ISSN 1870-4069

Classification and Annotation of Arabic Factoid Questions: Application to Medical Domain

Essia Bessaies, Slim Mesfar, Henda ben Ghezala

University of Manouba,
Tunisia

essiabessaies@gmail.com, mesfarslim@yahoo.fr,
henda.benghezala@ensi.rnu.tn

Abstract. Our question answering system is based on a linguistic approach, using NooJ's linguistic engine in order to formalize the automatic recognition rules and then apply them to a dynamic corpus composed of medical journalistic articles. We started by putting in place rules that identify and annotate the different medical entities. The module called entity recognition is able to find references to people, places and organizations, diseases, viruses, as targets to extract the correct answer from the user. These annotations are used in our system in order to identify these answers associated with the extracted named entities. The system is mainly based on a set of local grammars developed for the identification of different structures of phrases to extract the right answer. In addition, we present a method for analyzing medical questions and the approach to finding an answer to a submitted question based on the linguistic approach. The precision and recall show that the actual results are encouraging and could be integrated in a more general Arabic question answering system.

Keywords. Information extraction, medical questions, Arabic language, local grammar, named entities, journalistic articles.

1 Introduction

Nowadays, the medical domain has a high volume of electronic documents. The exploitation of this large quantity of data makes the search of specific information complex and time consuming. This complexity is especially evident when we seek a short and precise answer to a human natural language question rather than a full list of documents and web pages. In this case, the user requirement could be a Question Answering (QA) system which represents a specialized area in the field of information retrieval.

The goal of a QA system is to provide inexperienced users with flexible access to information allowing them to write a query in natural language and obtain not the documents which contain the answer, but its precise answer passage from input texts. There has been a lot of research in English as well as some European language QA systems. However, Arabic QA systems (Brini et al, .2009) could not match the pace

Table 1. Question answering system.

	Arabic language	Linguistic	learning-based	Factoid Question	No-Factoid	Semantic analysis of the question	Without Semantic
QARAB (Hammou et al. 2002)	✓		✓	✓	✓		✓
ArabiQA (Ben NAjiba et al. 2007)	✓	✓		✓		✓	
QALC (Ferret et al., 2000; Ferret et al., 2001a)		✓		✓		✓	
QUANTUM (Plamondon et al. 2002)			✓	✓	✓	✓	
AQAS (Mohammed F, et al 1993)	✓		✓	✓	✓	✓	
Qristal (Laurent et al., 2005)			✓	✓	✓	✓	
Esculape (Embarek,2009)		✓	✓	✓	✓	✓	
Piquant (Chu-Carroll et al., 2002)			✓	✓		✓	
JAVELIN (Nyberg et al., 2002)			✓	✓	✓	✓	
InsightSoft (Soubbotin et al., 2002)			✓	✓		✓	
PowerAnswer (Moldovan et al., 2002)		✓	✓	✓	✓	✓	
Webcoop(Ben amara, 2004)			✓		✓	✓	
Citron (Falco et al, 2014)			✓	✓	✓	✓	

due to some inherent difficulties with the language itself as well as due to lack of tools available to assist researchers. Therefore, the current project attempts to design and develop the modules of an Arabic QA system. For this purpose, the developed question answering system is based on a linguistic approach, using NooJ's linguistic engine in order to formalize the automatic recognition rules and then apply them to a dynamic corpus composed of medical journalistic articles. In addition, we present a method for analyzing medical questions (for a factoid questions). The analysis of the question asked by the user by means of the syntactic and morphological analysis.

The linguistic patterns (grammars) which allow us to extract the analysis of the question and the semantic features of the question of extracting the focus and topic of

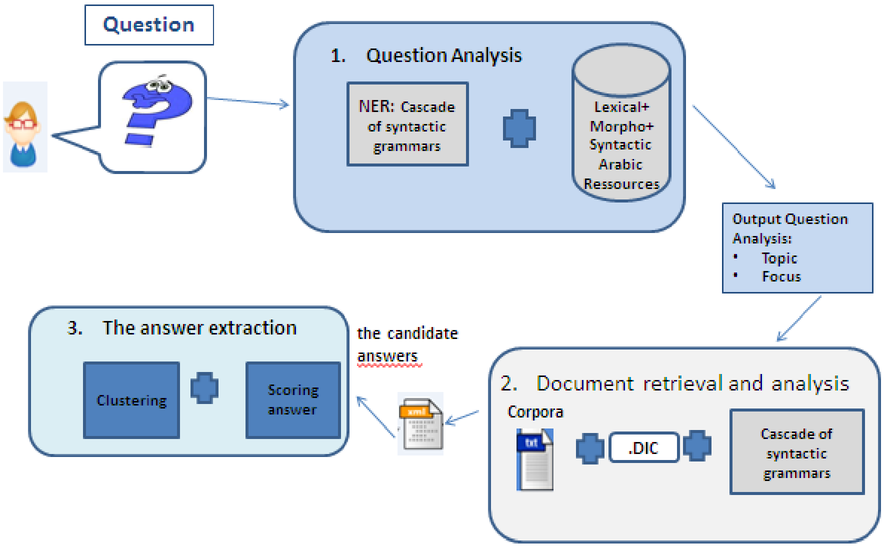

Question
1. Question Analysis
NER: Cascade of syntactic grammars
Lexical+ Morpho+ Syntactic Arabic Ressources
Output Question Analysis:
• Topic
• Focus
2. Document retrieval and analysis
Corpora
txt
.DIC
Cascade of syntactic grammars
the candidate answers
xml
3. The answer extraction
Clustering
Scoring answer

Fig. 1. Architecture of Question Answering System.

the question. In the next section, an overview of the state-of the art describes related works to question answering system. The section 3, we describe the generic architecture of the proposed QA system. In section 4, introduces our approach to annotation of medical factoid question and extraction of right answer.

2 State of Art

As explained in the introduction, Question-Answering systems present a good solution for textual information retrieval and knowledge sharing and discovery. This is reason why a large number of Q-A systems has been developed recently. The table below shows some work of question answering system by criterion.

After this investigation, to solve the problem of question answering system, the developed question answering system is based on a linguistic approach, using NooJ's linguistic engine in order to formalize the automatic recognition rules and then apply them to a dynamic corpus composed of Arabic medical journalistic articles.

The named entity recognizer (NER) is embedded in our question answering system in order to identify these answers and questions associated with the extracted named entities. For this purpose, we have adapted a rules-based approach to recognize Arabic named entities and right answers, using different grammars and gazetteer.

3 Architecture of Question Answering System

From a general viewpoint, the design of a QA system (Fig 1) must take into account three phases:

1. **Question analysis:** This module performs a morphological analysis to identify the question class. A question class helps the system to classify the question type to provide a suitable answer. This module may also identify additional semantic features of the question like the topic and the focus.
2. **Document retrieval and analysis:** The second motivation behind the question classification task is to develop the linguistic patterns for the candidate answers. These patterns are helpful in matching in parsing and identifying the candidate answers.
3. **The answer extraction:** This module selects the most accurate answers among the phrases in each corpus. The selection is based on the question analysis. The suggested answers are then given to the user as a response to his initial natural language query.

4 Our Approach

4.1 Named Entity Recognition (NER)

We think that an integration of a Named Entity Recognition (NER) module will definitely boost system performance. It is also very important to point out that an NER is required as a tool for al-most all the QA system components. Those NER systems allow extracting proper nouns as well as temporal and numeric expressions from raw text (Mesfar, 2007). In our case, we used our own NER system especially formulated for the Arabic medical domain. We have considered six proper names categories:

1. Organization: named corporate, governmental, or other organizational entity.
2. Location: name of politically or geographically defined location.
3. Person: named person or family.
4. Viruses: Names of medical viruses.
5. Disease: Names of diseases, illness, sickness.
6. Treatment: Names of Treatments.

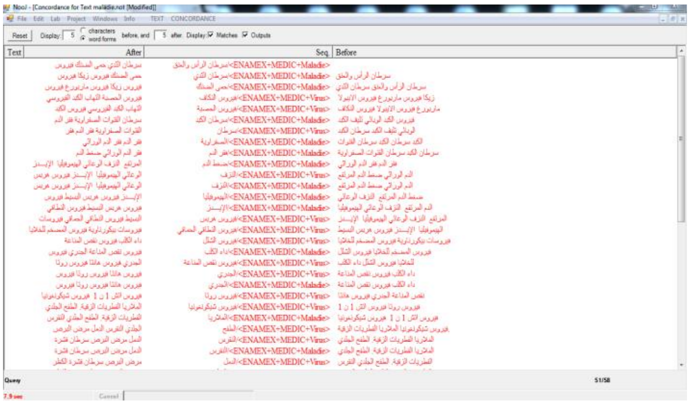
4.2 Automatic Annotation of Factoid Question in Standard Arabic

Our approach focuses on the problem of finding document snippets that answer a particular category of facts-seeking questions namely factoid questions. Simple interrogative questions which await an answer related to a named entity. The choice of factoid questions versus other types of questions is motivated by the following factors:

- Majority of the questions actually submitted to a search engine are factoid questions. Current search engines are only able to return links to full-length

Table 3. NER grammar experiments on our corpus.

Precision	**Recall**	**F-Measure**
0,90	0,82	0,88


NooJ - [Concordance for Text maladie.not [Modified]]
File Edit Lab Project Windows Info TEXT CONCORDANCE
Reset
Display: 5 ○ characters ● word forms before, and 5 after. Display: ☑ Matches ☑ Outputs
Query
51/58
7.9 sec
Cancel

Text	After	Seq.	Before
	سرطان الثدي حمى الضنك فيروس	سرطان الرأس والعنق/<ENAMEX+MEDIC+Maladie>	
	حمى الضنك فيروس زيكا فيروس	سرطان الثدي/<ENAMEX+MEDIC+Maladie>	سرطان الرأس والعنق
	فيروس زيكا فيروس ماربورغ فيروس	حمى الضنك/<ENAMEX+MEDIC+Maladie>	سرطان الرأس والعنق سرطان الثدي
	فيروس الحصبة التهاب الكبد الفيروسي	فيروس النكاف/<ENAMEX+MEDIC+Virus>	زيكا فيروس ماربورغ فيروس الايبولا
	التهاب الكبد الفيروسي فيروس الكبد	فيروس الحصبة/<ENAMEX+MEDIC+Virus>	ماربورغ فيروس الايبولا فيروس النكاف
	سرطان القنوات الصفراوية فقر الدم	سرطان الكبد/<ENAMEX+MEDIC+Maladie>	فيروس الكبد الوبائي تليف الكبد
	القنوات الصفراوية فقر الدم فقر	سرطان/<ENAMEX+MEDIC+Virus>	الوبائي تليف الكبد سرطان الكبد
	فقر الدم فقر الدم الوراثي	الصفراوية/<ENAMEX+MEDIC+Maladie>	الكبد سرطان الكبد سرطان القنوات
	فقر الدم الوراثي ضغط الدم	فقر الدم/<ENAMEX+MEDIC+Maladie>	سرطان الكبد سرطان القنوات الصفراوية
	المرتفع النزف الوعائي الهيموفيليا الإيـدز	ضغط الدم/<ENAMEX+MEDIC+Maladie>	فقر الدم فقر الدم الوراثي
	الوعائي الهيموفيليا الإيـدز فيروس هربس	النزف/<ENAMEX+MEDIC+Virus>	الدم الوراثي ضغط الدم المرتفع
	الوعائي الهيموفيليا الإيـدز فيروس هربس	النزف/<ENAMEX+MEDIC+Maladie>	الدم الوراثي ضغط الدم المرتفع
	الإيـدز فيروس هربس البسيط فيروس	الهيموفيليا/<ENAMEX+MEDIC+Maladie>	ضغط الدم المرتفع النزف الوعائي
	فيروس هربس البسيط فيروس النطاقي	الإيـدز/<ENAMEX+MEDIC+Maladie>	الدم المرتفع النزف الوعائي الهيموفيليا
	البسيط فيروس النطاقي الحماقي فيروسات	فيروس هربس/<ENAMEX+MEDIC+Virus>	المرتفع النزف الوعائي الهيموفيليا الإيـدز
	فيروسات بيكورناوية فيروس المضخم للخلايا	فيروس النطاقي الحماقي/<ENAMEX+MEDIC+Virus>	الهيموفيليا الإيـدز فيروس هربس البسيط
	داء الكلب فيروس نقص المناعة	فيروس الشلل/<ENAMEX+MEDIC+Virus>	فيروسات بيكورناوية فيروس المضخم للخلايا
	فيروس نقص المناعة الجدري فيروس	داء الكلب/<ENAMEX+MEDIC+Maladie>	فيروس المضخم للخلايا فيروس الشلل
	الجدري فيروس هانتا فيروس روتا	فيروس نقص المناعة/<ENAMEX+MEDIC+Virus>	للخلايا فيروس الشلل داء الكلب
	فيروس هانتا فيروس روتا فيروس	الجدري/<ENAMEX+MEDIC+Virus>	داء الكلب فيروس نقص المناعة
	فيروس هانتا فيروس روتا فيروس	الجدري/<ENAMEX+MEDIC+Maladie>	داء الكلب فيروس نقص المناعة
	فيروس اتش 1 ن 1 فيروس شيكونغونيا	فيروس روتا/<ENAMEX+MEDIC+Virus>	نقص المناعة الجدري فيروس هانتا
	الملاريا الفطريات الزقية الطفح الجلدي	فيروس شيكونغونيا/<ENAMEX+MEDIC+Virus>	فيروس روتا فيروس اتش 1 ن 1
	الفطريات الزقية الطفح الجلدي النقرس	الملاريا/<ENAMEX+MEDIC+Maladie>	فيروس اتش 1 ن 1 فيروس شيكونغونيا
	الجلدي النقرس السل مرض البرص	الطفح/<ENAMEX+MEDIC+Virus>	فيروس شيكونغونيا الملاريا الفطريات الزقية
	السل مرض البرص سرطان قشرة	النقرس/<ENAMEX+MEDIC+Virus>	الملاريا الفطريات الزقية الطفح الجلدي
	السل مرض البرص سرطان قشرة	النقرس/<ENAMEX+MEDIC+Maladie>	الملاريا الفطريات الزقية الطفح الجلدي
	مرض البرص سرطان قشرة الكظر	السل/<ENAMEX+MEDIC+Virus>	الفطريات الزقية الطفح الجلدي النقرس

Fig. 4. Result of NER NooJ syntactic grammar.

documents rather than brief document fragments that answer the user’s question.

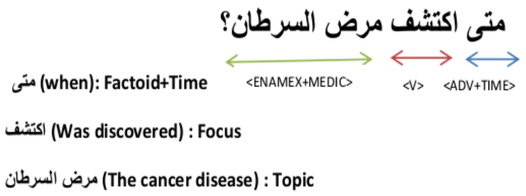
- The frequent occurrence of factoid questions in daily usage is confirmed by the composition of the question test sets in the QA track at Text Retrieval Conference.
- Most approaches of QA use NER as a foundation for detecting candidate answers.

As far as current research is concerned, our QA module accepts, as input, only Arabic factoid questions. Then, in order to look for the best answer, it gives the maximum amount of information (syntactic, semantic, distributional, etc.) from the given question, such as the expected answer the focus and topic of the question. This information will play an important role in the phase of extraction candidate answers.

- **Topic:** the topic corresponds to the subject matter of the question.
- **Focus:** the focus corresponds to the specific property of the topic that the user is looking for.

The following example shows the detailed annotation of the identified parts of a question. Example:

- When was cancer discovered?



متى اكتشف مرض السرطان؟

متى (when): Factoid+Time

اكتشف (Was discovered) : Focus

مرض السرطان (The cancer disease) : Topic

5 Experiments and Results

5.1 Named Entity Recognition (NER)

5.1.1 Evaluation

To evaluate our NER local grammars, we analyse our corpus to extract manually all named entities. Then, we compare the results of our system with those obtained by manual extraction. The application of our local grammar gives the following result:

According to these results, we have obtained an acceptable identification of named entities. Our evaluation shows F-measure of 0.88. We note that the rate of silence in the corpus is low, which is represented by the recall value 0,88 because journalistic texts of our corpus are heterogeneous and extract-ed from different resources.

5.1.2 Discussion

Despite the problems described above, the used techniques seem to be adequate and display very encouraging recognition rates. Indeed, a minority of the rules may be sufficient to cover a large part of the patterns and ensure coverage. However, many other rules must be added to improve the recall.

5.2 Automatic Annotation of Factoid Question in Standard Arabic

5.2.1 Evaluation

To evaluate our automatic annotation question local grammars, we also analyze our user's queries to extract manually the question analysis. Then, we compare the results of our system with those obtained by manual extraction. The application of our local grammar gives the following result:

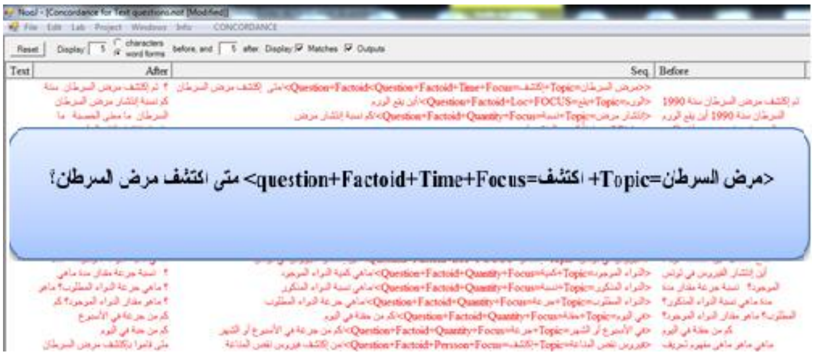
According to these results, we have obtained an acceptable annotation of question. Our evaluation shows F-measure of 0.73. We note that the rate of silence in the corpus is low, which is represented by the recall value 0.72. This is due to the fact that this assessment is mainly based on the results of the NER module.

5.2.2 Discussion

Errors are often due to the complexity of user’s queries sentences or the absence of their structure in our system In fact, the Arabic sentences are usually very long, which sets up obstacles for the question analysis. Despite the problems described above, the developed method seems to be adequate and shows very encouraging extraction rates. However, other rules must be added to improve the result.

Table 4. Annotation question grammar experiments.

Precision	Recall	F-Measure
0,75	0,72	0,73


NooJ - [Concordance for Text questions.not [Modified]]
File Edit Lab Project Windows Info CONCORDANCE
Reset
Display 5 ○ characters ● word forms before, and 5 after. Display ☑ Matches ☑ Outputs
Text
After
Seq.
Before

<مرض السرطن=Topic+ اكتشف=question+Factoid+Time+Focus> متى اكتشف مرض المرطان؟

Fig. 5. Result of Annotation NooJ syntactic grammar.

6 Conclusion

Arabic Question Answering Systems could not match the pace due to some inherent difficulties with the language itself as well as upon to the lack of tools offered to support the researchers. The task of Question Answering can be divided into three phases:

- Question analysis,
- Document retrieval,
- Analysis, and answer extraction.

Each of these phases plays crucial roles in overall performance of the Question Answering Systems. As a future work, we work on the two other phases of Question Answering Systems; that is “Answer Extraction” and “Document Analysis”. In Document Analysis, we can look for such methods used in information retrieval including tools, evaluation, and corpus. In Answer Analysis, we can look for such methods used in this phase including evaluation, tools, and corpus. Finally, as a long-term ambition, we intend to consider studying the processing of the “why” and “how” question types.

References

1. Benajiba, Y., Rosso, P., Lyhyaoui, A.: Implementation of the ArabiQA Question Answering System's components. In: Proc. Workshop on Arabic Natural Language Processing. In: 2nd Information Communication Technologies Int. Symposium, ICTIS-2007, Fez, Morroco (2007)

2. Ben A. F.: Un système de Question Réponse coopératif sur le web. Thèse de doctorat de l'université Paul Sabatier (2014)
3. Chu-Carroll, J., Krzysztof, C., Prager, J., G., S. B.: IBM's PIQUANT II in TREC2004. In: TREC-13 (2002)
4. Doaa S., Moreno-Sandoval, A., Bueno-Díaz, C., Garrote-Salazar, M., Guirao, J. M.: Medical term extraction in an Arabic medical corpus. In: Proceedings of LREC'12 (2012)
5. Ferret, O., Grau, B., Illouz, G., Jacquemin, C., Masson, N.: QALC - the Question Answering Program of the Language and Cognition Group at LIMSI-CNRS. In: Proceedings of the 8th Text Retrieval Conference, NIST Special Publications, pp. 465–475 (1999)
6. Hammou, B., Abu-Salem, H., Lytinen, S., Evens, M.: A Question answering system to support the Arabic language. In: Proc. of the workshop on Computational approaches to Semitic languages, ACL, pp. 55–65 (2002)
7. Laurent, D., Séguéla, P., Nègre, S.: Cross lingual question answering using Qristal for Clef. In: Working Notes (2005)
8. Mohammed, F. A., Nasser, K., Harb, H. M.: A knowledge-based Arabic Question Answering System (AQAS). In: ACM SIGART Bulletin, pp. 21–33 (1993)
9. Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R., Rus, V.: The Structure and Performance of an Open-Domain Question Answering System. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 563–570 (2002)
10. Nyberg, E, Mitamura, T.: Evaluating QA systems on multiple dimensions (2002)
11. Plamondon, L., Kosseim, L., Lapalme, G.: The quantum question answering system at trec-11. In: Proceedings of the Eleventh Text Retrieval Conference (TREC-2002), pp. 750–757 (2002)
12. Neifar, W., Ben-Ltaief, A.: Acquisition terminologique en arabe: État de l'art, Actes de la conférence conjointe JEP-TALN-RECITAL (2016)
13. Silberztein, M.: NooJ manual. Available at the WEB site http://www.nooj4nlp.net (2003)
14. Silberztein M.: NooJ's Linguistic Annotation Engine. INTEX/NooJ pour le Traitement Automatique des Langues, Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comté, pp. 9–26 (2006)