Enhancing HeidelTime for Time expression Annotation in Arabic News

DAHMRI Khaled
Department of computer science
USTHB, LRIA
Alger, Algérie
email: dahmrik@gmail.com

ALIANE Hassina
Department of computer science
CERIST
Alger, Algérie
email: haliane@hotmail.com

AZZOUNE Hamid
Department of computer science
USTHB, LRIA
Alger, Algérie
email: azzoune@yahoo.fr

***Abstract*— Temporal expression is considered as a key information in many tasks of modern information retrieval field. This paper presents a rule based approach for identification of temporal entities in Arabic text. Our approach is founded on HeidelTim[1] tool, where lists of regular expressions are used. Based on some of those lists and in order to cover more Temporal expressions in Arabic Language we add some regular expressions to enhance our list. For this approach, we use a linguistic preprocessing which consists of POS tagging, a list of regular expression which eases the processes of entities identification and makes it efficient, and a set of manually developed rules are employed to analyze the recognized temporal entities in a sentence. This approach is showing to have an F-measure of 0.93. Tested on news corpus.**

Keywords- Temporal expressions ; temporal annotation; Arabic language; HeidelTime; News corpus.

I. INTRODUCTION

Temporal annotation of documents plays an important role in many of natural language processing areas, such as question answering system, automatic summarizing system, event detection, and information retrieval…etc.

Temporal information is considered as crucial information that helps for the understanding, and exploration of NLP text [2]. For event detection, identifying temporal information can be useful since an event is defined as something that happens at a particular time and place. In question answering, it is useful to resolve temporal references such as the date in the sentence "Algeria got its independence in 1962" to answer the question "when did Algeria get independence from France?". In summarization, temporal annotation is used to establish a timeline for all events in multiple documents to create a coherent summary [7].

Due to the importance of temporal information, temporal annotation has been a challenging problem for NLP communities, and most of the research have been dealing with English language only. In contrast there are hardly researches done in Arabic language, therefore, that explains the lack of the availability of free resources and tools for general and temporal tagger in particular.

[1] https://code.google.com/archive/p/heideltime/

We propose in this paper a rule-based approach for identifying temporal information from a raw news corpus in Arabic text, our approach is based on time regular expression matching. The remainder of the paper is organized as follows: section 2 is a survey of related work, section 3 is a description of our approach, section 4 reports the results and evaluation of our system, and finally we end up with a conclusion and future work in section 5.

II. RELATED WOR

Much of research has been done in annotation of temporal expression, namely English [9], German [11], and Chinese [10]. This is not the case in Arabic where we hardly find such research that covers the Arabic language. Nevertheless, some industry tools already exist but the techniques underlying these tools haven't been disclosed and evaluated academically yet [13].

There are two major approaches for identifying temporal expression: rule-based approaches and machine learning approaches. [7] developed a machine learning method to identify temporal and numerical expressions in Arabic language, Support Vector Machine (SVM) were used for detecting temporal and numerical expressions in Arabic Penn Treebank (ATB) [4], they obtained F-measure of 88.5% for temporal expressions and 96% of numerical expressions.

[1] proposed a rule-based approach, they used an unsupervised segmentation algorithm and a minimalist set for rules to get a partial POS of their corpus which will be used as a basis for the recognition process that implements a set of rules using specific linguistic markers to recognize events, and expressions of time and place, they achieved F-measure of 84% for temporal expressions, and 45% for place expressions. [6] developed and analyzed two approaches (rule-based approach and CRF based classifier) for identification and classification of temporal entities in Hindi, For the rule-based approach they used a set of hand-crafted rules which are modeled as regular expressions, and for the other approach they used human tagged data to learn a classifier and then recognize temporal expression, for their system they report F-measure of 83% for the rule-based approach and 78% for CRF based classifier approach. [5] Used a set of linguistic rules, and syntactic analysis to recognize temporal and event expressions according to TimeML Standard [12]. [13] used a morphological analysis and a finite state transducer to detect Arabic temporal entities, for their method they obtained 94.6% of recall and

84.2% of precision. In [10] developed a multi-language tool (HeidelTime) using a rule-based approach, they used a rule-based and regular expression matching for Extraction and Normalization of Temporal expressions, and they obtained an F-measure of 86% for temporal expressions identification. Recently they developed their tool to cover all languages in [8].

III. DESCRIPTION OF OUR APPROACH

A. HeidelTime

HeidelTime is the first multilingual temporal tagger, developed at University of Heidelberg using a rule base and employs regular expression matching for extraction and normalization of temporal expressions according to TimeML Standard [12].

Initially, it was developed for English, but with the strict separation of source code and language dependent resources, several languages were thus integrated, namely: German, Dutch, Spanish, French, Italian, Arabic, Vietnamese, Chinese, Russian, and Croatian.

The language dependent resources are divided into three types:

1. Pattern files which contain regular expressions for frequently used terms to form temporal expressions.
2. Normalization files which contain normalization information of the patterns in the pattern file.
3. Rules, there is one rule file for each possible temporal expression type (Date, Time, Duration, and Set).

B. Arabic language in HeidelTime

HeidelTime is the first publicly available temporal tagger for Arabic language.

Arabic language is integrated to HeidelTime by developing the pattern, and Normalization files through the translation of these files from English to Arabic, taking into consideration the characteristics of Arabic language. And for the rules, they developed a few simple rules and processed them with the training documents in order to improve or modify them until they could not be improved or modified. In addition, they tried to translate more complex English rules to achieve high coverage of time expressions in Arabic language.

C. The corpus

The corpus used in the current study is a news corpus in culture category with 50 documents, written in Modern Standard Arabic language.

Arabic is a Semitic language written from right to left, spoken across Western Asia, North Africa, and the Horn of Africa, with an estimated 330 million native speakers. The Arabic alphabet contains 28 letters, these letters change their form according to their place in the word (at the beginning, middle or end).

Modern Standard Arabic language is divided from the holy Quran (known as Classical Arabic), is used in schools, universities, newspapers, television, and government.

D. Our System

The aim of the task is to detect and extract temporal expression in a given Arabic raw text. Our proposed approach is based on regular expressions matching which eases the process of handling a large number of rules and makes the recognition process efficient. For our work, we used regular expressions list of HeidelTime [10], and in order to cover more temporal expressions, we manually added some other regular expressions. A set of manually developed rules is employed to analyze the identified temporal expressions in a sentence.

A list of Potential Temporal Expression (PTE) is used, this list contains regular expressions of terms when most of the time are used with other temporal expressions, for example, numbers (numerical or letter), خلال, اول, عام …etc. these expressions cannot be used alone in a sentence because probably they have another meaning rather than temporal expression meaning.

Example: المدير العام للشركة الوطنية (Director General of the National Company).

In this case, the word 'العام' means ***General***, not a ***year***.

Linguistic preprocessing text like POS tagging for Arabic language is carried out, for our system we implemented Stanford pos tagger Arabic.

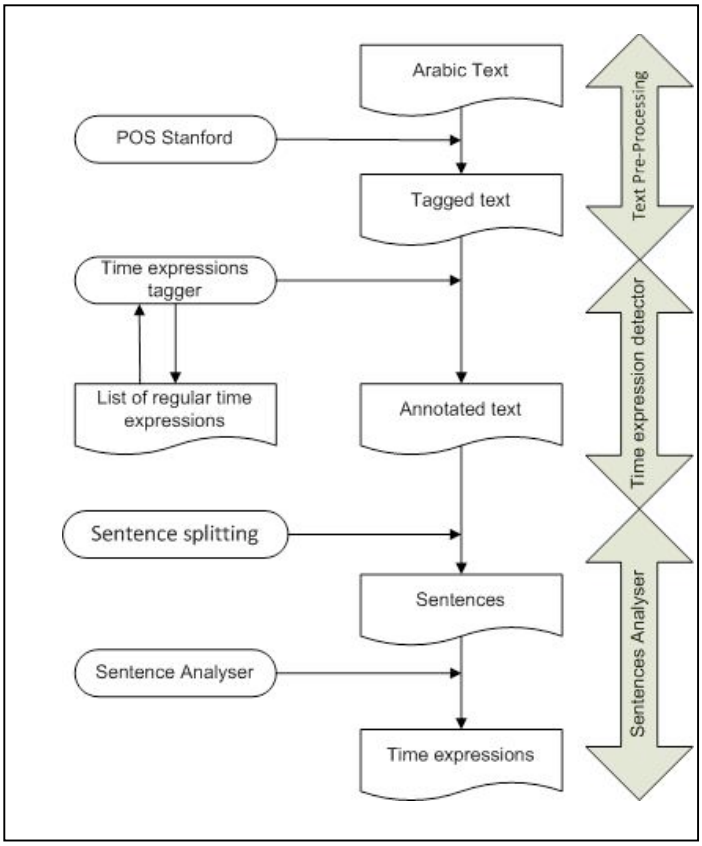

Figure 1. Architecture of the system

Our system is divided into three steps as follows:

Text Pre-Processing:

The first step consists in taking an Arabic raw text as an input and generate a tagged text using Stanford Arabic pos tagger. For this paper, the use of pos tagger is restricted to identify Preposition, subordinating conjunction, and Coordinating conjunction, which will be used in the last step.

We believe, we can extend the use of POS tagger when we cover temporal relations in further work.

Temporal expressions detector:

The second step takes the tagged text which was generated earlier in the first step as an input and identifies all the temporal expressions in the text using our temporal regular expression list.

Sentence Analyzer:

The first process of this step is sentence splitting, for this process we based on punctuation marks to identify the end of a sentence. The second process resides in sentence analysis which consists of analyzing the temporal expressions detected in a sentence, and it is based on four rules as follows:

Rule 0: if there is only one temporal expression in a sentence then we examine it to check if it is not a potential temporal expression, but if it is the case, thus we don't consider it as a temporal expression.

The last step takes the annotated text from the previous step and analyzes the temporal expressions of each sentence in the text.

Example 1: يوم العلم (Science Day)

TABLE I. EXAMPLE 1 FOR RULE0

Rule	sentence	Temporal expression	comment
/	يوم العلم	**يوم**	/
R0	يوم العلم	**يوم**	**يوم ∉ PTE**

Example 2: وحيد القرن (rhinoceros)

TABLE II. EXAMPLE 2 FOR RULE0

Rule	sentence	Temporal expression	comment
/	وحيد القرن	**القرن**	/
R0	وحيد القرن	ϕ	**القرن**∈PTE

Rule 1: if there is no word between two temporal expressions then we consider them as one temporal expression.

Example 1: 5 جويلية (5^{th} of July)

TABLE III. EXAMPLE 1 FOR RULE1

Rule	sentence	Temporal expression		comment
/	**5 جويلية**	**جويلية**	**5**	/
R1	**5 جويلية**	**5 جويلية**		/

Rule 2: if there is only one word between two temporal expressions, and this word is tagged as IN (Preposition or subordinating conjunction), or as CC (Coordinating conjunction) then we consider them as one temporal expression. Otherwise we examine the first temporal expression to check if it is not a potential temporal expression.

Example 1: الثامن من مايو (The eighth of May)

TABLE IV. EXAMPLE 1 FOR RULE2

Rule	sentence	Temporal expression		comment
/	الثامن من مايو	**مايو**	**الثامن**	/
R2	الثامن من مايو	**الثامن من مايو**		**من** is tagged as **IN**

Example 2: ذي الحجة هو شهر الحج (Dhu al-Hijjah is the month of pilgrimage)

TABLE V. EXAMPLE 2 FOR RULE2

Rule	**sentence**	**Temporal expression**		**comment**
/	ذي الحجة هو شهر الحج	**شهر**	**ذي الحجة**	/
R2	ذي الحجة هو شهر الحج	**شهر**	**ذي الحجة**	هو is tagged as **PRP**
R2	ذي الحجة هو شهر الحج	**شهر**	**ذي الحجة**	**ذي الحجة ∉ PTE**
R0	ذي الحجة هو شهر الحج	**شهر**	**ذي الحجة**	**شهر ∉ PTE**

Rule3: if there are more than one word between two temporal expressions that means they are not associated, and then we examine the first temporal expression to check if it is not a potential temporal expression.

We keep repeating the rules until they are no longer relevant.

Example 1: يوم 5 جويلية 1962 هو يوم استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة (5^{th} of July 1962, is the day of independence of Algeria from France after colonization of 132 years)

Rules	Sentence	Temporal expressions
/	يوم 5 جويلية 1962 هو يوم استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة	يوم | 5 | جويلية | 1962 | يوم | بعد | 132 | سنة
R1	**يوم 5** جويلية 1962 هو يوم استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة	**يوم 5** | جويلية | 1962 | يوم | بعد | 132 | سنة
R1	**يوم 5 جويلية** 1962 هو يوم استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة	**يوم 5 جويلية** | 1962 | يوم | بعد | 132 | سنة
R1	**يوم 5 جويلية 1962** هو يوم استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة	**يوم 5 جويلية 1962** | يوم | بعد | 132 | سنة
R2	**يوم 5 جويلية 1962** هو يوم استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة	**يوم 5 جويلية 1962** | يوم | بعد | 132 | سنة
R3	**يوم 5 جويلية 1962** هو **يوم** استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة	**يوم 5 جويلية 1962** | **يوم** | بعد | 132 | سنة
R3	**يوم 5 جويلية 1962** هو **يوم** استقلال الجزائر من فرنسا بعد استعمار دام 132 سنة	**يوم 5 جويلية 1962** | سنة | 132 | **يوم**
R1	**يوم 5 جويلية 1962** هو **يوم** استقلال الجزائر من فرنسا بعد استعمار دام **132 سنة**	**يوم 5 جويلية 1962** | **132 سنة** | **يوم**

Figure 2. Example 1 for Rule3

IV. RESULTS AND EVALUATION

We tested our system on a corpus of 50 documents, these documents are news articles taken from the web, written in Modern Standard Arabic language. In order to measure the capability of our system, we first annotated our corpus using heideltime Demo to identify the temporal expressions, then we identified manually the temporal expressions which were missing from heideltime Demo, as a result we find out that our corpus has 371 temporal expressions. The system was able to recognize 348 temporal expressions and shows an F-measure of 93%. To our knowledge, the result obtained are among the highest score achieved for temporal extraction.

The overall results are summarized in Table VI.

TABLE VI. RESULTS

/	**temporal expressions identification**
heideltime Demo	267
Manually identification	104
Total	371
Our system	348
missing	23
False positive	29
Precision	0,92
Recall	0,94
F-measure	0,93

The missing expressions are related to rules or regular expressions absence, for example: a year value like 2015, 1962 which appears alone is missing because the rule consists to tag a 4 digit when they preceded by words like خلال (During), قبل (before), في (in), or surrounded by a temporal expression.

The False positive expressions detected are related to ambiguity, or partially correct.

V. CONCLUSION

We presented in this paper a rule-based approach to temporal expression identification in news corpus of Arabic text, for this approach we used Stanford Arabic pos tagger, a list of regular expression for temporal expression recognition, and four rules to analyze the identified temporal entities in a sentence. This approach reports an F-measure of 93%, and the result is competitive with other works, whether they used the same method use or different methods. As future work, we aim to extend our temporal regular expression list and divide it into sub lists which will allow us to create, and employed more rules. Temporal relations and normalization information will be addressed in further works.

REFERENCES

[1] Aliane, H., Guendouzi, A., & Mokrani, A. (2013). Annotating Events, Time and Place Expressions in Arabic Texts. proceedings of recent advances in natural language processing, 25-31.

[2] Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121.

[3] Li, H., Str¨otgen, J., Zell, J., & Gertz, M. (2014). Chinese Temporal Tagging with HeidelTime. Conference of the European Chapter of the Association for Computational Linguistics, 133-137.

[4] Mohamed Maamouri, A. B. (2009). Creating a Methodology for LargeScale Correction of Treebank Annotation. International conference on Arabic language resources and tools. Egypt.

[5] Parent, G., Gagnon, M., & Muller, P. (2008). Annotation events and time in French language. Traitement Automatique des Langues. Avignon: Université d'Avignon.

[6] Ramrakhiyani, N., & Majumder, P. (2013). Temporal expression recognition in Hindi. Lecture notes in artificial intelligence springer , 740-750.

[7] Saleh, I., Tounsi, L., & Genabith, J. v. (2011). ZamAn and Raqm: Extracting Temporal and numerical expressions in Arabic. Asia Information Retrieval So- cieties Conference, 564-573.

[8] Str¨otgen, J., & Gertz, M. (2015). A Baseline Temporal Tagger for all Lan- guages. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 541-547.

[9] Strotgen, J., & Gertz, M. (2010). HeidelTime: High Quality Rule-based Ex- traction and Normalization of Temporal Expressions. Association for Computational Linguistics ACL, 321-324.

[10] Strötgen, J., Armiti, A., Canh, T. V., Zell, J., & Gertz, M. (2014). Time for more languages: temporal tagging of Arabic, Italian, Spanish, and Vietnamese. ACM Transactions on Asian Language Information Processing, 1.

[11] Strotgen, J., ogel, T. B., Zell, J., Armiti, A., Canh, T. V., & Gertz, M. (2014). Extending HeidelTime for Temporal Expressions Referring to Historic Dates. Language Resources and Evaluation Conference.

[12] TimeML 1.2.1. (2005, October). Récupéré sur TimeML 1.2.1: http://www.timeml.org/site/publications/timeMLdocs/timeml_1.2.1.html.

[13] ZARAKET, F., & MAKHLOUTA, J. (2012). Arabic Temporal Entity Extraction using Morphological Analysis. International Journal of. Computational Linguistics and Applications, 121-36.