Improving question analysis for Arabic question answering in the medical domain

Sondes Dardour[1], Héla Fehri[1], and Kais Haddar[1]

[1] MIRACL Laboratory, University of Sfax, Tunisia

dardour.sondes@yahoo.com
hela.fehri@yahoo.fr
kais.haddar@yahoo.fr

Abstract. Question analysis is a basic module in a question answering (QA) system, and its quality affects the performance of QA system. In this paper, we address the problem of Arabic question analysis in the medical domain where several specific challenges are met. The major challenging issue in processing Arabic medical question is the need for ambiguity resolution. Nevertheless, this issue has not been well studied in related works. Our question analysis uses dictionaries and transducers to analyze any medical question, factoid or complex. This module detects important elements of the question, including: different words in the question that identify what the user wants to ask for, and the nature of the expected answer. To identify well these elements a step of disambiguation is applied. Then, the words used in the question will be extended by adding new words that connect semantically to those in the question. Experimentations of the question analysis module of our Arabic medical question answering system show interesting results.

Keywords: Question answering, Arabic, Disambiguation, Medical domain, dictionary, transducer

1 Introduction

Nowadays, due to the continuous exponential growth of information produced in the medical domain, and due to the important impact of such information upon research and upon real world applications, there is a particularly great and growing demand for Question Answering (QA) systems that can effectively and efficiently aid users in their medical information search [1].

QA system takes a question posted in natural language instead of a set of keywords, analyzes and understands the meaning of the question, and then provides the exact answer from a set of knowledge resources [2]. The QA system consists of three main processing modules, namely, question processing, passages retrieval processing, and answer processing. A question processing is the primary and basic source through which a search process is directed for answer. Therefore, an accurate and careful analysis to the question is required. Thus, question processing is the most fundamen-

tal module in any QA system, and the performance of its results significantly impacts on the following modules of information retrieval and answer extraction.

To our knowledge, proposed Arabic medical QA systems are so limited either in terms of their performance as well as in terms of the types of questions they are designed to answer. Moreover, the most attention in Arabic has been paid to answering factoid questions, in which the answer is a single word or a short phrase [3].

Ambiguity is a common phenomenon in human natural language. In QA, ambiguity is a critical challenge in extracting what the user looking for in his question. Therefore, ambiguity can cause confusion in interpretation of the question, and then impacts negatively the performance of the QA system.

In this paper, we propose a new approach to handle medical questions (factoid and complex questions) for the Arabic language. Moreover, our approach overcomes the ambiguity in the question processing module, an issue that has not been appropriately addressed in the field of Arabic QA.

The remainder of this paper is structured as follows. Section 2 presents the related works. Ambiguity problems are presented in section 3. Section 4 describes our approach. Section 5 deals with the experimentation carried out to evaluate the efficiency of our question analysis module. Finally, section 6 draws the main contributions and proposes further perspectives.

2 Related works

The problem of answering questions formulated in natural language has been studied in the field of Information Retrieval (IR) since the mid-1990s [4]. However, unlike IR, the QA system returns simple and precise answer to a natural language question instead of a large number of documents [5] [6]. As we mentioned, the QA system is composed of three modules: question analysis, passage or document retrieval and answer extraction. Different QA systems may use different implementation for each module [7] [8]. In this section, we focus on some studies for the question analysis module.

Until now, very little effort was directed toward the development of QA system for the medical domain in the Arabic language, compared to other languages such as French and English. This is mainly attributed to the particularities of the medical domain and the language (see Section 3). The situation is further aggravated by the lack of linguistic resources and Natural Language Processing (NLP) tools that is available for Arabic [9] [10].

In an effort to achieve a better question analysis, [2] analyzed the question to extract type and category of desired answer whether it is a place, a quantity, a name or a date, which makes the answer extraction easier.

[10] analyzed Arabic questions by formulating the query, extracting the expected answer type, the question focus and the question keywords. The focus is the noun phrase of the question which the user wants to ask about. For instance, if the user's question is "What is the capital of Canada?" then the question focus is "Canada" and

the keyword is “capital” and the expected answer type is a named entity for a location.

[11] analyzed the question by:

- Tokenization and normalization,
- Determining answer type by question words (When, What...),
- Named entity recognition,
- Focus determination by extracting the main named entity,
- Keywords extraction,
- Removing stop words using the Khoja stop list,
- Query expansion using the Arabic dictionary of synonyms. Named entities are not expanded to avoid ambiguity,
- Stemming by Khoja’s Stemmer and named entities are not stemmed,
- Query generation of keywords into a boolean formula.

[3] made six steps to process Why-questions. They tokenized the question, then normalized it, then removed stop words (optional step). After that, they applied khoja’s stemmer to obtain the root of each non-stop word in the question. Then, they used the extracted keywords to formulate and generate the query. Finally, they extended the list of keywords by including synonyms and words that share the same root.

The system of [12] and [13] are developed for the medical domain in Arabic language. These systems analyze only factoid questions by extracting the topic and the focus of the question, and extracting named entities. The system of [12] classifies the questions into organization, location, person, viruses, diseases, treatment.

We can confirm, from literature reviews, that most Arabic QA systems ensure analysis of factoid questions. Nevertheless, there are few studies that have addressed the problem of answering complex questions. In addition, there are few works that have integrated semantic analysis and treated the medical field in the Arabic language, which makes the development of a new Arabic QA system is crucial.

3 Ambiguity

A study of different questions showed us the existence of several linguistic phenomena which can cause ambiguities in the question processing. Indeed, if we solve these problems, the errors will be so minimal and our system will be more relevant compared to existing Arabic QA systems.

3.1 Specific Arabic difficulties

Arabic specific difficulties consist in its richness that needs special processing, which makes regular NLP systems, designed for other languages, unable to process it. One of the Arabic-specific difficulties is the lack of diacritics (i.e. kasra, fatha, damma), which leads to more ambiguous situations than any other language. This issue can be explained through the question “من الذي قتل في أوغندا؟” (Who was killed in Uganda?).

The lack of diacritics in verb "قتل" (to kill) presents at least two cases for the question processing:

- "قُتِلَ" *qutila*[1] which means that the question is "Who was killed in Uganda?", so "قتل" in this question means "was killed".
- "قَتَلَ" *qatala* which means that the question is "Who did kill in Uganda?", so "قتل" in this question means "kill".

Arabic language morphology is challenging when compared to other languages. This is because Arabic is a highly agglutinative and derivational language where a word token can replace a whole sentence in other languages. For example, for the question "أيمكننا منع الجلطة؟" (Do we can prevent the clot?), the sentence "Do we can" can be expressed in one Arabic word "أيمكننا" which includes the verb "يمكن" (can), the prefix "أ" (do) and the pronoun "نا" (we). Therefore, extracting keywords from an Arabic question will be more complex than any other language. Furthermore, in a question like "من الطبيبان اللذان منحا جائزة نوبل في الطب لعلاج السرطان؟" (Who are the two scientists who won the Nobel Prize in medicine for cancer treatment?), the user looks for the name of two persons (i.e. James P. Allison and Tasuku Honjo). In English, the system catches this user require through the word "two". In Arabic QA, this keyword is embedded in the word "الطبيبان" *Alt~abiybaAni* (two scientists) thanks to the suffix "ان" *Ani*. Actually, the question above is just an example; the morphology of an Arabic word may contain multiple information (basic POS, number, gender, etc.) which are important for each module of Arabic QA.

Unlike English and most Latin-based languages, Arabic does not have capital letters which makes Named Entity Recognition (NER) harder [14].

3.2 Specific difficulties of medical domain

Apart from ambiguity in Arabic language, ambiguity also appears in medical terms. We observed that the more ambiguous terms are diseases names. For example, the term "القمل" means both an insect and a dermatological disease. This issue can be explained through the question « ماهو القمل؟ » (What is a louse?); such system can extract the following answers:

القمل هو نوع من الحشرات الضارة التي تتغذى على دم الإنسان (1)

(Louse is a kind of harmful insect that feeds on human blood)

Definition of an insect

القمل هو مرض يصيب فروة الرأس (2)

(Louse is a disease that affects the scalp)

Definition of a disease

[1] Buckwalter Arabic transliteration

In fact, to extract the right answer, the system must understand the context. For example, in (2), the keyword "مرض" (disease) indicate that it is a definition of the disease "القمل" (Louse).

Furthermore, in open-domain, the nature of the expected answer is known from the interrogative pronouns. For instance, in a When-question « متى اكتشفت امريكا؟ » (When America discovered?), the nature of the expected answer is a time. Nevertheless, in medical-domain, a When-question can indicate an age, a condition or a time. Table 1 gives an example.

Table 1. Ambiguity of the question "متى" (When)

Question	Translation	Question type	Expected answer
متى يتكون قلب الجنين؟	When is the fetus heart developed?	Wh-question	Age
متى يجب زيارة طبيب نفسي؟	When should visit a psychiatrist?	Wh-question	Condition
متى اكتشف مرض الزهايمر؟	When did Alzheimer's disease be discovered?	Wh-question	Time

To extract the correct answer, we must define a sequence of keywords which define the question and disambiguate it in the sense that it indicates what the question is looking for.

4 Proposed Method

The challenges discussed in the previous section make clear the need for new method to deal with Arabic medical QA. In addition, the most of previous studies are based on a superficial analysis of factoid questions (i.e. where, when, how much/many, who and what). The originality of our approach lies in the disambiguation and the semantic analysis of factoid and complex questions (i.e. why and how to). In our proposal, the question analysis module is based on five steps as illustrated in Fig 1: Corpus study, Named Entity Recognition (NER), Stop word removal, Disambiguation, and Question Expansion (QE). In the first step, questions are gathered and studied to define the disambiguation patterns. These patterns are transformed into transducers to process any type of medical question in Arabic language. Questions will be processed by the **parallel** steps (NER, Stop word removal, and disambiguation) using dictionaries, syntactic grammars, and morphological grammar in order to get some useful information. Finally, the last step will extend the extracted keywords.

4.1 Corpus study

The need to have an Arabic corpus is a necessity for processing Arabic QA systems. Indeed, the questions are gathered from several sources, namely, discussion forums, frequently asked questions (FAQ) and some questions translated from Text REtrieval Conference (TREC). Currently, we collected 350 questions which contain seven categories (see Table 2). The questions are then subjected to an analysis step.

Table 2. Collected questions

Question type	ما What	متى When	أين Where	من Who	كم How many/much	كيف How	لماذا Why
Number	85	53	42	38	45	39	48

According to our study, we identify 158 question disambiguation patterns. Table 3 shows some patterns of the question "متى" (When). These patterns will be transformed into transducers to parse the questions.

Table 3. Some question disambiguation patterns of the question "متى" (When)

Question	Expected answer
متى<Verb>جنين| طفل |رضيع؟ When<Verb>Fetus| Child |Infant?	Age
متى<Noun><Verb>طفل |جنين|رضيع؟ When <Verb><Noun><Child| Fetus| Infant?	Age
متى<Verb>< Noun ><Condition>؟	Condition
متى<Verb>< Trigger ><Virus>؟	Time
متى<Verb><Virus>؟	Time

4.2 Named Entity Recognition (NER)

The previous studies emphasize that the NER is important for all the QA system components. Indeed, the integration of a NER step will definitely boost our system performance because the answer of a factoid question is a named entity.

In our case, we developed our own NER tool especially formulated for our proposal. This step is based on dictionaries and transducers. We have considered five categories:

- **Organ**: Names of medical organs
- **Location**: names of location;
- **Disease**: Names of diseases, sickness, illness;
- **Virus**: Names of medical viruses;
- **Treatment**: Names of Treatments.

4.3 Stop Words Removal

This step removes the conjunctions, prepositions and interrogative pronouns. After removing the stop words, the important terms in the question will be remaining. In our proposal, the stop words are eliminated from the outputs of the syntactic transducers (see Fig 3).

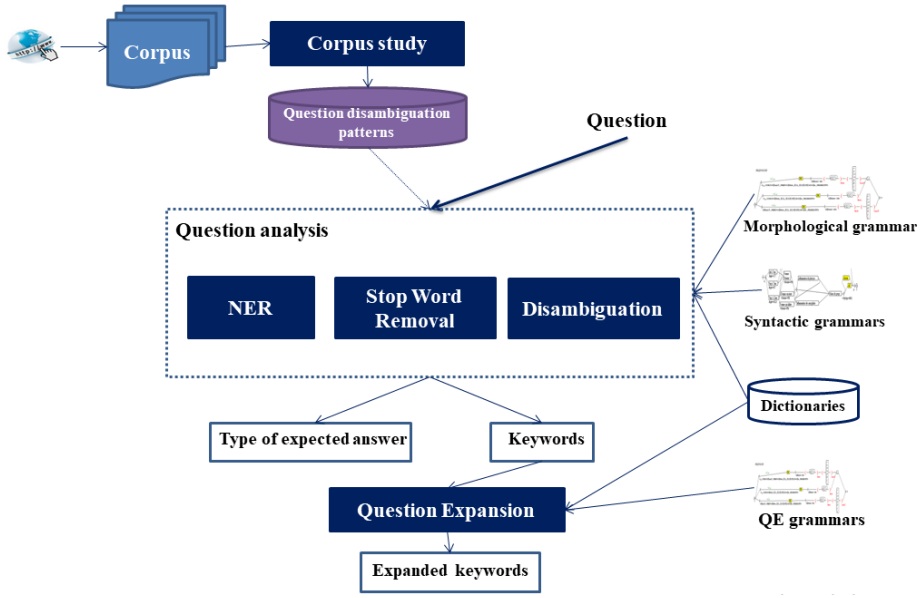

Fig. 1. Proposed method for question analysis module

4.4 Disambiguation

Our system is based on dictionaries and transducers. These resources allow us to disambiguate ambiguous words and the nature of the expected answer (Problems mentioned in the previous section).

Word Sense Disambiguation (WSD). WSD process is required in application such as a QA application [15]. Some ambiguous words which have a different sense influence negatively the extraction of the correct answer. Let's take the following questions as an example:

متى يولد الدماغ الطاقة؟ (3)

(When does the brain generate electricity?)

متى يولد الجنين الذي يعاني من تشوهات خلقية؟ (4)

(When a baby who has congenital anomalies is born?)

As shown above, the verb "يولد" have the sense of "generate" in the question (3) and the sense of "born" in the question (4). To resolve this problem, as shown in our dictionary in Fig 2, each ambiguous word is associated with semantic feature to identify the sense of the entry (sens_generer, sens_naitre). This feature is used in the syntactic transducers (see Fig 3).

```
ضَرَبَ,V+Tr+Correct+Salem+FLX=V_daraba1+DRV=N_daraba1:FlxDRV
نَـزَلَ,V+Tr+Correct+Salem+FLX=V_daraba1+DRV=N_daraba1:FlxDRV
وَلَـدَ,V+Tr+Sens_generer+FLX=V_allama4+DRV=N_allama4:Flx1
أَوْلَـدَ,V+Tr+Sens_naitre+FLX=V_akrama5+DRV=N_akrama5:FlxDRV
```

Fig. 2. Extract of dictionary

The WSD process allows also our system to define the correct stem. For instance, the stem of "يولد" *ywld* in question (3) is "وَلَّدَ" *wal~ada* and in the question (4) is "أَوْلَدَ" *>awolada*.

Disambiguation of the nature of the expected answer. For a reliable disambiguation, each defined pattern in the corpus study step is transformed into transducers. The identification of the nature of the expected answer is related to the focus of the question. For example, the transducer of Fig 3 describes the paths of the pattern "متى<Verb>جنين|طفل|رضيع؟" (see Table 3). This transducer can analyze a question like "متى يحبو الطفل؟" (When the child crawls?). The focus in this example is "طفل" (child), so the nature of the expected answer is "Age".

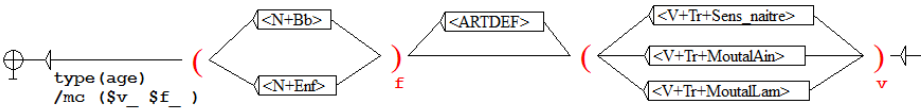

Fig. 3. Transducer of pattern "متى<Verb>جنين|طفل|رضيع؟"

4.5 Question Expansion

Extraction only original question keyword is proved to have some limitations. To get rid of these limitations, we need to define the meaning the user looking for. Therefore, in question expansion (QE), we extend the list of the exact words of the user's question by adding new words that connect semantically to those in the question. Since the documents may not contain the terms that the user used in his question, expanding question will increase the chance of getting the answer [16].

In the previous works, QE is achieved using Arabic WordNet[2]. In our dictionaries, the feature Syno (for synonyms) is used to expand questions. This feature is called in the QE transducer as shown in Fig 4.

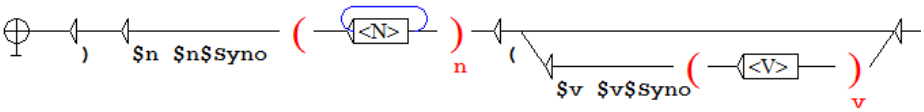

Fig. 4. QE transducer to extract synonyms

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

After processing the question "ما هو الاستخدام المناسب لواقي الشمس للرضيع؟" (What is the appropriate use of sunscreen for a baby?) with the previous steps, the transducer of Fig 4 can extract the synonym of "استخدام" <*isotixdaAm* which is "استعمال" <*isotiEomaAl*, the synonym of "مناسب" *munaAsib* which is "ملائم" *mulaA}im* and the synonym of "رضيع" *raDiyEo* which is "طفل" *Tifolo*.

Expanding question can be applied also in order to overcome the situations where the Passage Retrieval (PR) module eliminates relevant passages containing other forms of the question keywords. The idea now is adding other forms of the keywords that share the same root (see Fig 5).

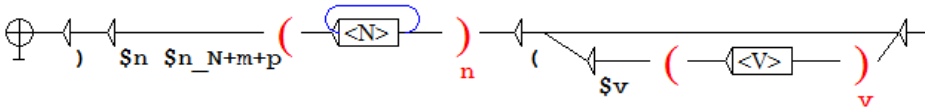

Fig. 5. QE transducer to extract different forms

Let's continue with the same question "ما هو الاستخدام المناسب لواقي الشمس للرضيع؟". Thanks to QE process, the PR module can extract not only the passages that contain the keyword "رضيع" *raDiyEo* but also its broken plural form "رضع" *ruD~aEo*. This process is applied also to extracted synonyms. Therefore, we consider each keyword with its synonym and its different forms since the QE would theoretically generate all these terms.

The expanded list of terms extracted from the question will be sent to the PR module to extract the passages that may contain the answer.

5 Experimentation and evaluation

In our proposal, linguistic resources are built with the linguistic platform NooJ [17]. We conduct a set of experimentations to evaluate the performance of our question analysis module. Therefore, we exploit a test corpus which contains 399 questions. For each question type (ما "What", متى "When", أين "Where", من "Who", كم "How many/much", كيف "How",لماذا "Why") a set of 57 questions is used. The results of applying the transducer that extracts the type of the expected answer and keywords are illustrated in Fig 6. This transducer allows the NER, stop words removal, and disambiguation. Then, the keywords are expended by the QE transducers.

After applying the analysis on the test corpus using our linguistic resources, we obtain the results illustrated in Table 4.

Table 4. Summarizing the measure values

Method	Without disambiguation	With disambiguation
Precision	0.66	0.93
Recall	0.58	0.87
F-Measure	0.61	0.89

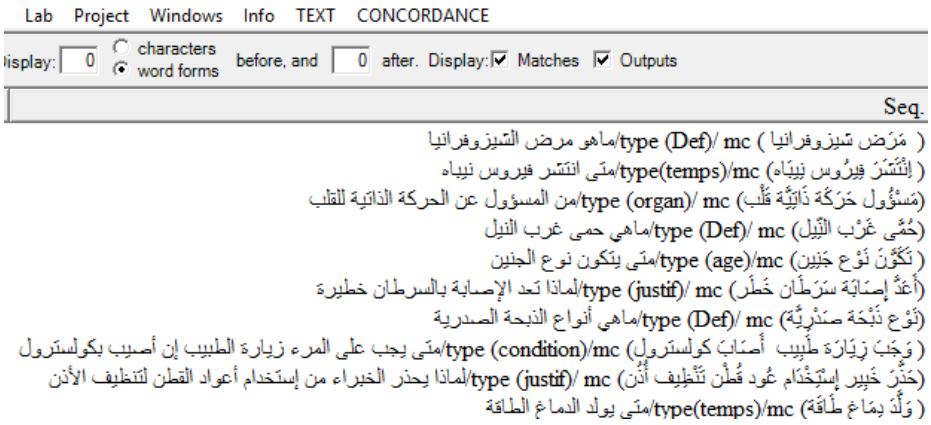

Fig. 6. Extract of concordance table

Table 4 shows that the disambiguation process enhances the F-Measure by 28%. It is then concluded that by reducing ambiguity, especially when processing the medical domain in the Arabic language, the obtained results will be increased.

Errors are often due to the problem in writing some Arabic letters such as the letter "ا" "*A"* which can also be writing like "أ" ">" or "آ" "|" or "إ" "<". For example, in some question, we can find the word "inflammation" written like "التهاب" *AlotihaAbo* or "إلتهاب" <*ilotihaAbo*. To resolve this problem, we need to rewrite the question by unifying all variants of a letter into a single form. Furthermore, the presented errors in the question analysis are due to dictionaries' coverage that must be improved and the complexity of some questions that requires special handling techniques.

6 Conclusion

In the present paper, we have developed a question analysis module (QAM) for our system to analyze an Arabic medical question. Our QAM is mainly concerned with the identification of four factors, namely, keywords extraction, disambiguation, question expansion, and nature of the expected answer extraction. This analysis of question allows extracting all the necessary information that will be used as inputs for the other QA components. Our proposed method achieves satisfactory results.

In the future work, we seek to add a pre-processing to normalize the question. We also seek to improve our linguistic resources by adding new terms in the dictionaries.

7 References

1. Athenikos, S. J., Han, H.: Biomedical question answering: A survey. In: Computer methods and programs in biomedicine, vol. 99, n. 1, pp. 1-24 (2010).

2. Hammo, B., Abuleil, S., Lytinen, S., Evens, M.: Experimenting with a question answering system for the Arabic language. In: Comput. Human. Vol. 38, n. 4, pp. 397–415 (2004).
3. Azmi, A. M., Alshenaifi, N. A.: Lemaza: An Arabic why-question answering system. In: Natural Language Engineering, vol. 23, n. 6, pp. 877-903 (2017).
4. Verberne, S.: In Search of the Why. PhD Thesis, University of Nijmegen, The Netherlands (2010).
5. Kanaan, G., Hammouri, A., Al-Shalabi, R., Swalha, M.: A new question answering system for the Arabic language. In: American Journal of Applied Sciences, vol. 6, n. 4, pp. 797 (2009).
6. Trigui, O., Belguith, L. H., Rosso, P.: DefArabicQA: Arabic definition question answering system. In: Workshop on Language Resources and Human Language Technologies for Semitic Languages, 7th LREC, Valletta, Malta, pp. 40-45 (2010).
7. Benajiba, Y., Rosso, P., Soriano, J. M. G.: Adapting the JIRS passage retrieval system to the Arabic language. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 530-541. Springer, Berlin, Heidelberg (2007).
8. Ezzeldin, A. M., Shaheen, M.: A survey of Arabic question answering: challenges, tasks, approaches, tools, and future trends. In: Proceedings of the 13th International Arab Conference on Information Technology (ACIT 2012), pp. 1-8 (2012).
9. Abouenour, L., Bouzoubaa, K., Rosso, P.: Improving Q/A using Arabic wordnet. In: Proc. The 2008 International Arab Conference on Information Technology (ACIT'2008), Tunisia, December (2008).
10. Brini, W., Ellouze, M., Mesfar, S., Belguith, L.H.: An Arabic question-answering system for factoid questions. In: IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009, pp. 1–7 (2009)
11. Abdelbaki, H.; Shaheen, M.; Badawy, O.: ARQA high performance arabic question answering system. In: Proceedings of Arabic Language Technology International Conference (ALTIC) (2011).
12. Bessaies, E., Mesfar, S., Ghzela, H. B.: Processing Medical Binary Questions in Standard Arabic Using NooJ. In: International Conference on Applications of Natural Language to Information Systems, pp. 193-204. Springer, Cham (2018).
13. Ennasri, I., Dardour, S., Fehri, H., & Haddar, K.: Question-Response System Using the NooJ Linguistic Platform. In: International Conference on Automatic Processing of Natural Language Electronic Texts with NooJ, pp. 190-199, Springer, Cham (2017).
14. Shaheen, M., Ezzeldin, A. M.: Arabic question answering: systems, resources, tools, and future trends. In: Arabian Journal for Science and Engineering, 39(6), pp. 4541-4564 (2014).
15. Ferrandez, S., Roger, S., Ferrández, A., Aguilar, A., López-Moreno, P.: A new proposal of Word Sense Disambiguation for nouns on a Question Answering System. Advances in Natural Language Processing. In: Research in Computing Science, vol. 18, pp. 83-92 (2006).
16. Al-Chalabi, H., Ray, S., Shaalan, K..: Semantic Based Query Expansion for Arabic Question Answering Systems. In: Arabic Computational Linguistics (ACLing), 2015 First International Conference on IEEE, pp. 127-132 (2015).
17. Silberztein, M.: Using Linguistic Resources to Evaluate the Quality of Annotated Corpora. In: Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pp. 2-11 (2018).