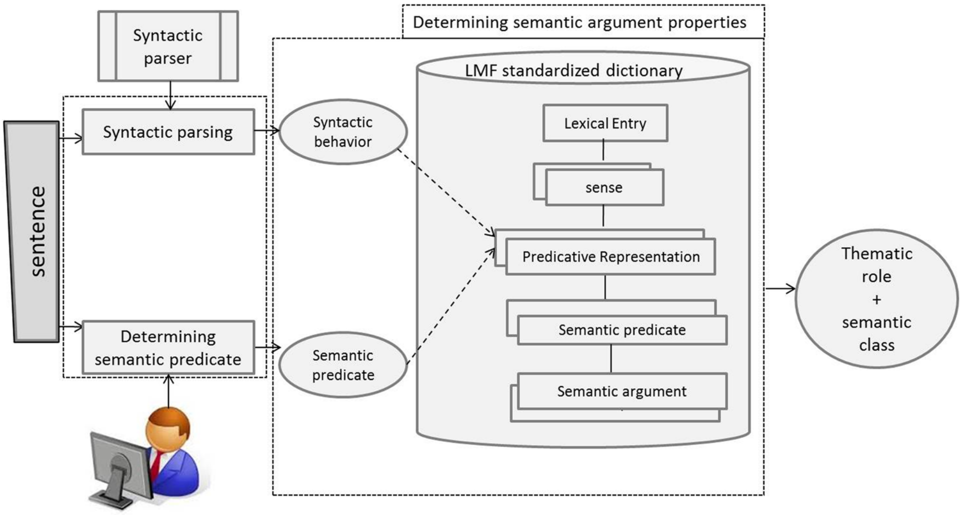

Syntactic parser
Syntactic parsing
sentence
Determining semantic predicate
Determining semantic argument properties
Syntactic behavior
Semantic predicate
LMF standardized dictionary
Lexical Entry
sense
Predicative Representation
Semantic predicate
Semantic argument
Thematic role + semantic class

Fig. 4 Determination diagram semantic argument properties

where:

ASC is the number of common semantic arguments between the two sentences,
ASS1 is the number of semantic arguments contained in sentence S1 and
ASS2 is the number of semantic arguments contained in sentence S2.

3.3 Supervised learning

We propose using supervised learning to define the appropriate coefficients of the similarity scores described below [20]. In this context, the aim is to apply, in the first time a hyperplane equation (decision boundary, such as similar or not similar) on sentences S1 and S2, and deduct a total score similarity "Sim (S1,S2)" in the second time. The total score similarity "Sim (S1,S2)" aggregates the lexical, semantic and syntactico-semantic scores. The process of determining the suitable coefficients includes two phases: the first is the training phase that aims at getting a hyperplane equation via the learning algorithm, such as supervised vector machine (SVM). The second is the test phase that validates the generated equation (hyperplane equation) by the cross-validation method. In the training phase, we first prepare the extraction vectors "V" where each vector describes a pair of sentences "S1" and "S2". Generally, any vector "Vi" is described by the collection of lexical (SL), semantic (SM) and syntactico-semantic similarity (SSM) scores. Each vector "Vi" is completed by a Boolean criterion, namely D. This criterion class is determined by an expert who decides whether the pair of sentences is similar or not. However, the repetition of a number of identical vectors and other contradiction is possible. The extraction vector "Vi" is defined as follows:

Vi (SL, SM, SSM, Di),

where:

Vi is the vector extracting the pair of sentences,
SL is the lexical similarity score between the elements of a pair of sentences,
SM is the semantic similarity score between the elements of a pair of sentences,
SSM is the syntactico-semantic similarity between the elements of a pair of sentences and
Di is the Boolean criterion representing the class of the similar or dissimilar vector Vi .

Then, the SVM learning algorithm is applied to the generated extraction vectors to have an optimal hyperplane that separates two classes (similar and not similar). Indeed, the use of the standard SVM learning algorithms was limited to a group of researchers as these algorithms were long and difficult to implement.

Platt [17] developed a learning algorithm called SMO, "sequential minimal optimization" that can quickly solve the problem of quadratic optimization (QP). This algorithm is usually faster and easier to implement and requires a reduced memory space [9]. The classification equation defined by SMO function is presented as follows:

$$\mathrm{Sim}(\mathrm{S1}, \mathrm{S2}) = \alpha * \mathrm{SL} + \beta * \mathrm{SM} + \gamma * \mathrm{SSM} + C, \qquad (5)$$

where

α is the weight attributed to lexical similarity,
β is the weight attributed to semantic similarity,
γ is the weight attributed to syntactico-semantic similarity and
C is constant.

The test phase consists in validating the classification equation generated in the training phase by the cross-validation method. Indeed, this cross-validation method is a model validation technique of assessing how the results of a statistical analysis generalize to an independent data set.

After the computing process of the sentence similarity score, the similarity class is detected as follows:

If Sim (S1, S2)$\geq$ threshold, then the sentences are similar.
If Sim (S1, S2) $<$ threshold, then the sentences are not similar.

4 Experiments and results

Experiments use on the one hand the LMF standardized Arabic dictionary [10] as a resource to exploit [10] the synonymy of words and properties of semantic arguments (semantic class and thematic role)and, on the other hand, the Stanford Parser [4], the MADAMIRA tool [16] to reduce words to their stem or lemma by removing the suffix, the prefix. After that, they match the remaining word with verbal or noun patterns and the Weka software package [5] to find out the optimal parameters in the learning phase.

4.1 The databases

There are currently no suitable Arabic benchmark data sets (or even standard text sets) for the evaluation of sentence (or a very short text) similarity methods. Building such a data set is not a trivial task due to subjectivity in the interpretation of a language, which is in part due to the lack of deeper contextual information. To evaluate our similarity measure, a preliminary data set of sentence pairs is constructed with human similarity scores provided by five participants. Indeed, each participant is asked to rate the sentences on the scale 0.0–4.0 according to the similarity of the meaning.

These sentences consist of dictionary definitions and examples of words. Then, a further data set of sentences is produced from the Arabic dictionaries for human use such as Lissan Al-Arab, Al-Wassit, Al-Muhit and Tj Al-Arous. Our selection is composed of 690 pairs of sentences as indicated in Table 1.

Table 1 Data sets used in the evaluation of sentence similarity measure

Dataset	#Pairs
Lissan Al-Arab	480
Al-Wassit	266
Al-Muhit	178
Tj Al-Arous	456

4.2 An experiment with human similarities of Arabic sentence pairs

The participants were asked to complete the rating similarity of the sentence pairs on the scale from 0.0 (minimum similarity) to 4.0 (maximum similarity). A rubric containing linguistic anchors was provided for the five major scale points 0.0 (the sentences are unrelated in meaning), 1.0 (the sentences are vaguely similar in meaning), 2.0 (the sentences are very much alike in meaning.), 3.0 (the sentences are strongly related in meaning) and 4.0 (the sentences are identical in meaning). The values are taken from a study by Charles [1], which yielded psychometric properties analogous to an interval scale. The use of the linguistic anchors reconciles these wise conflicting requirements. Each of the 690 sentence pairs was assigned a sentence similarity score calculated as the mean of the judgements made by the experts. Table 2 presents a comparison of our similarity measure with the all human

Table 2 Arabic sentence data set results

Arabic sentence	English translation	Human similarity (mean)	Our proposed method
كُتب الّه النجاة للمريض	God decreed the patient's survival	0.7	0.75
كُتب الّه الشفاء للمريض	God decreed the patient's healing		
خط المتسول الطعام	The beggar took the food	0.6	0.75
خط المتسول في الطعام	The beggar took in the food		
أحس بالوجع	I feel pain	0.5	0.5
أحس باألم في بطني	I have an ache in my belly		
حلاء فلانا درهما	He gave a person the money	0.7	0.75
حلاءه درهما	He gave the money to him		
كتب له الأرض	He wrote him the ground	0.3	0.25
كتب له رسالة	He wrote him a letter		
أغمط المطر	The rain continues	0.4	0.25
أغمطت السماء بالمطر	The sky continues with the rain		
شجرة رفيفة	Shelving tree	1	1
شجر رفيف	Shelving trees		
حقنته المرضة إبرة	The nurse gave an injection	0.45	0.5
حقنت المرضة المريض بلئبرة	The nurse gave the patient an injection		
أبعده الّه	God kept him away	0	0
لا أبعده الّه	God did not keep him away		
أثخن أعداءه	He weakened his enemies	0.3	0.5
أثخن في عدوه	He weakened his enemies with wounds		

similarity scores provided as the score mean for each pair and scaled into the range [0,1].

Furthermore, the weight to lexical similarity “SL(S1,S2)” is 0.2, the weight to semantic similarity “SM(S1,S2)” is 0.35 and the weight to syntactico-semantic similarity “SSM(S1,S2)” is 0.45 in the total similarity score between sentences “Sim (S1,S2)”. Two sentences are considered similar if the total similarity score “Sim (S1,S2)” is superior to 0.85.

4.3 Results and discussion

To evaluate our sentence similarity measure, we used the correlation coefficient to link the scores computed by a measure to the judgements provided by humans in the database. The Pearson correlation coefficient *r* can be used as a metric evaluation. It indicates how well the results of a measure are similar to human judgements, where a value 0 means no correlation and 1 means perfect correlation. Pearson *r* is computed as follows:

$$r = \frac{n\left(\sum xiyi\right) - \left(\sum xi\right)\left(\sum yi\right)}{\sqrt{n\left(\sum xi^2\right)\left(\sum xi\right)^2}\sqrt{n\left(\sum yi^2\right)\left(\sum yi\right)^2}}, \tag{6}$$

where *xi* refers to the *i*th element in the list of human judgements, *yi* refers to the corresponding ith element in the list of sentence similarity computed by our proposed measure and *n* is the number of sentence pairs.

Our sentence similarity measure achieved a reasonably good Pearson correlation coefficient of 0.92, with the human ratings significant at the 0.01 level. In Table 3, we present the results that have calculated the correlation coefficient *r* for the judgements for each participant against the rest of the group and then kept the means.

The evaluation of our proposal is achieved following the cross-validation method and using the Weka tool. To realize this, we divided the training corpus into two distinct parts, one for learning (80 %) and one for testing (20 %). The results are given in Table 4.

The obtained results are encouraging and represent a good start to implement automatic learning in measuring sentence similarity in Arabic. We noticed that the analysis of short sentences (≤10 words) presents the highest measures of recall and precision. As the sentence gets longer, there will be a more complex computation, which reduces the system performance. We believe that these results can be improved. In fact, we think that we can improve the learning stage by adding other features besides the semantic argument and synonymy senses. As an example of additional features, we can incorporate other types of relations, such as hyponymy. We will explore the effects of the integration of phrase functions in the learning phase. During the implementation of our system, we noticed that the bigger the number of sentences, the higher are the recall and precision. Therefore, we believe that the enrichment of our database of Arabic sentences can significantly enhance the results. In addition, the performance of our system depends on the lemmatizer system, syntactical parser, synonyms and semantic predicates retrieved from the Arabic LMF dictionary [10]. According to a comparative evaluation study of Arabic language stemmers and syntactical parsers, MADAMIRA [16] and the Stanford parser [4] achieved the highest accuracy. Consequently, we do not expect to increase the performance of our proposed measure using other lemmatizers or syntactical parsers.

Table 3 Similarity correlations

	Correlation *r*
Our proposed measure	0.92
Mean of all participants	0.938
Worst participant	0.73
Best participant	0.947

Table 4 Evaluation results

Precision (%)	Recall (%)	*F*-score (%)
88.12	83.24	85.61

5 Conclusion

Sentence similarity measures are an old and valuable area for various applications. However, they have not considered some relevant semantic knowledge, such as, the thematic role, the semantic class and syntactico-semantic knowledge like the semantic predicate in computing the sentence similarity. In this paper, we proposed a method to extend the previous methods by enhancing the similarity measure between sentences with the semantic and syntactico-semantic knowledge profiting from the LMF standardized dictionaries. In fact, a standardized LMF dictionary is a finely structured source, rich in lexical, semantic and syntactic knowledge. Our method consists of three stages. It starts with preprocessing the sentence pairs; then it proceeds by attribution the following similarity scores lexical, semantic and syntactico-semantic, and finally ends with computation total score using supervised learning. Besides, the proposed method is proven to be reliable despite the illustrations carried out on the Arabic language, the choice of which is explained by three main motives. The first one is the great deficiency of works on the Arabic language measuring sentence similarity; the second is the processing within our research team of the Arabic language; and the third an LMF standardized Arabic dictionary is at hand where the syntactico-semantic component is available and well structured. Additionally, we evaluate our proposal on 690 pairs of sentences taken from various definitions and examples of Arabic dictionaries. In fact, we reached a good correlation $r = 0.92$ for the formed Arabic data set close to human judgements. As perspectives of our work, we aim to extend our sentence similarity measure in enriching the Arabic dataset by including other kinds of semantic relations, such as hyponymy. Finally, we propose to apply our method to other languages.

References

1. Charles, W.G.: Contextual correlates of meaning. Appl. Psycholinguist. **21**, 505–524 (2000)
2. Chatterjee, N. A.: Statistical approach for similarity measurement between sentences for EBMT. In: Proceedings STRANS-2001, IT Kanpur, pp. 122–1318 (2001)
3. Francopoulo, G.: LMF Lexical Markup Framework. ISBN: 978-1-84821-430-9. Wiley, New York, 288 pages (2013)
4. Green, S., Manning, C.D: Better arabic parsing: baselines, evaluations, and analysis. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, pp. 394–402 (2010)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
6. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans Knowl Discov Data TKDD **2**(2), 10 (2008)
7. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Impr, Corbaz (1901)
8. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. cmp-lg (1997)
9. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput. **13**(3), 637–649 (2001)
10. Khemakhem, A., Gargouri, B., Hamadou, A.B., Francopoulo, G.: ISO standard modeling of a large arabic dictionary. Natural Language Engineering, pp. 1–31 (2016). doi: 10.1017/S1351324915000224
11. Leaock, C., Chodorow, M., Miller G.: Combining local context andwordnet similarity forword sense identification. In: Fellbaum, C. (ed.) WordNet. An Electronic Lexical Database, pp. 265–283. MIT Press, Cambridge (1998)
12. Lee, C.M., Chang, J.W., Hsieh, T.C., Chen, H.H., Chen, C.H.: Similarity measure based on semantic patterns. Adv. Inf. Sci. Serv. Sci. AISS **4**(18), 10 (2012)
13. Li, Y., Mclean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE Trans Knowl Data Eng **18**(8), 1138–1150 (2006)
14. Mandreoli, F., Martoglia, R., Tiberio, P.: A syntactic approach for searching similarities within sentences. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, CIKM '02, ACM, New York, pp. 635–637. (2002)
15. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: an on-line lexical database. Int. J. Lexicogr. **3**(4), 235–244 (1990)
16. Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.: Madamira: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. LREC **14**, 1094–1101 (2014)
17. Platt, J.C.: 12 fast training of support vector machines using sequential minimal optimization. Adv. Kernel Methods, pp. 185–208 (1999)
18. Salton, G.: Automatic information organization and retrieval. McGraw Hill, New York (1968)
19. Taieb, M.A.H., Aouicha, M.B., Bourouis, Y.: Fm3s: features-based measure of sentences semantic similarity. In: Hybrid Artificial Intelligent Systems, pp. 515–529. Springer, Heidelberg (2015)
20. Wali, W., Gargouri, B., et al. Supervised learning to measure the semantic similarity between Arabic sentences. In: Computational Collective Intelligence, pp. 158–167. Springer, Heidelberg (2015)
21. Wali, W., Gargouri, B., Hamadou, A.B.: Using standardized lexical semantic knowledge to measure similarity. In: Knowledge Science, Engineering and Management, pp. 93–104. Springer, Heidelberg (2014)
22. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 133–138. Association for Computational Linguistics (1994)