A Temporal Arabic Question Answering System based on the construction of a temporal resource

Mayssa Mtibaa[1] Zeineb Neji[2] Mariem Ellouze[3]
Faculty of Economics and Management, 3018, Sfax Tunisia
Computer department, Miracl laboratory, University of Sfax

[1]maysamtibaa@gmail.com, [2]zeineb.neji@gmail.com,
[3]mariem.ellouze@planet.tn

Abstract. Approaches in Arabic temporal question answering systems are in their first steps compared to other languages. They are often confronted to complex temporal information or a real ambiguity. Our goal is to solve this problem by developing an Arabic temporal resource. This paper deals with answering questions about temporal information involving several forms of inference to obtain a relevant answer.

Ar-TQAS, an Arabic Temporal Question Answering System based on the construction of a temporal resource, has been implemented.

Keywords: Question answering system, Arabic language, temporal inference, Arabic temporal resource.

1 Introduction

In general, a Question Answering (QA) system aims to generate a precise answer to a given question in natural language. Indeed, many approaches are developed in this context. They deal with different domains, various types of questions and use different information resources (for example: a database, a collection of documents, a knowledge base and the Web) [1]. In this paper, we focus on the task of answering to temporal question in Arabic. We propose a new method which allows improving the performance of traditional Arabic temporal question answering systems. This new method is based on the construction of an Arabic temporal resource for the resolution of temporal inferences. Therefore, the challenge of developing a system capable of obtaining a relevant and concise answer is obviously of great benefit [2]. The challenge becomes huge when we try automatically process a complex natural language such Arabic. This complexity is mainly due to the inflectional nature of Arabic. Moreover, in our chosen field, research on temporal entity extraction in English, German, French, or Spanish, uses local grammars, finite state automata [3] and neural networks [4] to detect temporal entities. These techniques do not work well directly for Arabic due mainly to the rich morphology and high ambiguity rate of Arabic temporal information [5]. Thus we propose a solution to infer complex temporal information through an Arabic temporal resource.

This paper is organized as follows: in the next section, we give some earlier and related researches in Arabic question answering systems. Then, we present the temporal notion in the third section. After that, in the section 4, we detail our proposed method and its different stages. While section 5 details the experiment setup and the obtained results. And we close with a conclusion of this work and make suggestions for future researches.

2 Related works and motivations

In this section, we present previous work related to question answering systems in Arabic language. These systems are still few compared to the number of those developed in English, French languages.

In fact, the technology of Arabic question answering has been dealt since the 1993s with the AQAS system presented in [6]. This is the first system for the Arabic language. It is a knowledge based QA system that allows returning answers to questions (e.g.: Who, What, Where, When) from structured data and not from a raw text. ArabiQA (Arabic question answering system) is an Arabic question answering system developed by [7] that deals with factoid questions. It is based on a generic architecture based on three integrated modules; a NER (Named Entity Recognition) module, a passage retrieval system (JIRS) and an Answer Extraction (AE) module. Another work focusing on named entities provides AQuASys [8] which is a Question Answering System that makes it possible for the user to ask a question formulated in Arabic natural language. The correct answering question is related to a named entity that can be of any type: person, place, organization, time, quantity, location, etc. AQuASys takes as input questions starting with interrogative words (e.g.: من /who, ما /what, أين /where, متى /when, كم العددية /how many, كم الكمية /how much). QARAB [9] is a QA system that takes a set of questions in Arabic language and provides short answers. The source of knowledge of this system is a collection of Arabic newspaper texts taken from the AL-RAYA newspaper published in Qatar. QARAB uses superficial language understanding to deal with problems and it doesn't attempt to understand the content of the question at a deep semantic level. A QA system that deals with definitional questions is DefArabicQA (Arabic definitional Question Answering System) developed by [10]. The system searches for candidate definitions using a set of lexical patterns and categorizes them using heuristic rules. DefArabicQA classifies definitions using a statistical approach. Bakari et al, develop a new approach to questions answering system based on the automatic understanding of Arabic texts to transform them into representation logic [11]. They use techniques of textual entailment recognition.

The challenge becomes important when we try to create capabilities of the processing and recognition of Arabic temporal information for the answers to the questions. When the question asked by the user refers directly or indirectly to a temporal expression, the answer is expected to validate the temporal constraints. In this context, TPE [12], an Arabic QA system, has been proposed based on temporal inference. It deals with the relationships between temporal expressions and events mentioned in the question and relies on temporal inference to justify the answer. This system consists of answering questions that start with temporal signals by an answer pattern. During this work, Omri et al, found problems at the level of the temporal inference [12]. In the same context, we present a new method based on Arabic temporal resource for the resolution of temporal inferences. For example, the expected answer type of question Q1 is a Date:

Q1 : متى يحبذ حرث الأرض ؟
Q1: When is it favored to plow the land?

The expected answer type is an argument of the first event Evt1= يحبذ / favored and the second event Evt2= حرث الأرض / plow the land.
The answer A1: الليالي السود / Black Nights, extracted from the context P1:

P1 : يستحب حرث الأرض في الليالي السود و سميت بالسود لأنها تتميز بكثرة الغيوم والسحب مما يجعل الطقس باردا في النهار ودافئا أثناء الليل.

P1: It is desirable to plow the land in black nights and it’s called black nights because it is characterized by over clouds, which makes the weather cold in the day and warm during the night.

In the paragraph P1, the event Evt2 = يستحب / desirable which has an argument حرث الأرض / plow the land. The event Evt1 differs from the event Evt2, but they are related.
We need an analysis to get the simple temporal answer (e.g. 14 جانفي/ 14th January). This intelligent analysis is called Temporal Inference. It is concerned with building systems that automatically answer questions in a natural language by extracting a precise answer from a corpus of documents.
But, for the example above, the answer A1 presents complex temporal information which is not extracted by the analysis of temporal inference; it needs to refer to a temporal resource.

3 The temporal notion in Arabic language

There are several motivating factors for the choice to use the Arabic temporal notion. Such factors are:

- Arabic is a very rich and complex language,
- Temporal information is an important dimension of any information space,
- Arabic temporal information is an essential component in text comprehension and useful in a large number of automatic language processing applications,
- The temporal entities are expressed in different ways,
- For temporal information, we find several representations. This causes ambiguity.

Example of ambiguity: For the example of الثورة التونسية / Tunisian Revolution can be expressed as follows:

17 / 12 / 2010
17- 12 - 2010
2010 ديسمبر 17 / 17 dysmbr 2010
2010 كانون الأول 17 / 17 kAnwn Al>wl 2010
٢٠١٠ كانون الأول ١٧ / ١٧ kAnwn Al>wl ٢٠١٠
2010 السابع عشر من ديسمبر / AlsAbE E$r mn dysmbr 2010

For the word «ديسمبر /December/dysmbr», we also find the following words which are equivalent « دجنبر/ djnbr, ذو الحجة / * / w AlHjp» and for the year 2010 we can also find the year 1432/1432 هجري hjry /1432 Hijri or 2010/2010 ميلادي mylAdy /2010 gregorian.

An another example of ambiguity: Table 1 shows an example of the multiple designation of the months of the year.

Table1. Example of month representation variety

<table>
<tr><th>English</th><th>Arabic</th></tr>
<tr><td>March : مارس</td><td>مارس\ mArs ,آذار \|*Ar, المريخ \ Almryx</td></tr>
<tr><td>June : جوان</td><td>يونيو \ ywnyw ,يونيه\ ywnyh ,حزيران \ HzyrAn</td></tr>
<tr><td>August : أوت</td><td>أوت \ >wt , أغسطس\ >gsTs,غشت \g$t, آب \ |b</td></tr>
<tr><td>September : سبتمبر</td><td>سبتمبر\ sbtmbr, شتمبر\ $tmbr,شتنبر \ $tnbr,أيلول \ >ylwl</td></tr>
<tr><td>November : نوفمبر</td><td>نوفمبر\ nwfmbr ,تشرين الثاني \ t$ryn AlvAny, نونبر \ nwnbr</td></tr>
</table>

To extract temporal information, a model is needed to precisely define such information in documents (or rather the corresponding textual expressions) and to classify them. The classification is based on how the temporal terms are presented.

4 Proposed method

In this section, we propose a method for a question answering system based on the construction of a temporal resource for the resolution of temporal inference for the Arabic language. The proposed method involves three main stages presented in figure 1, namely: (1) question analysis, (2) document processing, (3) answer processing. In the following subsections, we will detail the different sub-stages used in this proposed method.

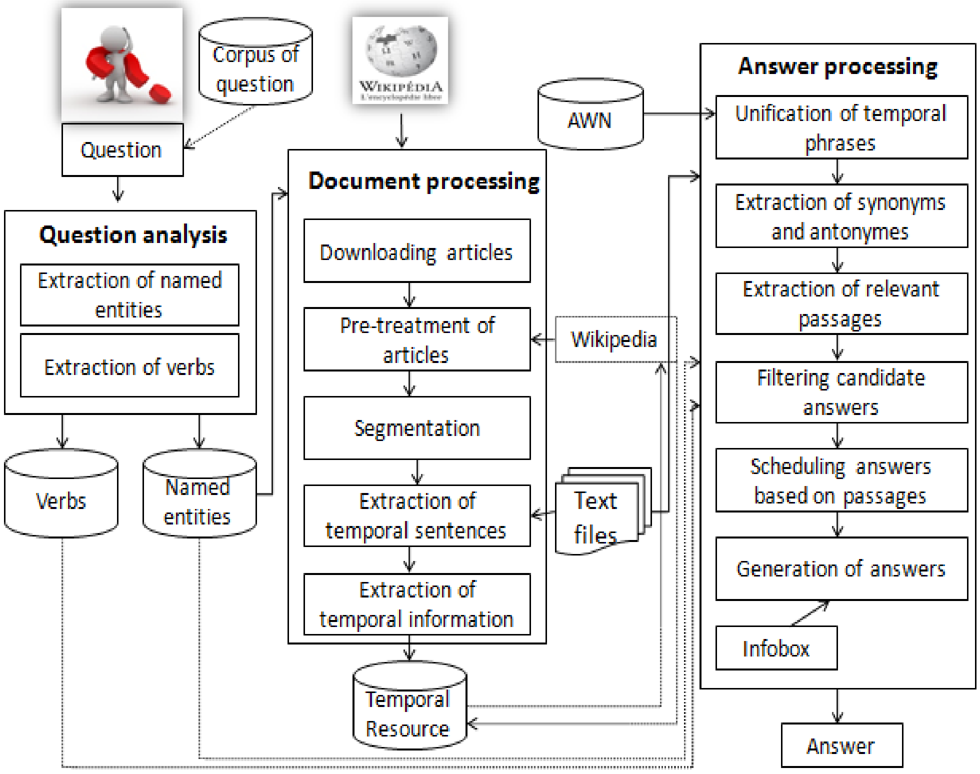
Corpus of question
Question
AWN
Answer processing
Unification of temporal phrases
Document processing
Extraction of synonyms and antonymes
Question analysis
Downloading articles
Extraction of named entities
Extraction of relevant passages
Pre-treatment of articles
Wikipedia
Extraction of verbs
Filtering candidate answers
Segmentation
Verbs
Named entities
Scheduling answers based on passages
Text files
Extraction of temporal sentences
Generation of answers
Extraction of temporal information
Infobox
Temporal Resource
Answer

Fig. 1. Process of the proposed method

4.1 Question Analysis

The objective of question analysis is to understand asked question. This stage consists of representing the characteristics of the question which can facilitate us the treatment of the following stages of the system. This analysis focuses on the extraction of named entities and verbs.

Indeed, our input is a question bearing temporal information only who starts with one of the signals presented in Table 2 and the other types of questions are not part of the subject of our research.

Table2. Temporal signals

Temporal signals in Arabic	Temporal signals in English
متى	When
منذ متى	Since when
كم (دام - مكث - أقام - كان - أصبح)	How(long-stay-many-much)
في أي (عام - وقت - زمن - شهر - يوم - تاريخ)	In Which (year-time-day-date-month-era)
إلى أي(عام - وقت - زمن - حقبة - شهر - يوم - تاريخ)	Until which(year-time-day-date-month-era)

Extraction of named entities.
Named entities are presented as a word or a group of words as semantically textual objects that can be names of people, names of organizations or companies, place, names, dates, quantities, distances, monetary units, percentages.

The detection of named entities (NE) in Arabic is a potential pretreatment and represents a serious challenge taken into account the specificities of the Arabic language. We proceed at this level to the extraction of the NE that occurs in each question for the purpose of building an NE base that will be useful to the following stage of answer processing.

Extraction of verbs.
The extraction of verbs is based on the decomposition of the question in order to extract the verb that exits. This verb will be stored in a base of verbs that will be useful for the following stage of answer processing.

4.2 Document Processing

Downloading articles.
This step consists in downloading articles from the online encyclopedia Wikipedia. In fact, we automatically download the articles which contain the extracted named entities (NE) from the previous stage and we will retrieve the Infobox, if it is existed because we will use it after for the verification of candidate answers. The structuring of Wikipedia articles requires pretreatment, a segmentation that leads to the extraction of relevant sentences.

Pre-treatment of articles.
We keep the textual content of previously downloaded articles. The content of the article requires a pre-treatment like; elimination of empty words that are presented as parentheses, symbols, spacious characters and words that are not in Arabic as well as links and images.

Segmentation.
Segmentation is a linguistic pretreatment of one or more cleaned texts to process them later. It consists of dividing the text into paragraphs and then into sentences. We applied the approach proposed by [13]. This approach is based on the contextual exploration of markers of punctuations, connective words as well as some particles such as coordination conjunctions.

Extraction of temporal sentences.
A determined selection of sentences contains one or more temporal information. This task consists of filtering useful information that is considered relevant to facilitate the retrieval of answers.

Extraction of temporal information.
This step consists on extracting all the existing temporal information in the sentences and presenting them according to their categorization in our temporal resource.

Construction of temporal resource.
This is the fundamental task of our research work; it’s a temporal resource specific to the Arabic language. It allows to identify and unify all the extracted complex temporal information substituted into categories (as illustrated in table 3), which purpose is to solve all the problems handled in temporal inferences and to facilitate the acquisition of the relevant answer.

Table3. Categorisation of time information

CATEGORY	NUMBER
Solar months : الأشهر الشمسية/ Al>$hr Al$msyp	42
Lunar months : الأشهر القمرية/ Al>$hr Alqmryp	12
Dates: التواريخ/ AltwAryx	1860
Hijri Year: سنة هجرية / snp hjryp	31
Gregorian Year: سنة ميلادية/ snp mylAdyp	1910
Agricultural seasons : المواسم الفلاحية/ AlmwAsmAlflAHyp	24
Feasts : الأعياد/ Al>EyAd	73
Temporal markers : علامات الوقت/ ElAmAt Alwqt	31
Islamic dates : التواريخ الإسلامية/ AltwAryx Al<slAmyp	56

We collected a set of Arabic temporal information of different categories from the Wikipedia, Internet, and through the temporal sentences extracted from NOOJ in order to build a well-structured database (table 4) named Temporal Resource. Its purpose is to facilitate the identification of temporal information, to solve all ambiguities and to have a quick answer in a standard format.

Table4. Extract from the temporal resource

TEMPORAL INFORMATION	Matches	Unification
غشت/ g$t	أوت – أغسطس -أب	أوت : August
نونبر / nwnbr	تشرين الثاني - نوفمبر	نوفمبر : November
قرة العنز/ qrp AlEnz	14فيفري - 14 فبراير - 14 شباط - 14 فورار	14 فيفري : 14th February
اليوم العالمي للمرأة /Alywm AlEAlmy llmr>p	8مارس - 8 آذار - 8 المريخ	8مارس : 8 March
الثورة التونسية/ Alvwrp Altwnsyp	17 دجنبر 2010 -17 كانون الأول 2010	17 ديسمبر 2010: 17th December 2010
الفاتح / AlfAtH	الأول – مطلع -غرة	1 : the 1st

4.3 Answer Processing

Unification of temporal sentences.
A task of unification is applied to the extracted temporal sentences whose aim is to detect, identify and normalize all the complex temporal information under the same writing format to facilitate the extraction of the relevant answers. We used for this phase our Arabic temporal resource.

P1: يستحب حرث الأرض في الليالي السود وسميت بالسود لأنها تتميز بكثرة الغيوم والسحب مما يجعل
الطقس باردا في النهار ودافئا أثناء الليل.

P1: It is desirable to plow the land in black nights and it’s called black because it is characterized by over clouds, which makes the weather cold in the day and warm during the night.

After UNIFICATION:

P1: يستحب حرث الأرض في 14 جانفي إلى 02 فيفري وسميت بالسود لأنها تتميز بكثرة الغيوم والسحب مما
يجعل الطقس باردا في النهار ودافئا أثناء الليل.

P1: It is desirable to plow the land in 14th January to 02nd February and it’s called black because it is characterized by over clouds, which makes the weather cold in the day and warm during the night.

Extraction of synonyms and antonyms verbs.
At this step, we will extract a list of synonyms and antonyms for the verbs detected in the analysis module of the question. This extraction is made from a lexical resource for the Arabic language called Arabic WordNet (AWN) based on the design and content of Princeton WordNet and linked to other WordNet for other languages.

Extraction of relevant passages.
We have a set of relevant sentences from which we will extract a set of candidate sentences. Indeed, a sentence is considered as a candidate when we can deduce the elements of answer to our question asked at the beginning. In our case:

- It comprises the detected NE of the starting question,
- It comprises both the NE or a name of signal and the same verb as that of the question or belonging to the list of synonyms of this verb.
- It comprises both the detected NE of the starting question or a name of signal and a verb belonging to the list of antonyms of the question verb.

Filtering candidate answers.
The filtering task makes it possible to obtain a set of candidate sentences containing the most maximum of answers likely to be correct. It eliminates answers that do not carry certain conditions and filters for the rest of the candidate answers based on a temporal inference mechanism.

Scheduling answers based on passages.
We calculated for each answer corresponding to the question the number of occurrences of the extracted named entity. A score S of relevance is based on the number of

occurrences of term t in the passage p with S: Score, t: Named entity and p: Passage associated with each answer.

We ranked the relevance of the passages according to the number of occurrences of terms (C t).The score varies between 1 and 15 named entities; we fixed this condition according to the results obtained during this step.

If the score S > = 1 and S < 4: the passage is considered to be weakly relevant,
S > = 4 and S <= 8: the passage is considered moderately relevant,
S > 8 and S <= 15: the passage is considered highly relevant.

Answer generation.

The generation of the answer is based on a comparison between the answer and the details extracted from infobox which generally contains the most important temporal information.

5 Evaluation

Evaluation is an essential task in the development of computer applications for the TALN. It aims to analyze the performance of our proposed method cited in the previous section. For this evaluation, we collected a corpus in Arabic language composed of a set of 500 temporal questions from the TREC corpus (Text REtrieval Conference) and from a list of questions produced in TERQAS Workshop. We got 25533articles download from Wikipedia for the questions collected.

Table 5. Experiment results

Number of questions	Number of articles	Temporal Relations
500	25533	12947

In our collection of downloaded articles, almost 0.48% of Wikipedia articles not contain an infobox, 0.30% are empty articles. Indeed, these noted temporal relations are relevant sentences whose correct answer is in one or some of them.

Subsequently, our temporal resource composed of 6 categories presented previously in step of construction of temporal resource.

Finally, the performance of an Ar-TQAS system can be measured by the metrics of evaluations introduced by [14].

Table 6. Results of System Ar-TQAS

Recall	Precision	F-measure
0.76	0.70	0.72

6 Conclusion

The representation of temporal information in Arabic language is one of the most important problems in the automatic processing of natural language (NLP). In this paper, we have presented a work that deals with temporal information involving several forms of inference in Arabic language. The notion of temporality has been dealt with in the context of a question and answer system. We proposed a new method for the resolution of temporal inference based on the construction of an Arabic temporal resource. Then, the overall objective is to solve the problems of temporal inference analysis. The Ar-TQAS system has been developed. This system makes it possible to deal with the complex temporal information and generates an answer from a corpus of texts. The evaluation Ar-TQAS system has shown encouraging results. In the future, we aim to extend the corpus of questions and the basis of the Arabic temporal information too, to obtain a specific temporal resource that must be useful in different domains.

7 References

1. A. Ben-Abacha, “ Recherche de réponses précises à des questions médicales : les systèmes de questions-réponses MEANS ”, PhD thesis. Universite PARIS-SUD 11 LIMSI-CNRS. JUIN 2012.
2. Y. Benajiba, “Arabic Named Entity Recognition”, PhD dissertation, Polytechnical University of Valencia, Spain, 2009.
3. D.B. Koen, W. Bender, “Time frames: temporal augmentation of the news”, IBM Systems Journal 39, PP.597-616, July 2000.
4. H. Li, Y. Gao, G. Shnitko, Y. Meyerzon, D. Mowatt David, “Techniques for extracting authorship dates of documents”, December 2009.
5. F. Zaraket, J. Makhlouta, “Arabic Temporal Entity Extraction using Morphological Analysis”, IJCLA VOL.3, NO. 1, PP.121-136, JAN-JUN 2012.
6. F. Mohammed, K. Nasser, H. Harb, “A knowledge based Arabic Question Answering system (AQAS)”, ACM SIGART Bulletin, PP.21-33, 1993.
7. Y. Benajiba, P. Ross, A. Lyhyaoui, “Implementation of the ArabiQA Question Answering System’s components”, Workshop on Arabic NLP, 2nd Information Communication Technologies ,3-5, April 2007.
8. S. Bekhti, A. Rehman, M. AL-Harbi and T. Saba,“AQUASYS: an Arabic question-answering system based on extensive question analysis and answer relevance scoring”, In International Journal of Academic Research, Vol. 3 Issue 4, p45, Jul2011.
9. B. Hammo, S. Ableil, S. Lytinen and M. Evens,“Experimenting with a Question Answering system for the Arabic language”, In Computers and the Humanities. Vol. 38, N°4. Pages 397 – 415, 2004.
10. O. Trigui, L.H. Belguith, P. Rosso,“Arabic definition question answering system”, In workshop on language resources and Human Language technologies for Semitic Languages, 7th LREC, Valleta, Malta, PP. 40-45, 2010.
11. W. Bakari, P. Bellot, M. Neji, “A logical representation of Arabic questions toward automatic passage extraction from the Web”, Int J Speech Technol, 2017.
12. H. Omri, Z. Neji, M. Ellouze, L. Belguith, “The role of temporal inferences in understanding Arabic text”, International Conference on Knowledge Based and IntelligentInformation and Engineering Systems, Marseille, France,2017.
13. L. H. Belguith, L. Baccour, M. Ghassan, “Segmentation de texts arabes basée sur l’analyse contextuelle des signes de ponctuations et de certaines particules”,6-10 juin 2005.
14. V. RIJSBERGEN, “Information Retrieval”, Butterworth & Co(Publishers) Ltd, London, second edition, 1979.