Temporal Pattern Extraction in Arabic language

Hajer Omri [1], Zeineb Neji [2], Mariem Ellouze [3]

Faculty of Economics and management, Tunisia, Sfax
Computer department, Miracl laboratory, University of Sfax

[1]hajer.omri2010@gmail.com
[2]zeineb.neji@gmail.com
[3]mariem.ellouze@planet.tn

Abstract. Despite the importance of temporal inference in several domains especially the question answering systems it remains still in its departure compared to other languages. This article deals with the automatic co-construction of patterns of temporal relations for the question answering systems.

We have implemented this approach in temporal inference called **TPE**: **T**emporal **P**attern **E**xtraction.

Keywords: inference; Question answering system; temporal inference; Arabic language.

1 Introduction

In previous years the main objective of researchers is to build machines that can learn, communicate, see and manipulate objects and essentially reason because it is considered one of the biggest stakes in different fields. Although reasoning or inferring has always been peculiar to the human being and will not be easy to reproduce, it constitutes a research objective and a motivation to continue imitating the functions of the human brain.

Inference is a mental operation that allows the reader to deduce the unspoken or implicit elements in a text by drawing on his knowledge of the world in his "personal encyclopedia". Making an inference means producing new information based on available information.

There are many kinds, all researchers don't agree on a single definition and classify inference according to non-mutually exclusive categories. Among the categories of inferences we distinguish temporal inference. This inference as called also temporal reasoning makes it possible to deduce temporal relations.

Example (all the examples in Arabic language are transliterated with Buckwalter[1]):

[1] http://www.qamus.org/transliteration.htm

" ولد موزارت في 27 يناير 1756 في سالزبورغ بالنمسا مؤلف موسيقي نمساوي يعتبر من أشهر العباقرة المبدعين في تاريخ الموسيقى رغم أن حياته كانت قصيرة، فقد مات عن عمر يناهز الـ 35 عاماً بعد أن نجح في إنتاج 626 عمل موسيقي
"

"Mozart was born on 27 January 1756 in Salzburg, Austria, an Austrian composer who is considered one of the most famous geniuses in the history of music, although his life was short. He died at the age of 35 after producing 626 musical works. "

Question: "**متى ولد موزارت؟**/ mtY wld mwzArt? / when was Mozart born?"

We need a smart analysis here to get the right answer. This intelligent analysis is called inference more particularly temporal inference since one is processing temporal information.

The temporal inference covers several domains and disciplines because of its importance. It is presented strongly in question answering systems, which is concerned with building systems that automatically answer questions in a natural language by extracting a precise answer from of a corpus of documents.

Any temporal information can be clearly expressed (explicit) or referred to as an unspoken (implicit) and which the interlocutor must understand by himself. A speaker may wish to pass over some temporal information and if we speak of a machine that extracts a response that is not clearly expressed, we encounter several difficulties, hence the need for a system that makes the extraction of any temporal information implicitly represented.

2 Related works

In this section, we present the previous work on temporal inference. Despite extensive research in Arabic and the volume of Arabic textual data has started growing on the Web in the last decade, it is considered as a starting point for the work of other languages such as English. Several criteria go into slower progress at Arabic research levels.

To understand that the information X is deduced from the information Y, is a simple deduction for the human being, but for the machine it is quite different. That's why the researchers proposed several approaches to solve this problem. The latter are classified into:

2.1 Rules-based methods

These methods, which are based on rules, are the oldest among the other types of extraction methods. The principle of this method is that the system designer manually establishes a set of rules for locating and extracting the desired data. These rules are extraction patterns, often implemented using automata, but the creation of these patterns is a long and costly job.

Among the researchers who have made systems based on rules are:

Reasoning [1] about time at different granularities while assuring the modeling of imprecise, gradual and intuitive relationships such as "just before" or "almost touch-

es". To deduce from the new relations it uses not only the classical operators but also its new operators of ascending granular conversion "↑" and descending "↓" which allows the conversion of one granularity to another.

Expresses temporal information [2] on different levels of granularity as well precision. It integrates it with other inferences, uses a uniform memory for declarative, episodic, and procedural knowledge. It distinguishes temporal inference by several characteristics: the use of a temporal window, temporal chaining, and interval manipulation, with projection, eternisassions and Anticipation.

2.2 Semantic methods

HUTO [3] is an ontology which provides a conceptual model in RDFS for modeling temporal expressions and annotating RDF resources. It proposes a set of the rules allowing standardizing the representation of the temporal data, but also rules of inferences and implications, expressed in the form of CONSTRUCT requests in SPARQL in order to deduce and explain the maximum temporal information so that to allow reasoning on the data.

CHRONOS [4]: is a system of reasoning on temporal information for the OWL ontologies. The latter represents both qualitative and quantitative temporal information. Based on Allen's relationships CHRONOS makes it possible to deduce the implicit relations and to detect the inconsistencies while retaining the solidity, the exhaustiveness and traceability on the whole of the supported relations.

2.3 Hybrid methods

Temporal inference has increased in recent years in several areas. Among the works are researchers who focus on the clinical field as the team of [5]. He develops a hybrid method for adapting the extraction of temporal expressions in a corpus of patient clinical records. Hybridization takes place between a symbolic approach which is a manual enrichment of the rules of the HeidelTime tool specific to the clinical field. A supervised approach to sequence prediction based on CRF (conditional random fields).

3 Particularity of Arabic language and time constraints

Arabic is a very rich language; However, this richness needs special manipulation which makes regular NLP systems, designed for other languages are unable to manage it. Arabic is a spoken language by nearly 300 million people in the world and it is the religious language for more than a billion people. It imposed itself with the Quranic revelation which conferred its status as a sacred language. Its unique character and beauty have forgotten the admiration of Muslims, beyond ethnic and geographical disparities.

Among the manifestations of the richness of this language is the fact that the names, notions and concepts benefit from a very wide palette of nuances which al-

lows to be expressed with extreme precision. Citing the example for the designation of the months of the year when one can note a significant variety of this word that's why we need a system of equivalence between the representations set which designates the same temporal information to resolve any ambiguity.

Table 1. The name of the solar months

English	Arabic
January	جانفي/ jAnfy,يناير / ynAyr, كانون الثاني / kAnwn AlvAny
February	فيفري/ fyfry,فبراير / fbrAyr
March	مارس/ mArs
April	أفريل / >fryl,أبريل / >bryl
May	ماي/ mAy,مايو / mAyw
June	جوان/ jwAn,يونيو / ywnyw
July	جويلية/ jwylyp,يوليوز / ywlywz
August	أوت / Awt,أغسطس / >gsTs,غشت / g$v
September	سبتمبر/ sbtmbr,شتنبر / $tnbr
October	أكتوبر / >ktwbr
November	نوفمبر/nwfmbr,نونبر /nwnbr
December	ديسمبر/ dysmbr,دجنبر /djnbr,كانون الأول/ kAnwn Al>wl

Table 2. The name of the lunar months

Arabic
محرم/ mHrm
صفر / Sfr
ربيع الأول/ rbyE AlAwl
ربيع الثاني/ rbyE AlvAny
جمادى الأول/ jmAdY >lAwl
جمادى الثاني/ jmAdY AlvAny
رجب/ rjb
شعبان/ $EbAn
رمضان/ rmDAn
شوال/$wAl
ذو القعدة/ *wAlqEdp
ذو الحجة / *w AlHjp

Example of ambiguity: for temporal information 03/12/2000 we find several representations:

- 03-12-2000
- 1421 هجري الثالث من دجنبر/AlvAlv mn djnbr 1421 hjry
- 1421 هجري الثالث من ذو الحجة/AlvAlv mn *w AlHjp 1421 hjry
- 2000 ميلادي اليوم الثالث من شهر ديسمبر/Alywm AlvAlv mn $hr dysmbr 2000 mylAdy

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 2000 we can

also find the year هجري 1421/1421 hjry /1421 Hijri or ميلادي 2000/2000 mylAdy /2000 gregorian.

4 Proposed approach

The proposed method presented in this section aims at automating the construction of temporal relationship patterns for question answering systems. This method is considered as a rule-based method and it's composed of three modules as shown in the previous figure (Fig1).

The first module in this method consists of the question analysis, which makes it possible to extract the various named entities as well as the verbs. In the second module, we proceeded to the construction of our corpus by automatically downloading the articles corresponding to the named entities already acquired through Wikipedia. After a set of corpus pre-processing us go on to the last module which consists in extracting the candidate sentences, which leads to a set of relevant sentences that is used to construct the patterns of temporal relations.

In the following we detail the various steps and the phases that constitute them.

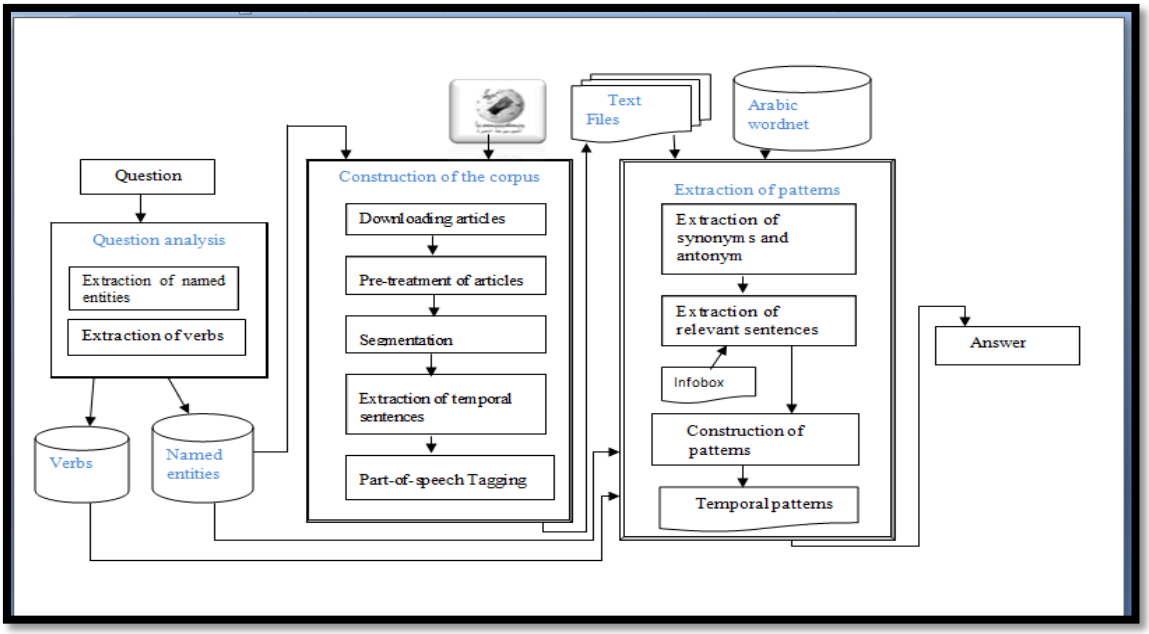

Fig. 1. Proposed approach

4.1 Question Analysis

This step consists in analyzing a question in the Arabic language that solicits temporal information only. This constraint must be highlighted at the level of our pro-

gram. Indeed, our starting point is a question bearing temporal information only; the other types of questions are not the subject of our research.

A question is called temporal if it begins with temporal signals. To find these temporal signals we have used the list of questions produced in TERQAS Workshop[2] (illustrated in Table 2) then this list has undergone an Arabic translation in order to have possible temporal signals.

This first stage contains two phases to be detailed.

Table 3. Temporal signals

Temporal signals in Arabic	Temporal signals in English
متى mtY	When
منذ متى mn* mtY	since when
كم(يقي,دام,مكث,أقام,ظل) km(bqy,dAm,mkv,>qAm,Zl)	How long{many|much}(stay, long, remain,resided,shadow)
في أي (عام وقت, زمن,حقبة,شهر,يوم,تاريخ) fy >y (EAm wqt, zmn,Hqbp,Shr,ywm,tAryx)	in which (year, month,day, date,period,time, era)
إلى أي(عام , وقت, زمن ,حقبة ,شهر , يوم , تاريخ) IIY >y(EAm , wqt, zmn ,Hqbp ,Shr , ywm ,tAryx)	Until which (year, month,day, date,period,time, era)

Extraction of named entities.

We proceed at this level to the extraction of the named entities by [9] that remains in each question for the purpose of building an EN base that we use for the construction of our corpus granting the corresponding articles.

Extraction of verbs.

Here it is a question of decomposing the question in order to extract the verbs that exist by [8]. The extraction of verbs will be useful for the following modules. More details will then be given in the following sections.

4.2 Construction of the corpus

Downloading articles.

From this phase, we put our first step for the construction of our corpus. This phase consists of hosting articles from the online encyclopedia Wikipedia. In fact, we will automatically download the corresponding articles to the extracted EN from the previous step in XML format.

[2] TERQAS was an ARDA Workshop focusing on Temporal and Event Recognition for Question Answering Systems, www.cs.brandeis.edu/_jamesp/arda/time/readings.html

Pre-treatment of articles.

The structuring of Wikipedia articles requires a pre-treatment of gender: elimination of parentheses and words that are not in the Arabic language and links and images.
In a first phase, we extract the textual content of the articles downloaded automatically to have our corpus.

During this phase, we will retrieve the Infobox, when it exists because we will use it for verification in the following steps. We retrieve the raw text from the article. The corpus becomes after this step of format TXT.

Segmentation.

In a third phase, we proceed to the segmentation [7] of the articles. The latter represents, in linguistics, a pre-processing of one or more textual documents in order to be able to subsequently process them (a morphological analysis, semantics, etc.).This operation is sensitive to each language because each has its own specificities that must be taken into account. It is considered to be important to locate segments containing the information.

The result of this stage will serve as input for the step of extracting the so-called temporal or candidate sentences.

Extraction of temporal sentences.

This is to get rid of unnecessary information and access those that are considered relevant to anticipate and act as quickly as possible in decision-making.

Once the articles, text part of the article precisely, cleaned up is segmented, we proceed to a selection to keep only those sentences that contain temporal information (relevant) [8].

Part-of-speech Tagging.

This stage consists in identifying the morphological characteristics [6] of the words of each temporal sentence of our corpus. What really interests us in this morphological analysis is to locate the verbs.

Let us return here to the first phase in which the verb of each question analyzed was detected. A comparison of the verbs of the identified phrases and the detected verb of the question will take place.

4.3 Construction of patterns

Extraction of synonyms and antonyms.

In this step we will extract the list of synonyms and antonyms for the verbs detected from our starting time questions from Arabic Wordnet (AWN).

We went through a coding phase for this extraction; in fact AWN is codified with Bluckwalter so we used a codification to have synonyms and antonyms in Arabic.

The antonyms serve us for temporal questions concerning duration.

For Example:

كم دامت الحرب العالمية الاولى
How long did the First World War last?
Km dAmt AlHrb AlEAlmyp AlAwlY

We find ourselves in front of two situations:

- We can have a direct answer from a relation of synonymy:

استمرت الحرب العالمية الاولى لمدة أربعة سنوات
The First World War continued for four years
Astmrt AlHrb AlEAlmyp AlAwlY lmdp >rbEp snwAt

[دام/ lasted /dAm =/ إستمر/ continued /Astmr].

- Or we can extract the response from an antonymic relation:

إنتهى لهيب الحرب العالمية الاولى بعد أربعة سنوات من جحيم
The flames of First World War ended after four years of hell
<nthY lhyb AlHrb AlEAlmyp AlAwlY bEd >rbEp snwAt mn jHym

[إنتهى /<nthY / ended≠دام /dAm/ lasted].

Extraction of relevant sentences.

This phase is the most difficult if we aim at a good evaluation of the patterns.
A sentence is considered relevant if:

- 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.

Certainly, we have a set of relevant sentences whose correct answer (s) exists in one or some of them. The solution envisaged for the correct answer is to make a comparison between the temporal information contained in these relevant sentences and the Infobox which generally contains the most important temporal information.

In case of equality, after solving the temporal constraints, the sentences will be considered that candidates.

The result of this module will serve as input for the last module which is the extraction of the patterns.

Construction of patterns.

The candidate sentences are considered to be responses to the temporal questions asked at the outset.

We can then associate with each question one or more regular answers (called patterns).
As an example to answer the question "**متى انتهت الحرب العالمية الأولى**/ When the First World War were ended / mtY Antht AlHrb AlEAlmyp AlAwlY ". The answer to this question can be presented differently in the text.

Example:

- الحرب العالمية الأولى (1914-1918)
- بدأت الحرب العالمية الأولى سنة 1914 و انتهت سنة 1918

The patters are:

- <اسم> <تاريخ بدأ> <تاريخ انتهاء>
- بدأت <اسم> سنة<تاريخ بدأ> و انتهت سنة <تاريخ انتهاء >

5 Evaluation

The aim of the evaluation is analyzing the detailed capabilities of our proposed method cited in the previous section. In this section we present their evaluation results.

As a first evaluation, we collected a corpus (in Arabic language) composed of set of 100 temporal and heterogeneous questions related to several domains at the beginning and the number of questions was increased each time to evaluate the results of our system TPE. Our corpus was extracted from the corpus of TREC international conference[3] (Text REtrieval Conference) for the years from 1999 to 2003 and from a list of questions produced in TERQAS Workshop.

Once the patterns are extracted, and for more precision we have asked the help of an expert in the domain to judge the semantics of the patterns.

Table 4. Experiment results

Number of question	Number of article	Temporal relation	Number of pattern	Validated pattern
100	35000	7000	1500	1150
200	63250	11500	3050	2520
350	90407	25630	4700	3100

6 Conclusion

The question of identifying temporal relations using a pattern approach is particularly on interesting entry point in several areas such as question-and-answer systems.

[3] http://trec.nist.gov/

The work that we have presented in this article is part of the work of the identification of temporal relations. In this context, we proposed a method for the identification of temporal relations based on a semantic approach based on patterns.

We began this article with an overview of temporal inferences. Next, we proposed a method for the automatic extraction of patterns for the identification of temporal relations. Then, we presented our system "TPE" which presents the result of development of the proposed method. This system allows defining the temporal patterns from a corpus of texts.

In this work, we aim at extending the temporal information base in order to build a specific time dictionary that can be useful in different domains.

Acknowledgment

We would like to record our appreciation to all people that involve in writing this article. First of all, our appreciation goes to Computer department for all the guidance especially my advisors Madam Mariem Ellouze and Zeineb Neji for guiding and assistaning us until we complete this article.

References

1. Quentin Cohen-Solal et al, “Une algèbre des relations temporelles granulaires pour le raisonnement qualitatif ”, 2015.
2. Patrick Hammer et al, “The OpenNARS implementation of the Non-Axiomatic Reasoning System”, 2015.
3. Papa Fary Diallo et al, “ HuTO: une Ontologie Temporelle Narrative pour les Applications du Web Sémantique”, 2015.
4. Eleftherios Anagnostopoulos et al, “CHRONOS: A Reasoning Engine for Qualitative Temporal Information in OWL”, 2014.
5. Mike Donald Tapi Nzali, Aurélie Névéol, Xavier Tannier, “analyse d'expressions temporelles dans les dossiers électroniques patients”, 2015.
6. Spence Green and Christopher D. Manning, “Better Arabic Parsing: Baselines, Evaluations, and Analysis”, 2010.
7. Lamia Hadrich Belguith, Leila Baccour et Ghassan Mourad, “Segmentation de textes arabes basée sur l’analyse contextuelle des signes de ponctuations et de certaines particules”, 2005.
8. Max Silberztein, Nooj: A Linguistic Annotation System for Corpus Processing, 2005.
9. Héla Fehri et al, “Reconnaissance et traduction d'entités nommées en arabe avec NooJ en utilisant un nouveau modèle de représentation”, 2011.