Factoid and definitional Arabic Question Answering system

Wissal Brini, Mariem Ellouze, Omar Trigui
ANLP Research Group- MIRACL Laboratory, University of Sfax, Tunisia.
wissal.brini@voila.fr, mariem.Ellouze@planet.tn, omar.trigui@fsegs.rnu.tn

Slim Mesfar
ISI, University of El-Manar, Tunisia.
mesfarslim@yahoo.fr

Lamia Hadrich Belguith
ANLP Research Group- MIRACL Laboratory, University of Sfax, Tunisia.
l.belguith@fsegs.rnu.tn

Paolo Rosso
Natural Language Engineering Lab. - ELiRF, Universidad Politécnica de Valencia, Spain
prosso@dsic.upv.es

ABSTRACT. *The amount of available information is becoming very huge, especially with the Web proliferation. The problem faced by the user is not the lack of documents or information but is the lack of time to find a short and precise answer among the variety of available documents. Search engines offer a lot of links toward web pages, but are not able to provide an exact answer. Thus, a new need is emerged: the possibility of obtaining a unique, brief and concise answer. It is the main goal of Question Answering systems. The proposed Question Answering deals factoid and definition question. It employs the NooJ platform which represents a valuable linguistic development environment.*

1. Introduction

1.1. Question Answering systems

Question Answering (QA) systems are relatively advanced for Latin languages. Among them, we cited QALC (Ferret et al., 2001) which is a QA system for English factoid questions in open domain. This system uses a syntactic and semantic analysis for each question. Nevertheless, it presents some errors due to incomplete syntactical rules; QUANTUM (Plamondon, 2002) is a bilingual QA system that

takes a natural language question as input, looks for the answer in a one million document collection, and generates five answer suggestions of up to 50 characters each. The question analysis module is robust, but the bilingualism causes a drop in its performance; Œdipe (Balvet et al., 2005) is a French QA system which applies a set of morpho-syntactical patterns. However, it presents a minimalist approach with respect to the used tools.

1.2. Arabic Question Answering systems

QA systems for Arabic language are very few. Mainly, it is due to the lack of accessibility to linguistic resources, such as corpora and basic NLP tools (tokenizers, morphological analyzers, etc.). Moreover, the Arabic language has a very complex morphology (inflectional and derivational characteristics) and texts suffer from the scarcity of vowels as well as the absence of capitalization. These specificities of the Arabic language introduce many processing problems related to the word tokenization, the identification and categorization of named entities, etc.

To our knowledge, there are three research works on Arabic Question Answering. AQAS (Mohammed et al., 1993) is a knowledge-based QA system that extracts answers only from structured data and not from raw text. QARAB (Hammo et al., 2002) is an Arabic QA system that uses both information retrieval and natural language processing techniques. The QARAB system reached a precision of 97.3% and a recall of 97.3%. The evaluation was done directly by four native Arabic speakers who presented 113 questions to the system and judged themselves the correctness of the answers. Note that such accuracy was not achieved in any other language in the QA state-of-the-art. We think that the obtained results could be reliable if a test-bed of questions in Arabic were provided in order to allow a comparison between different QA systems. ArabiQA (Benajiba, 2007) is an Arabic QA system that embeds an Arabic Named Entity Recognition (NER) system called *ANERsys* and it adapts the *Arabic-JIRS* to extract passages from Arabic documents.

2. Factoid questions

This section 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.

Examples of such questions are “How much costs a BMW?” or “Who is the president of USA?”, some other question are presented in Tabe 1.

The QA system which will be described in the next section recognizes the set of question types illustrated in Table 2.

Factoid answers	Factoid questions
1332	متي ولد ابن خلدون ؟
	When was Ibn Khaldoun born?
مصر	أين يوجد جامع الأزهر ؟
Egypt	Where is Al-Azhar Mosque?

Table 1. *Example of Factoid question*

The choice of factoid questions versus other types of questions is motivated by the following factors: (i) a considerable percentage of the questions actually submitted to a search engine are factoid questions. Current search engines are only able to return links to full-length documents rather than brief document fragments that answer the user's question. (ii) The frequent occurrence of factoid questions in daily usage is confirmed by the composition of the question test sets in the QA track at TREC[1]. The percentages of questions that are factoid questions grew in TREC. (iii) Most recent approaches to open-domain QA use NER as core process for detecting candidate answers.

Question forms		**Type of Answer expected**	
Who+…….. ?	من هو.....؟	Person name	شخص
When +….. ?	متى .....؟	Temporal expressions	زمان
Where +…. ?	أين .....؟	Localization names	مكان
How much +…. ?	كم .....؟	Numeric expressions	كميّة

Table 2. *Types of the factoid question processed*

In fact, the main task of the NER process is to allow the identification and

[1] Text Retrieval Conference http://trec.nist.gov/

categorization of proper names as well as temporal and numeric expressions, in an open-domain text. NER systems proved to be very important for many tasks in NLP such as IR and QA tasks. NER process was firstly introduced during the MUC-6[2]. In this conference, three types of entities distinguished to be recognized and categorized (Grishman and Sundheim, 1995).

- *ENAMEX:* detection and classification of proper names and acronyms. The classes considered in this subtask are:

- Organization: named corporate, governmental, or other organizational entities such as "Bridgestone", "Mips" or "Language Computer Corporation".

- Person: named person or family such as "Mahatma Ghandi", "Marie Curie" or "Bill Clinton".

- Location: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.) such “Morocco”, “Italy” or “Spain”.

- *TIMEX:* Detection and classification of temporal expressions. The classes considered in this subtask are:

- Date: complete or partial date expression such as "January 2008", "summer".

- Time: complete or partial expression of time of day such as "5 p.m.", "eleven o'clock" or "12h45 a.m.".

- *NUMEX:* Detection and classification of numeric expressions monetary expressions and percentages. The classes considered in this subtask are:

- Money: monetary expression such as "9,000 Euros", "million-dollar" or "$16,000".

- Percent: percentage such "5%", "20 pct" or "20.3%".

The problem of proper names identification is particularly difficult for Arabic language (Benajiba et al. 2008). This is due to some specific problems related to Arabic NER:

- *Non-Vocalization:* It is due to a lack of short vowels in usual texts from which a high degree of ambiguity ensues. In theory, only the Koran, and children’s books are fully vowelled; our automatic analysis allows parsing of fully vowelled, partially vowelled and unvowelled texts. Non-vocalisation can affect NER systems when potential vocalizations can lead to different senses which can designate trigger

[2] The 6th Message Understanding Conference

words for two or more different NE types such as the case of unvowelled form "مؤسسة" [mowass'sat] that can accept, between others, the two vocalizations:

" مُؤَسَّسَة " [mowassasat – a company] => trigger word of an organization name.

" مُؤَسِّسَة " [mowassisat – a founder] => trigger word of a personal name.

- *Lack of capitalization:* The problem of identifying named entities is particularly difficult for Arabic, since names in the Arabic language do not start with capital letters and, therefore, we cannot mark them in the text by looking at the first letter of the word (Mesfar, 2007).

- *Delimitation problems:* They are related to the lack of information about unknown words with NEs, an antonomastic usage where proper names are substituted with a phrase or conversely as well as the presence of some homonyms which increases ambiguity when trying to mark NE constituents (Plamondon, 2002) such as:

- "أَشْرَف" [achrafa] which can be a first name, an inflected verbal form meaning "he supervised", an elatif adjective which means "the most honorable", etc.

- "أَحْمَد" [ahmadu] which can be a first name, an inflected verbal form meaning "I thank", etc.

3. QASAL architecture

In this section, we describe our factoid QA system named *QASAL* for Arabic language. This QA system uses natural language processing techniques to processing Arabic factoid questions.

QASAL has a pipeline architecture consisting of three components: Question Analysis, Passage Retrieval, and Answer Extraction modules (Figure 1).

(i) Question Analysis module: this module accepts, as input, any Arabic factoid question. Then, in order to look for the best answer, it gives the maximum amount of information from the concerned question, such as the expected answer type (used by the Answer Extraction module), the question focus (i.e., the named entities appearing in the question, which can play an important role in the extraction of potential answers) and the list of important keywords (used by the passage retrieval module as a query). For instance, the analysis of the question: متى استقلت تونس ؟ (When Tunisia became independent?) generates the following information: expected answer type is time entity, focus of the question is تونس (Tunisia-noun-) and the keyword is استقلّت (become independent-verb-).

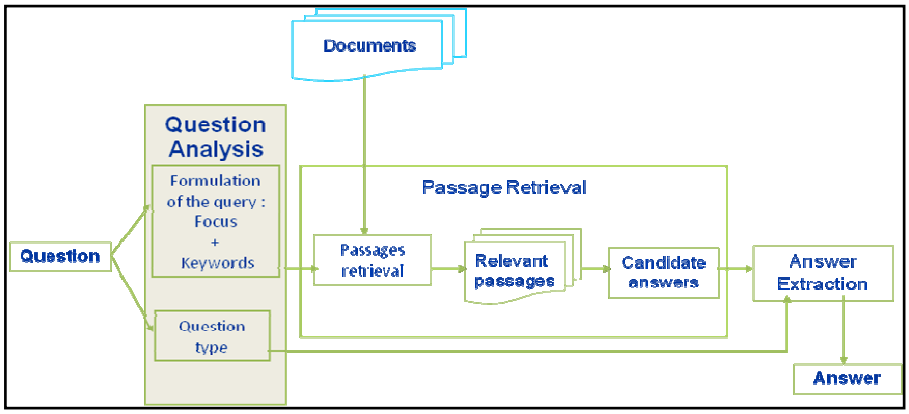

Figure 1. *QASAL architecture*

Figure 2 shows the NooJ's text annotation structure that gives the linguistic analysis of each word form in our sample question: متى استقلّت تونس ؟ ("When Tunisia became independent?").

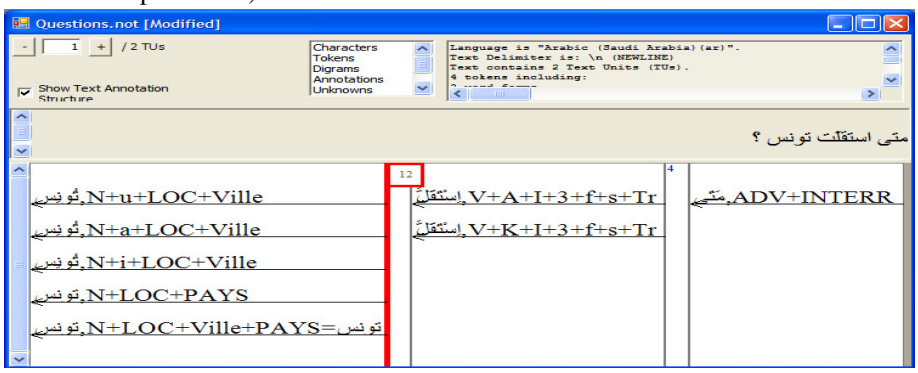

Figure 2. *Text Annotation Structure of* متى استقلّت تونس ؟

This analysis is carried out through a couple of syntactic grammars built using the linguistic platform NooJ. These syntactic grammars could translate each input question into one or more regular expressions representing its corresponding patterns (Figure 3).

(ii) Passage Retrieval module: this second module is the core of the QA system. It retrieves the passages which are estimated as relevant to contain the expected answer. Indeed, text passages are an important intermediary between full documents and exact answers. They form a very natural unit of response for QA systems. Thus, the quality of a Q/A system heavily depends on the effectiveness of the second step of the pipeline: if the system fails to find any relevant passage for a question, further processing steps to extract an answer will inevitably fail too.

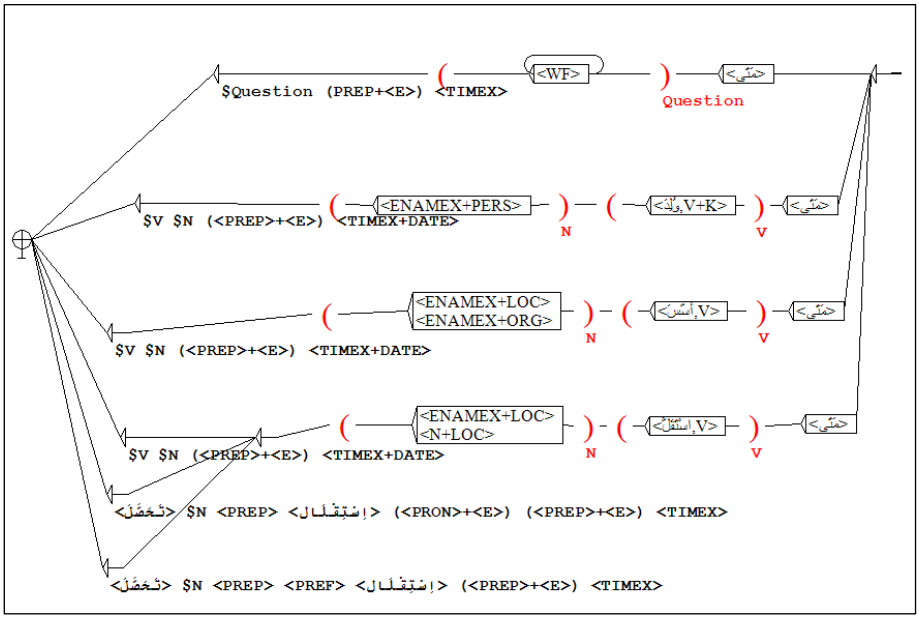

Figure 3. *NooJ syntactic grammar (Time sub-graph)*

(iii) Answer Extraction module: this module extracts the answer from the retrieved passages taking into account the constraint of the Question Analysis module. The task has to take into consideration the type of answer expected by the user, and this means that the Answer Extraction module should perform differently for each type of question.

The third module uses the displayed concordance table to filter the number of context's tokens. Then, we use NooJ's facilities to export the retained token sequences. The result is saved in a text format (.txt) or an XML file.

Figure 4 shows the concordance table for answer to the factoid analyzed question: متى استقلّت تونس؟ ("When Tunisia became independent?").

The corpus used was extracted from Tunisian books for basic education. It is first segmented by the STAr tokenizer (Belguith et al., 2005) and then tagged by the MORPH-2 morphological analyser (Belguith et al., 2006). This corpus is used to constitute the pairs of questions-answers.

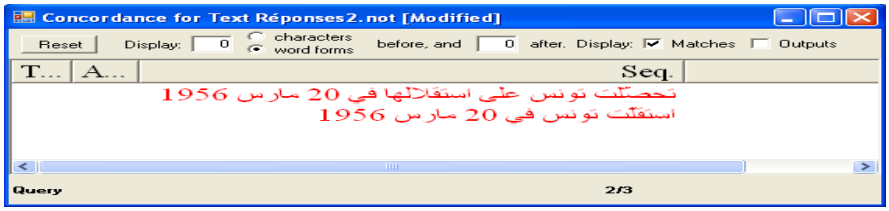

Figure 4. *Concordance table for the answer*

4. Definition questions

Generally, a definition question ask for a profile of a person, full name of an organization or a concept. They account for a significant number of queries submitted to Web search engines (Figueroa and Atkinson, 2009) as for the factual questions.

In Table 3 we present examples of definition questions about a person and an organization.

Definition answers	Definition questions
مؤسس علم الاجتماع ومؤرخ مسلم من إفريقية في عهد الحفصيين وهي تونس حالياً ترك تراثاً مازال تأثيره ممتداً حتى اليوم.	من هو ابن خلدون ؟
Ibn Khaldoun is considered a forerunner of several social scientific disciplines. He is born in North Africa in present-day Tunisia.	Who is Ibn Khaldoun?
هو من أهم المساجد في مصر و أشهرها في العالم الإسلامي. وهو جامع وجامعة منذ أكثر من ألف عام	ما هو جامع الأزهر ؟
Al-Azhar Mosque is one of the most important mosques in Egypt and in the Islamic world. A mosque and the University for more than a thousand years.	What is Al-Azhar Mosque?

Table 3. *Two examples of definition questions with their respective answers.*

The difference between a definition question and a factoid question are principally the interrogative pronoun used (e.g., “who” and “what”) and the number of words. In a definition question, globally there is few words in a questions: interrogative pronoun/verb to be/topic/ ? (Table 3).

Question patterns		Type of Answer pattern expected	
Who+be+<topic> ?	**من هو <topic>؟**	Person	شخص
Who+be+<topic> ?	**من هي <topic>؟**	Person	شخص
What+be+<topic> ?	**ما هي <topic>؟**	Organisation	مؤسسة

Table 4. *Patterns expected for a definition question*

4.1. *Definition Question Answering system*

In this section, we detail the modification made to the factoid QA system QASAL to process the definition question. Figure 5 illustrate the new architecture of the version of QASAL dealt Arabic definition questions.

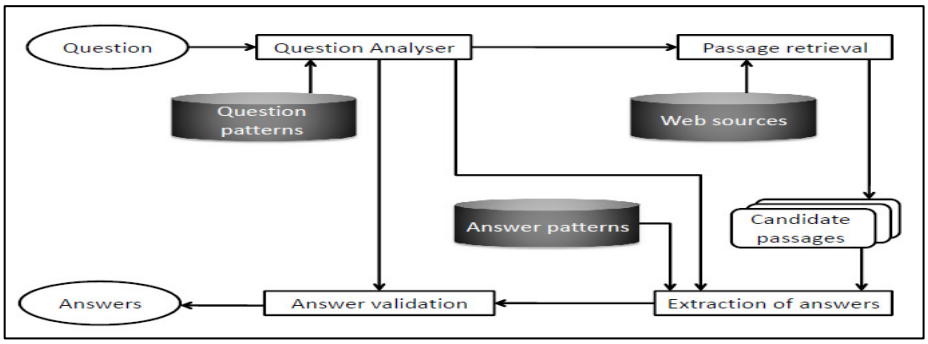

Figure 5. *Architecture of the definition QA.*

The main modification done to QASAL are as follows: (i) the use of question patterns to extract the focus of the question, (i) the use of the search engine Google and the Web as resource, and finally the use of a list of lexical patterns to extract the specific answers.

4.2 *Preliminary experiments*

We have used Google search engine as Web resource and 43 definition questions. We have not used the collection of Tunisian books as corpus as for we have used to test of the factoid QA system, because this collection don't contain pairs of definition questions and answers.

The preliminary results obtained are recall equal to 100% and precision equal to 94%.

5. Conclusion and future work

In this paper, we have proposed an Arabic Question Answering System called QASAL for factoid and definition questions using NooJ local grammars. QASAL takes advantage of some linguistic techniques from IR and NLP to process a collection of Arabic documents and the Arabic version of search engine Google. The overall success of the system is quite encouraging with the NooJ's platform. We have obtained a precision equal to 94% for the definition questions.

As future work, we plan to investigate more to study the influence of the platform NooJ in the Arabic QA systems.

Acknowledgments

This research work started thanks to the bilateral Spain-Tunisia research project on "Answer Extraction for Definition Questions in Arabic" (AECID-PCI B/017961/08). The work of the sixth author was carried out in the framework of the AECID-PCI C/026728/09 and the MICINN TIN2009-13391-C04-03 (Plan I+D+i) research projects.

References

Balvet A, Embarek M and Ferret O. 2005. Minimalisme et question-réponse: le système OEdipe. TALN 2005, Dourdan, 6-10 juin 2005.

Bahou Y., Belguith H.L, Aloulou C., Ben Hamadou A, (2006), "Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés", In : 15th AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, January 2006 Tours - France. pp 25-27.

Belguith H.L, Baccour Leila, Mourad Ghassan, "Segmentation de textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules". In : TALN'2005, Dourdan, France, 6-10 Juin 2005, pp 451–456.

Benajiba Y. 2007. Arabic Question Answering. Diploma of advanced studies, Technical University of Valencia, Spain.

Benajiba Y. Diab M. and Rosso P. 2008. Arabic Named Entity Recognition using Optimized Feature Sets. In: Proceeding of the Conference on Empirical Methods in Natural Language Processing, EMNLP-2008, Waikiki, Honolulu, USA.

Ferret O, Grau B and Hurault-Plantet M. 2001. Finding an answer based on the recognition of the question focus. In: Proceedings of TREC 10, Gaithersburg MN.

Figueroa A., and Atkinson J. 2009. Using Dependency Paths for Answering Definition Questions on the Web. In: Proceeding of the 5th International Conference on Web Information Systems and Technologies.

Grishman R and Sundheim B. 1995. Design of the MUC-6 Evaluation. In: Proceedings of the 6th Conference on Message Understanding, pp 1-11.

Hammo B, Abu-Salem H and Lytinen S. 2002. QARAB: A Question Answering System to Support the Arabic Language. In: Proc. Of the workshop on computational approaches to Semitic languages, ACL, pages 55-65, Philadelphia.

Mohammed F.A, Nasser K and Harb H.M. 1993. A knowledge-based Arabic Question Answering System (AQAS). In: ACM SIGART Bulletin, pp. 21-33.

Mesfar S. 2007. Named Entity Recognition for Arabic Using Syntactic Grammars. In: Natural Language Processing and Information Systems: Springer Berlin / Heidelberg, p. 305-316.

Plamondon L. 2002. Le système de question-réponse QUANTUM. Mémoire de maîtrise, University of Montreal. Mars 2002