Information Extraction- based on Arabic Information Retrieval using RDF Graphs: A Preliminary Study

Mohammad Khaled A. Al-Maghasbeh
School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

Mohd Pouzi bin Hamzah
School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia

ABSTRACT

This study introduces a method to facilitate Arabic information retrieval based information extraction from Arabic text. In this study propose model of Arabic information retrieval to improve information access. This proposed model attempts to enhance the performance of Arabic information retrieval from unstructured texts. This extracted information that expressed about the text will improve the retrieval of the information needs by the user and makes retrieval systems more efficient than other current systems.

Keywords

Knowledge representation, information extraction, text representation, semantic knowledge representation

1. INTRODUCTION

Although Artificial intelligence (AI) appeared over 60 years, there are many challenges of knowledge extraction from text from Arabic languages. Information extraction is an important task to discover the embed information of text by extract the semantic relation between words, and sentence. Information extraction, also concern with extract the sentiment analysis, and words disambiguation from WordNet. There are many techniques of knowledge representation such as semantic net, a Bayesian network, facts production rules, frames, conceptual dependency, neural networks, script, and hybrid representation [7]. Text representation to produce an abstract summary of input documents is a critical field in NLP. Text summarization helps to access the various information and discover the embedded knowledge that covered inside the document [3] [4]

This paper organized as follows. Section 2 briefly describes the related works in the area of information extraction and knowledge representation. Section 3 describes Arabic corpus. Section 4 provides a brief explanation of the proposed system. Section 5 shows the discussion with an example. In section 6 it summarizes the work and future work.

2. RELATED WORK

Represent the knowledge and inference is an essential branch of artificial intelligence to make the machine intelligent as a human being. The study produced by Tanwar et al. showed and discussed the importance of knowledge representation and how to build a knowledge base structure [9]. Atwell et al. in their work showed the essential challenges in Quranic texts retrieval to build a tool for improving search in the Holy Quran to understand the Quranic texts and attempt to develop the Quranic conceptual map to facilitate of extraction the Quranic knowledge [1].

The knowledge representation is a meaningful way to solve the particular problem through uses and employs information about this problem. Pike et al., the paper discussed how to represent the scientific knowledge via map the concepts, and its relations of human's cognitive into structured knowledge [6]. Robinson aimed in his study to understand of the effect of the visual representation on the visual interact simulation (VIS) in knowledge elicitation operation [8].

3. ARABIC CORPUS

There more than one websites provide the Arabic corpus with free. In our training and testing experiments, we used Arabic Newswire corpora. A group of textual Newswire will use the first corpus from CNN Arabic, BBC Arabic. Other corpora are token from Khaleej-2004, Watan-2004 Newswires.

4. PROPOSED SYSTEM

This proposed system includes several phases. It starts with the pre-processing phase. The second one is knowledge extraction which comprises; extract the NER, semantic annotations, and ontology extraction. The third one is a semantic graph which consists of knowledge base creation using OWL, and RDF. These phases will be explained as follows.

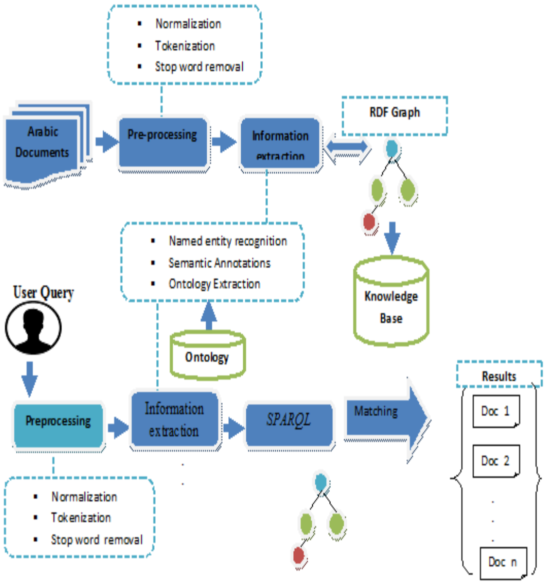

Figure 1: The proposed system architecture

4.1 System phases

- **Pre-processing phase**

The input into this phase is the text of modern standard language of Arabic Newswire. It used to reduce the noise in the texts, through remove irrelevant or not essential words such as stop words, prepositions, punctuation marks, digits from Arabic texts. After that, replace some Arabic characters into other characters to be more understandable and readable by a computer. These subtasks can be summarized as follows:

- Text normalization
- Automatic tokenization
- Stemming

• Information Extraction Phase

It is a process that responsible for information extraction from a vast database or unstructured data such as texts through understanding the different patterns in the texts. Information extraction mechanism is using to mining for data and knowledge in colossal DB or written texts of natural languages by identifying the necessary annotations, such as names entities, and relations among them [5]. In other hand, knowledge tags extraction represents a part of the information that used to describe the data sources such as image or texts document. It also helps the system to extract the knowledge and classification in the knowledge base.

In other words, KE is a way to determine specific information in Arabic textual documents to be suitable for representing the embedded knowledge in KB. This phase designed to seek the knowledge from different resources. It can obtain to capture the knowledge from unstructured data such as XML, relational DB and unstructured data such as image and texts. IE requires a correctly parsing of texts, passage, or sentences to determine the critical information such as name entities, the relation between central concepts [2]. There are subtasks of information extraction such as an ontology extraction and named entity recognition (NER) as following.

- Named entity recognition (NER)
- Semantic annotation
- Ontology extraction

Named entity recognition (NER) is one of information extraction tasks that concerned to extract the names labels to classify them into their classes such as locations, people, organization, and expression of date time and others. It applied to describes and identify the words type to become more evident. Semantic annotations extraction represents metadata that makes the text understandable. It used to extract the semantic relationships between ontologies and different concepts. SA includes two subtasks which followed to extract the best sense of texts as following

Ontology Extraction: the ontologies are extracted to capture all synonyms of each concept. The ontology includes several concepts which related to each other in class hierarchies. It concerns to determine the relevant concepts in an ontology, and semantic relations between of them. An ontology represents the formal way of knowledge representation and semantic relation between concepts. It uses to capture the knowledge in a particular domain and reduce the ambiguity of concepts to make the machine able to understand and interpret them. Ontology-based knowledge extraction (OBKE) depends on formal ontology to find the semantic relations among concepts and entities. It may be considered a guide to extract information and knowledge from unstructured texts.

• Modeling and representation phase

This section, the representation task comes after extract the unstructured knowledge from text. This knowledge can be represented and modeled into knowledge base as RDF, and OWL. After that, transformed it into structured knowledge that makes the document and text more understandable. The representation of the subtasks is showing as a figure.

• Knowledge base creation

The creation of knowledge bases through the representation of text relations between concepts and words which enables many applications of the discovery of the facts and derive new knowledge from pre-existing facts, and therefore the updated knowledge base [10]. The knowledge base includes the knowledge that extracted from the textual document, and expressed in one of the following methods:

Web Ontology Language (OWL)

OWL represents a logic method that used to express the semantic relations between things, and entities through g vocabulary and formal semantics. OWL also used to represent the complex knowledge about several things in a particular domain. OWL is a part of semantic web technology

RDF (Resource Description Framework)

RDF also is a part of semantic web technology. It used for encoding knowledge on the semantic web. RDF can be considered a formal model to make the machine more understandable of metadata. It uses the standard description to describe the web resources.

SPARQL Phase

SPARQL indicate SPARQL Protocol and RDF Query Language. SPARQL represents a query language that used to express the user need to access the RDF Graph

5. DISCUSSION

The following example (1) has taken from BBC Newswire.

Example

أعلنت وزارة الداخلية العراقية الأحد، اعتقال آخر هاربين من أحد السجون بمدينة الرمادي، ومقتل هارب ثالث خلال معركة مع قوات الأمن، فيما أسفرت عدة هجمات عن سقوط ستة قتلى على الأقل، وأكثر من 24 جريحاً. وقال مسؤول بوزارة الداخلية إن أحد السجناء الثلاثة الذين فروا من سجن "الرمادي" الجمعة، قُتل خلال تبادل لإطلاق النار مع قوات الأمن في وقت سابق السبت، أثناء محاولة إلقاء القبض عليه، دون الكشف عن مزيد من التفاصيل

- ➢ Knowledge extraction: The second step, we have to apply all K-extraction tasks, start from ontology extraction till to Named entity recognition (NER).

○ Ontology Extraction

Seq No#	Term	Synonyms
1	أعلن	اظهر, صرّح, جهر, اخبر, أذاع,أشعل, نشر
2	وزر	ملجأ, أصبح وزير, تاب, قسم في الإدارة المركزية, حال الوزير ومنصبه
3	دخل	الشك, خلل, صار داخله
4	عرق	اسم دولة عربية, عسل, مودة, رشح, شدّة, شراب مخمر, نَدِي, أكل, كِدّ
5	احد	لبس,حدّه, وحده, أول العدد, يوم من أيام الأسبوع
6	اعتقل	حبس, سجن
7	هرب	فرَّ, ترك, لاذ, اشتدَّ, تنصلَّ,غاب, سَاحَ, أدخل, صدّر, استورد, أبعد فيها
8	سجن	اعتقل, حبس, أخفى, دخل, لم يتكلم
9	مدن	جمع مدينة , حضارة, يثرب,تجمعات,بنى منطقة
10	رمد	داء, هاج, لون من الألوان, اسم مدينة
11	قتل	أمات, ذبح, أزهق روحه, فتك به, أذل
12	ثلث	جعل, طبخ, عدد من الأعداد
13	خل	جعل, طبخ, عدد من الأعداد
14	عرك	دلك, حَكّة, دار, بطش
15	قات	أعطى, عال, اطعم, قوة
16	امن	أمين, الأمان, اطمئن, ضمن
17	سفر	جمع سفرة,رحل, انكشف, وضح, أضاء, أشرق, نتج
18	عدّ	حسب, ظنَّ, راقب, عدد, حصر
19	هجم	دخل, أتى بسرعة, طرد, فتر, ذهب, غار, هدّم, اقتحم
20	سقط	وقع, رسب, أحتل, فقد,غاب, أقبل,مات
21	ست	الأنثى, اسم شجرة, نبات له زهر حسن, عدد ما بين الخمسة والسبعة
22	قلّ	قليل, نذر, فرغ, صَغُرَ, حدث, ظن, اخبر,روى
23	كثر	جمع كثرة, غلب, زاد, ارتفع
24	جرح	شقّ, شتم, كسب, اقترف, سبب له حزنا, أصاب
25	سألَ	حاسب, طرح, استخبر, استعلم,المنوط به, ذو مسؤولية, راعي الشيء
26	فر	هرب, تراجع, أسرع, كشف
27	جمع	أضاف,عزم, حشد, اجتمع, ضم, لملمَّ, ألّفَ,قوم
28	بادل	أعطى, مبادلة
29	أطلق	حلّ, أعاد,ترك, أرسل, أباح, قذف
31	نار	اللهب, رأي,أمر, جهنم, احتراق
32	سَبَتَ	حلق, أرسل, دخل, ضرب, أول أيام الأسبوع بعد الجمعة وقبل الأحد
33	حاول	أراد, بذل جهدا, أدرك, أنجز
34	القى	طرح, رمى, سلَم, توقف, امسك,قبض
35	قبض	سلّم, أخذ, امسك به
36	كشف	اظهر الشيء, رفع, فضح, أفاد, وضح,نتج, ابلغ
37	زاد	الطعام, أضاف, ضاعف,كثُرَ, نما, ارتفع
38	تفصيل	معلومة, جزء, إسهاب

○ Named entity recognition (NER)

The word “عرق” classify into Location

The word “سجن” classify into Location

The word “24” classify into Number

The word “سبت” classify into Date

The word “رمد” classify into Location

The word “امن” classify into Persons

Representation: The RDF graph will be as following:

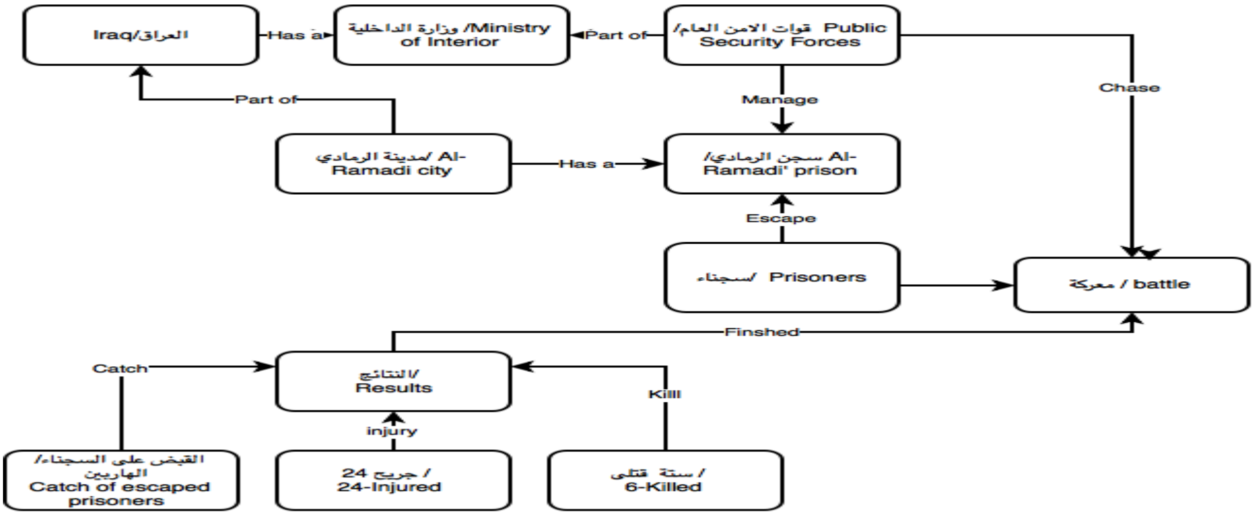
Iraq/العراق
Has a
وزارة الداخلية /Ministry of Interior
Part of
قوات الامن العام/ Public Security Forces
Chase
Part of
Manage
مدينة الرمادي/ Al-Ramadi city
Has a
سجن الرمادي/ Al-Ramadi' prison
Escape
سجناء/ Prisoners
معركة / battle
Finshed
Catch
النتائج/ Results
Killl
injury
القبض على السجناء/ الهاربين Catch of escaped prisoners
24 جريح / 24-Injured
ستة قتلى / 6-Killed

Figure 2: The RDF graph

Outputs will be as following

InfoNo#	Event	Reason	Place	Organization	Results
1	اعتقال هاربين	الهروب من السجن	مدينة الرمادي- العراق	وزارة الداخلية- العراق	مقتل الهاربين
2	هجوم ومعركة مع قوات الأمن	هروب السجناء	مدينة الرمادي- العراق	وزارة الداخلية- العراق	سقوط 6 قتلى &24 جريح
3	ملاحقة الهاربين	الهروب من السجن وقتل المدنيين	مدينة الرمادي- العراق	قوات الأمن	القبض عليهم

Convert the output into RDF

```
<rdf:RDF

<rdf:Description rdf:about="http:// www.bbc.com/arabic /rdf/Arabic IR based KR">
<rdf:type rdf:resource="http:// www.bbc.com/arabic "/>
xmlns:rdf="http://www.w3.org/2017/02/22-rdf-syntax-ns#"

xmlns:feature="http://www.linkeddatatools.com/clothing-features#">

< rdf:text> Detention of prisoners </ rdf:text>

 < rdf:name> Al-Ramadi' City-Iraq </ rdf:name>
 < rdf:text> Battle between prisoners with security forces-Al-Ramadi </ rdf:text>

 < rdf:text> Catch of escaped prisoners </ rdf:text>
<feature:size>12</feature:size>
```

Query: Assume that have the following query:

“من هي مدينة السجناء الهاربين في العراق”

The query will pass through several stages until the phase matching with the relative answer as following

Knowledge extraction: The second step, we have to apply all K-extraction tasks, start from ontology extraction till to Named entity recognition (NER).

○ Ontology Extraction

Seq No#	Term	Synonyms
1	مدن	جمع مدينة , حضارة, يثرب,تجمعات,بنى المنطقة
2	سجن	حبس, دخل حياة الرهبنة, لم يتكلم,شقّق,حفر, طين
3	هرب	فرَّ, ترك, لاذ, اشتدَّ, تنصلَّ,غاب, سَاحَ, أدخل, صدّر, استورد, أبعد فيها
4	عرق	اسم دولة عربية, عسل, مودة, رشح, شدّة, شراب مخمر, نَدِي, أكل, كِدّ

○ Named entity recognition (NER)

The word “مدن” classify into Location
The word “سجن” classify into Location
Convert the query in SPARQL query, and then match it with the relevant result as following:

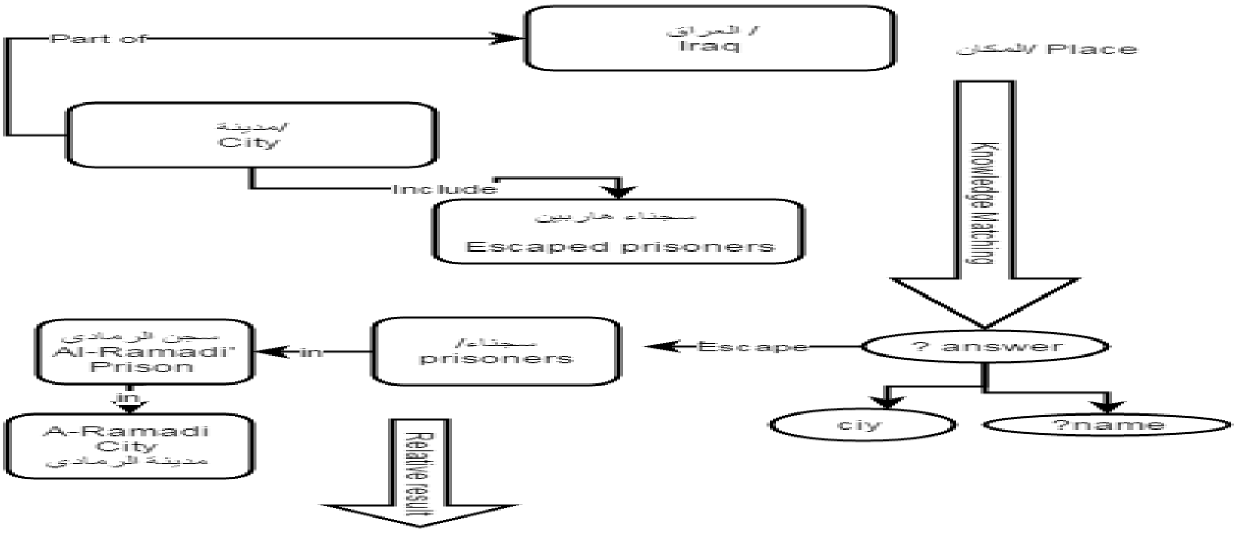

Figure 3: SPARQL Graph

InfoNo#	Event	Reason	Place	Organization	Results
2	هجوم ومعركة مع قوات الأمن	هروب السجناء	مدينة الرمادي- العراق	وزارة الداخلية- العراق	سقوط 6 قتلى &24 جريح

6. CONCLUSION

The Arabic text differs from that other texts the written in different languages. This coherent Arabic text covered embedded knowledge and information. So the Arabic information retrieval system needs intelligent characteristics such as intelligent reasoning, to be able of extraction of semantic relations, and others that help the user to retrieve the needs documents easily. This study introduced a method to develop the Arabic information retrieval via using information extraction with an OWL, SPARQL, and RDF Graphs in both documents, and query to improve the performance the system and make it able to access the text knowledge

7. ACKNOWLEDGMENT

We would like to thank University Malaysia Terengganu (UMT) for providing us with an excellent scientific environment to produce this simple work

8. REFERENCES

[1] Atwell, E., Brierley, C., Dukes, K., Sawalha, M., & Sharaf, A.-B. (2011). An Artificial Intelligence approach to Arabic and Islamic content on the internet. Paper presented at the Proceedings of NITS 3rd National Information Technology Symposium.

[2] Fan, J., Kalyanpur, A., Gondek, D. C., & Ferrucci, D. A. (2012). Automatic knowledge extraction from documents. IBM Journal of Research and Development, 56(3.4), 5: 1-5: 10.

[3] Ismail, S. S., Aref, M., & Moawad, I. F. (2013). Rich semantic graph: A new semantic text representation approach for Arabic language. Paper presented at the 17th WSEAS European Computing Conference (ECC'13).

[4] Moawad, I. F., & Aref, M. (2012). Semantic graph reduction approach for abstractive Text Summarization. Paper presented at the Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on.

[5] Mooney, R. J., & Bunescu, R. (2005). Mining knowledge from text using information extraction. ACM SIGKDD explorations newsletter, 7(1), 3-10.

[6] Pike, W., & Gahegan, M. (2007). Beyond ontologies: Toward situated representations of scientific knowledge. International Journal of Human-Computer Studies, 65(7), 674-688.

[7] Prasad, T. (2012). Hybrid systems for knowledge representation in artificial intelligence. arXiv preprint arXiv:1211.2736.

[8] Robinson, S., Lee, E. P., & Edwards, J. S. (2012). Simulation-based knowledge elicitation: Effect of visual representation and model parameters. Expert Systems with Applications, 39(9), 8479-8489.

[9] Tanwar, P., Prasad, T., & Datta, D. K. (2010). An Effective Knowledgebase system Architecture and issues in representation techniques. International Journal of Advancements in Technology http://ijict. org/ISSN, 0976-4860.

[10] Thabtah, F. (2008). VSMs with K-Nearest Neighbour to categorise Arabic text data.