Extracting Named Entities Using Named Entity Recognizer and Generating Topics Using Latent Dirichlet Allocation Algorithm for Arabic News Articles

Tarek Kanan[a], Souleiman Ayoub[a] , Eyad Saif[b] , Ghassan Kanaan[b] ,
Prashant Chandrasekar[a], Edward A. Fox[a]

[a] *Virginia Polytechnic Institute and State University (Virginia Tech), McBryde Hall Room 114, Department of Computer Science (M/C 0106), Blacksburg, VA 24061, USA*
[b]*Amman Arab University, Amman, Jordan*
[a]*Email: {tarekk, siayoub, peecee, fox}@vt.edu*
[b]*Email: e.hodiani@hotmail.com, ghassan.kanaan@yahoo.com*

Abstract

This paper explains for the Arabic language, how to extract named entities and topics from news articles. Due to the lack of high quality tools for Named Entity Recognition (NER) and topic identification for Arabic, we have built an Arabic NER (RenA) and an Arabic topic extraction tool using the popular LDA algorithm (ALDA). NER involves extracting information and identifying types, such as name, organization, and location. LDA works by applying statistical methods to vector representations of collections of documents. Though there are effective tools for NER and LDA for English, these are not directly applicable to Arabic. Accordingly, we developed new methods and tools (i.e., RenA and ALDA). To allow assessment of these, and comparison with other methods and tools, we built a baseline corpus to be used in NER evaluation, with help from volunteer graduate students who understand Arabic. RenA produces good results, with accurate Name, Organization, and Location extraction from news articles collected from online resources. We compared the RenA results with a popular Arabic NER, and achieved an enhancement. We also carried out an experiment to evaluate ALDA, again involving volunteer graduate students who understand Arabic. ALDA showed very good results in terms of topics extraction form Arabic news articles, achieving high accuracy, based on an experimental evaluation with participants using a Likert scale.

Keywords: Arabic Language; Named Entity Recognizer; Topic Extraction; Latent Dirichlet Allocation, Natural Language Processing

Preprint of paper to appear in Proceedings of the International Computer Sciences and Informatics Conference (ICSIC 2016)

1. Introduction

1.1. Arabic: Language, Encoding and Morphology

1.1.1. Arabic Language

Arabic is a widely used global language that has major differences from most popular languages, e.g., English and Chinese. The Arabic language has many grammatical forms, varieties of word synonyms, and different word meanings that vary depending on factors like word order. In spite of such complexities, limited work has been devoted to natural language processing involving Arabic, especially in comparison to the English language, which has been addressed by numerous studies. Most of the software packages, tools, and APIs for information retrieval and natural language processing do not address Arabic language requirements. To allow these software packages and tools to handle Arabic language data, substantial modification and extra work would be required for tailoring to Arabic.

Unlike most languages, Arabic is written from right to left, with no capitalization, and with 28 alphabetical characters as well as diacritics. According to Nizar Habash et al. [1], there are multiple forms of the Arabic language such as:

- Classical Arabic – This form is used in reading / reciting the holy books.
- Modern Standard Arabic (MSA) – Standard Arabic, which is commonly used in writing, speech, interviewing, broadcasting, etc. It should be noted that throughout this report, implementation is based on MSA.
- Spoken – oral dialects that vary significantly from region to region.

Arabic also employs Vowel Marks (*Tashkeel* or "*Harakat*" [1] (known as diacritics)), such as those shown in Table 1 for one of the letters.

Table 1: Diacritics for the Letter "Alef"

ا	آ	أ	إ

Diacritics for letters such as "Alef" are used to signify or distinguish sounds that are not fully specified by the Arabic letters. These characters can be used interchangeably, and change the meaning of the word. Since they are mostly used in the context of verbal exchanges or recitation, they hold very little value in the analysis carried out on texts in connection with computational linguistics.

1.1.2. Arabic Encoding

One of the first challenges faced while working with texts is the ability to recognize the characters programmatically or via a computer program. Encoding tends to be problematic; the most common and effective way to solve this difficulty is to use Unicode (UTF8 for example). Alternatives include Windows CP-1256 or X-MacArabic

1.1.3. Arabic Morphology

The Arabic language has a complex morphology due to its derivational and inflectional nature [2]. Arabic verbs and nouns are derived from a root word, and usually consist of the root word followed by a pattern to form a lemma [2]. Consider the words in Table 2.

Table 2: Derivation forms of the word “Read” in Arabic

Read	قرأ
Reading	قراءة
Reader	قارئ
I read	قرأت
Peruse	قرأ بتمعن
Legible	مقروء

Both the word “read,” and the other forms shown in Table 2, have the same base root. This is very common in the language. In Figure 1, observe how both words share the same root word, but semantically, have two different meanings. The first word is the root word, with the second word derived from it, but the meaning is changed by inflection, due to the suffix indicating singularity or plurality. In other cases, it may indicate gender or both, i.e., singularity/plurality and gender.

نيباتک بتک

Figure 1: Wrote two books

1.2. Named Entity Recognizer – NER

Consider the English quote,

“Go back, Sam. I’m going to Mordor alone.”

— Frodo, The Lord of the Rings: The Fellowship of the Ring

In theory, an NER should be able to extract "Sam" as a name and "Mordor" as a location (and arguably an organization). This is very useful as it extracts useful keywords in context.

Named Entity Recognition of Arab(ic) names of persons, organizations, and locations requires modification of available tools, e.g., the Stanford Named Entity Recognizer (SNER), or creation of new tools to extract names of entities from text, e.g., from news articles. Extracting the named entities for any text may help point out key elements. We believe these three main entities (persons, organizations, and locations) reflect the most important entities in the text and serve as the main features for future work in Arabic news article text summarization. Toward extracting the appropriate Arabic named entities, we have modified one of the available Arabic NER tools, i.e., the one created by Yasine Benajiba that is called ANER (Arabic Name Entity Recognition) [2].

1.3. Latent Dirichlet Allocation – LDA

As more information becomes available, it becomes more important to access what we are interested in. It requires advanced implementations to help us organize, search, and understand these large amounts of information. Topic Modeling, for example LDA, can assist with automatic organization, understanding, searching, and summarization of massive amounts of electronic data. A collection of documents may cover a variety of topics and sometimes each document addresses a mixture of those topics. In order to determine or "select" topics through a computational algorithm, topic modeling will be required, for example using the Latent Dirichlet allocation (LDA) algorithm.

Suppose we have a document that consists of various subjects. This document contains words that may refer or correspond to a specific subject. Basically the LDA algorithm attempts to map each word to its corresponding topic(s). LDA is a form of probabilistic model that will collect a set of words that establish a statistical relation based on the document collection. This generative model uses Bayesian inference, and involves collapsed Gibbs sampling to collect topics from a collection of documents [3, 4].

Figure 2 below appears in Blei's work [3] and provides further explanation of the LDA algorithm.

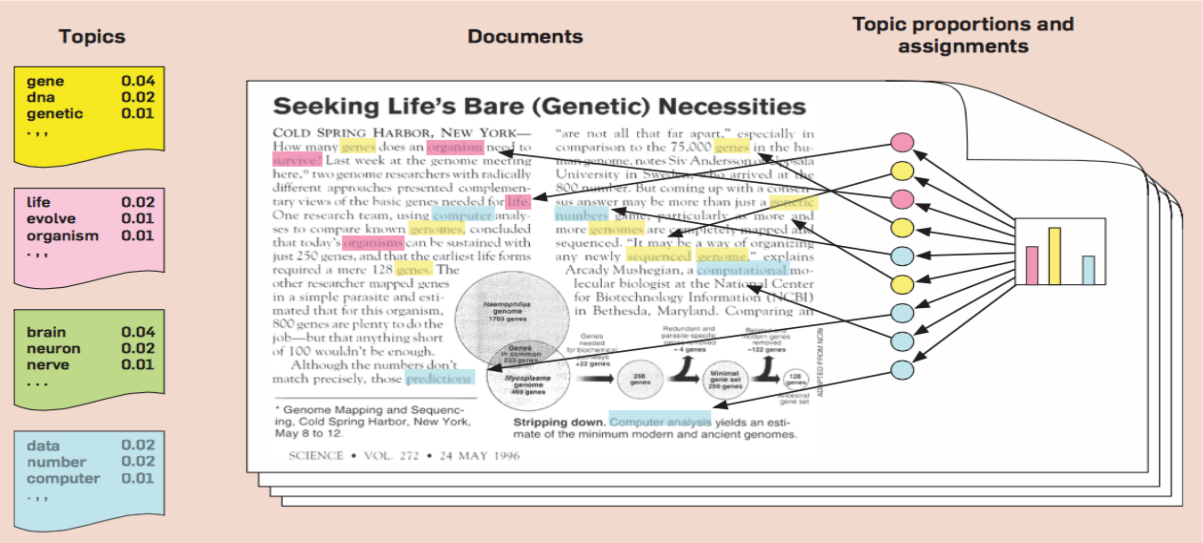


Figure 2: LDA Algorithm Demonstration (adapted from [3])

There is no implementation of the LDA algorithm that is publicly available and supports the Arabic language for topic modeling. So, to support the Arabic language, we have extended an open source LDA implementation (written in C Sharp) for the Chinese language that is publicly available on GitHub [5].

2. Literature Review

2.1. Named Entity Recognizer – NER

The Stanford Named Entity Recognizer (SNER) is a Java implementation of a Named Entity Recognizer as defined by Manning et al. [6]. A Named Entity Recognizer (NER) labels sequences of words in a text, namely proper nouns, such as person and company names, or gene and protein names. It comes with feature extractors for Named Entity Recognition, and with many options for defining additional feature extractors. Included with the download are good named entity recognizers for English, particularly for the 3 classes (PERSON, ORGANIZATION, and LOCATION) [6]. In his dissertation "Arabic Name Entity Recognition", Benajiba describes a system he has developed to extract Arabic name entities within an open domain Arabic text. In order to create his ANER system, he examines the different aspects of the Arabic language related to NER tasks and the state-of-the-art of NERs [2]. [7] describes a new technique to extract names from Arabic text by Abuleil et al. They build graphs to describe relationships between words. The proposed technique extracts some names, but misses others; they believe if they re-run the technique on more articles, the system would extract the missing names. Kanaan et al. use an existing tagger to identify proper names and other crucial lexical items and build lexical entries [8]. Shaalan develops a Named Entity Recognition system for Arabic (NERA) using a rule-based approach [9]. He uses a whitelist to represent a dictionary of names, and he includes a grammar, in the form of regular expressions. NERA has been evaluated using special tagged corpora, yielding satisfactory results in terms of precision, recall, and F1-measure. In his work, Kareem Darwish tries to enhance Arabic named entity extraction by using cross-lingual resources (Arabic/English) for Wikipedia links [10]. He shows a positive effect on recall using his method compared with Benajiba [2].

2.2. Latent Dirichlet Allocation – LDA

Allan et al. [11] define temporal summaries of news stories extracting as few sentences as possible from each event within a news topic, where the stories are presented one at a time. They define an evaluation strategy and describe simple language models for capturing novelty and usefulness in summarization. They show that their simple approaches work well. Topic discovery based on text mining techniques is discussed by Pons-Porrata et al. [12] the authors present a topic discovery system that aims to reveal the implicit knowledge present in news streams. This knowledge is expressed as a hierarchy of topic/subtopics, where each topic contains the set of documents related to it. A summary is then extracted from these documents. The summaries they build are useful to browse, with topics of interest selected from the generated hierarchies [12]. An approach to building topic models based on a formal generative model of documents, Latent Dirichlet Allocation (LDA), is frequently discussed in the machine learning literature, but its practicality and effectiveness in information retrieval are mostly unknown [13]. Liu et al. [14] consider two aspects in their paper. These aspects are the design and development of a time-based, visual text summary that effectively conveys complex text summarization results produced by the Latent Dirichlet Allocation (LDA) model, and the description of a set of rich interaction tools that allow users to work with a created visual text summary to further interpret the summarization results in context and test the text collection from multiple angles. They have applied their work to a number of text corpora and their evaluation shows promise, especially in support of complex text analysis. Brahmi et al. observe that the topic model judges each document (considered as a bag of words) as a combination of topics defined by a probability distribution over words [15].

The LDA model has been introduced within a general Bayesian framework where the authors have developed a vibrational method and expectation–maximization (EM) algorithm for learning the model from the aggregation of discrete data [16]. Since the original Prolog version of the LDA model, several contributions have been proposed. However, few studies on finding latent topics in Arabic text have been identified. For integration with works related to Arabic topic detecting and tracking [17, 18], a segmentation method that utilizes the Probabilistic Latent Semantic Analysis [19] has been applied to an AFP_ARB corpus for monolingual Arabic document topic analysis [20]. In Larkey et al. [18], the researchers compare different topic tracking methods. They claim that the utilization of separate language for building concrete topic models is preferred. Good topic models are obtained when native Arabic stories are available. However, Arabic topic tracking has not been satisfactory in texts translated from English stories. In fact, studies on Arabic IR are insufficient and the few works carried out for topic modeling lack strong evaluation. Considering the high inflectional morphology in Arabic, it seems more opportune to learn an LDA model in a mono-language context, taking more care with linguistic aspects.

3. Methodology

3.1. Building a Baseline Dataset

Since we could not find a judged Arabic news article corpus to be used as a baseline corpus to test and evaluate the NER results and to compare the implementation of our NER with other existing NER systems, we decided to build a new baseline judged Arabic news article corpus.

We began with roughly 5,200 PDF archived documents from Al-Raya, a Qatari news site covering various topics such as sports, politics, etc. Since each document contains multiple news articles, we analyzed the files to separate the articles. Files that would require OCR or where encoding was problematic were discarded. Remaining documents were processed to extract individual news articles. However, since some encoding issues were not caught by the first filter, another layer of filtering was added to discard illegible articles. The result was roughly 120,000 articles. We randomly selected one thousand articles as a test sample toward building our judged baseline corpus, as shown in Figure 3.

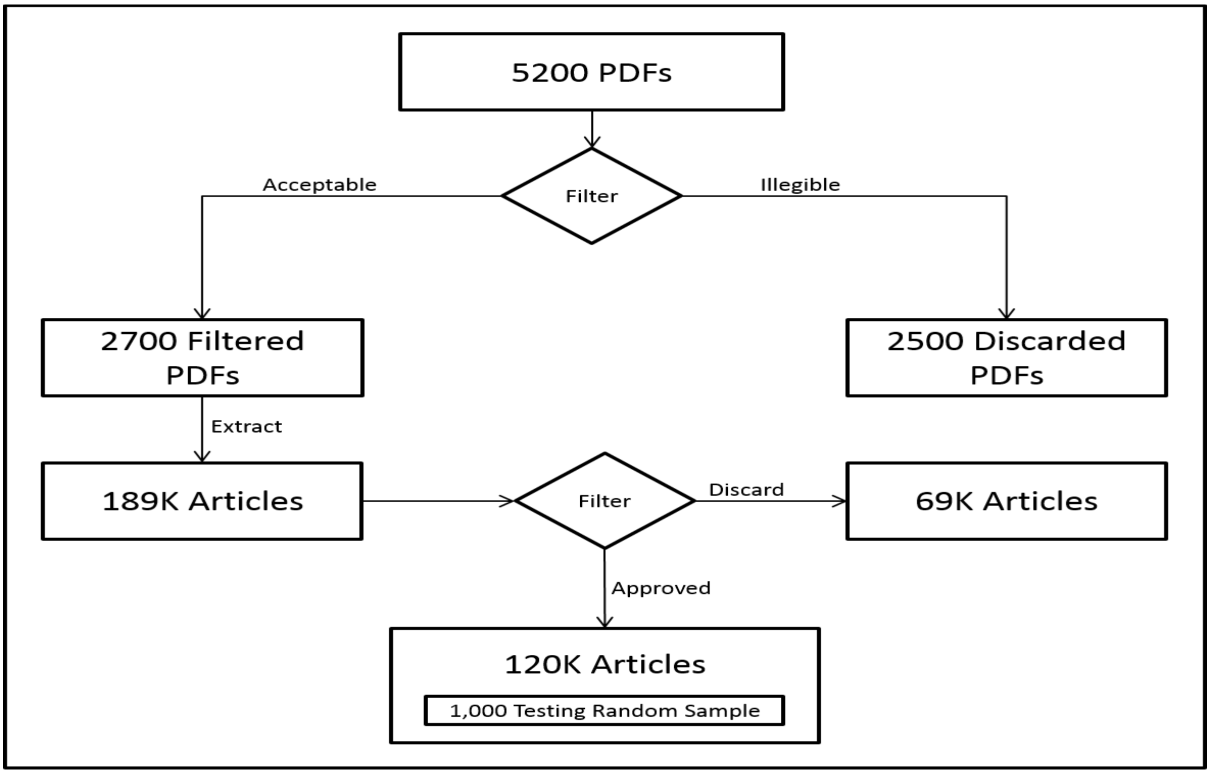
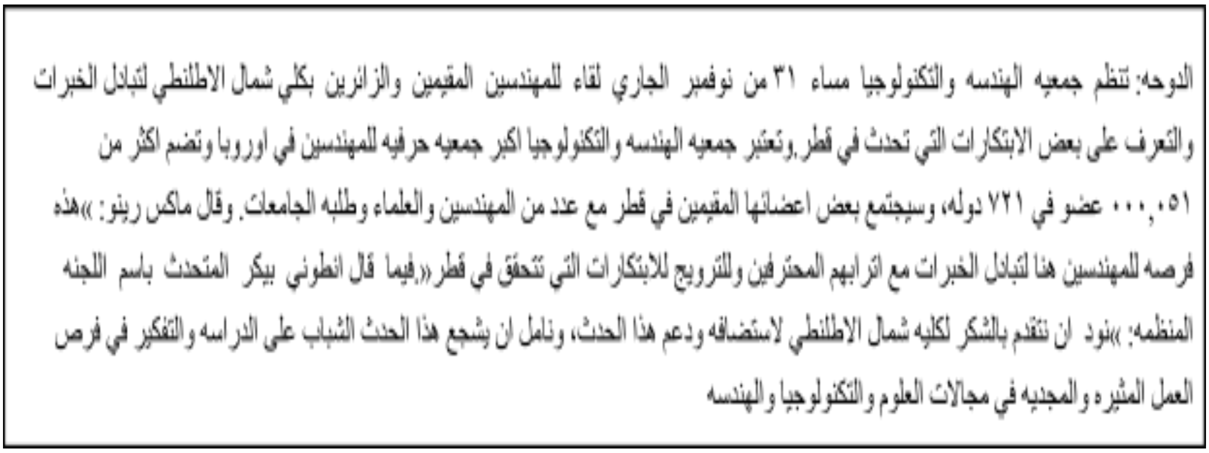


Figure 3: Flow of Dataset Extraction

We recruited graduate students who understood Arabic, because user experiences, behaviors, and task performances could differ depending on the participants' academic level and familiarity with Arabic reading and writing. For this study, we recruited ten participants, each participant was assigned a one hundred articles from the one thousand random sample we collected in our dataset extraction. For each article, they were tasked to read the article and then label the named entities. Later, they were to evaluate the articles’ topics.

3.1.1. Dataset Examples

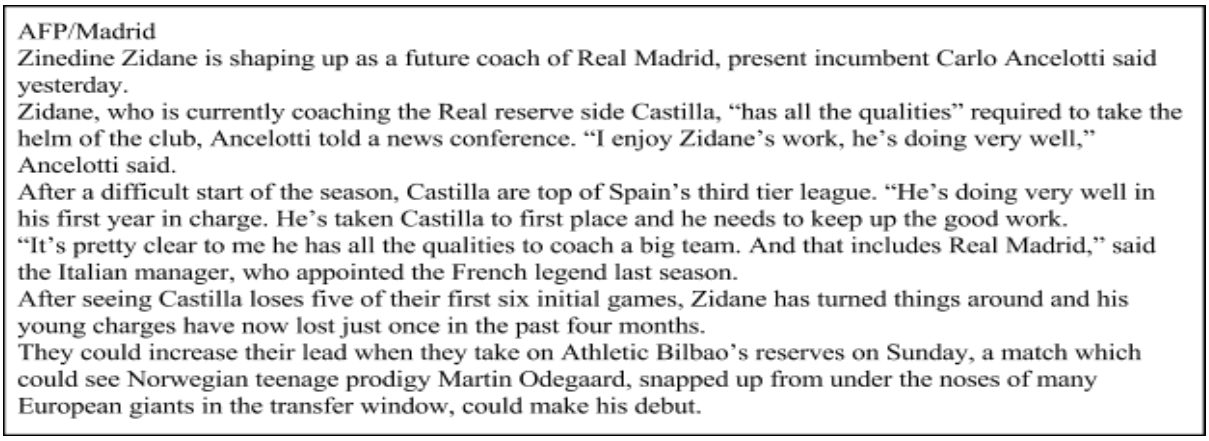
Figure 4 below introduces an example news article from our dataset. This article has been chosen from one of the many documents we collected.



الدوحه: تنظم جمعيه الهندسه والتكنولوجيا مساء ٣١ من نوفمبر الجاري لقاء للمهندسين المقيمين والزائرين بكلي شمال الاطلنطي لتبادل الخبرات
والتعرف على بعض الابتكارات التي تحدث في قطر,وتعتبر جمعيه الهندسه والتكنولوجيا اكبر جمعيه حرفيه للمهندسين في اوروبا وتضم اكثر من
٠٠٠,٠٥١ عضو في ٧٢١ دوله، وسيجتمع بعض اعضائها المقيمين في قطر مع عدد من المهندسين والعلماء وطلبه الجامعات. وقال ماكس رينو: »هذه
فرصه للمهندسين هنا لتبادل الخبرات مع اترابهم المحترفين وللترويج للابتكارات التي تتحقق في قطر«,فيما قال انطوني بيكر المتحدث باسم اللجنه
المنظمه: »نود ان نتقدم بالشكر لكليه شمال الاطلنطي لاستضافه ودعم هذا الحدث، ونامل ان يشجع هذا الحدث الشباب على الدراسه والتفكير في فرص
العمل المثيره والمجديه في مجالات العلوم والتكنولوجيا والهندسه

Figure 4: An Arabic News Article from our dataset

For better understanding, we include an English news article example to explain how our process works. See Figure 5.



AFP/Madrid
Zinedine Zidane is shaping up as a future coach of Real Madrid, present incumbent Carlo Ancelotti said yesterday.
Zidane, who is currently coaching the Real reserve side Castilla, "has all the qualities" required to take the helm of the club, Ancelotti told a news conference. "I enjoy Zidane's work, he's doing very well," Ancelotti said.
After a difficult start of the season, Castilla are top of Spain's third tier league. "He's doing very well in his first year in charge. He's taken Castilla to first place and he needs to keep up the good work.
"It's pretty clear to me he has all the qualities to coach a big team. And that includes Real Madrid," said the Italian manager, who appointed the French legend last season.
After seeing Castilla loses five of their first six initial games, Zidane has turned things around and his young charges have now lost just once in the past four months.
They could increase their lead when they take on Athletic Bilbao's reserves on Sunday, a match which could see Norwegian teenage prodigy Martin Odegaard, snapped up from under the noses of many European giants in the transfer window, could make his debut.

Figure 5: Example of English News Article

3.1.2. Stopword Removal

Similar to the English language, Arabic includes words that can be treated as stopwords. It is necessary to filter out those stopwords. However, due to language requirements, sometimes a stopword can have multiple meanings, such as the Arabic word "ال". In English, this translates to "the". Obviously, this keyword is considered as a stopword in English, which gives us a reasonable reason to discard it. However, this is not the case for Arabic, as the word can be used to represent a family name. It is important to consider these special stopwords when processing and extracting key entities. There are multiple freely available Arabic stopword lists. For this research we have merged two lists of Arabic stopwords to produce a richer collection of stopwords to meet the requirements of our work. One of the lists we use is adopted from UniNE [21] and is also used by Lucene Apache [22]. The other list is from a Google project called "Stop-Words", where stopwords are provided for 28 different languages, including Arabic [23].

3.1.3. Stemming

The main goal of a stemmer is to map different forms of the same word to a common representation called the “stem”. Stemming can significantly improve the performance of topic extraction systems by reducing the dimensionality of word vectors. The goal of an Arabic light stemmer is to find the representative form of an Arabic word by removing prefixes and suffixes, while maintaining infixes. Thus, the meaning of the word remains intact, which results in improved topic identification effectiveness.

For our experiments we used the stemmer of the Al-Khalil Morphological System, an open source Arabic analyzer, to stem our corpus [24].

3.1.4. Normalizing the Text

As discussed in Section 1.1.1, we need to consider every possible form of each word, since we rely on a knowledge base. For example, consider the following in Table 3.

Table 3: Examples of Harakat

Word with Vowel	**Root Word**	**English**
كَتَبَ	كتب	Write
كُتُب	كتب	Books
قَرَأ	قرأ	Read
قارِئ	قرأ	Reader
رَكَضَ	ركض	Run
رَكّاض	ركض	Runner

In order to produce more precise results, the collection of content that needs to be processed will have to be normalized. However, in the example provided in Table 3 above, when such words as “write” and “books” are normalized to the same root form, the meaning changes.

3.2. Arabic Named Entity Recognizer - RenA

3.2.1. Building the NER

There are 3 ways to build an NER, as shown below:

- Knowledge Base – The use of a collection of words used to identify entities based on a predefined dataset; such a collection contains a set of words mapped to a specific entity.
- Machine Learning – Uses statistical models to classify and identify grammars to deterministically identify

entities, with Conditional Random Fields (CRF) [2] being the plausible choice.

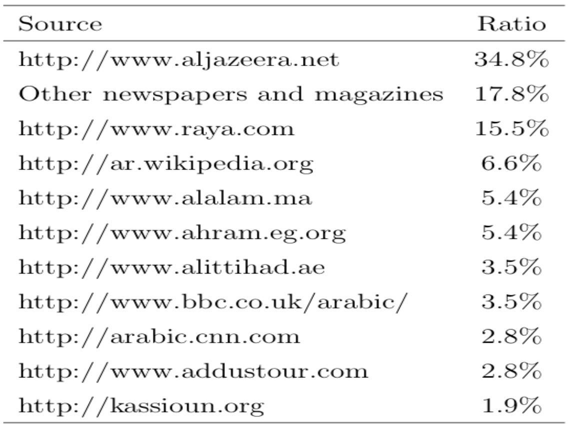
- Training – manually or automatically train a classifier to deterministically identify the entities.

Initially, this research relies on a knowledge base, but it is later improved by training.

3.2.2. Knowledge Base

We are using ANERCorp [25], which is Benajiba's [2] freely distributed corpus for Arabic, Table 4, which consists of roughly 150,000 tokens that are tagged.

Table 4: Ratio of sources used to build the ANERCorp [2, 25]



Source	Ratio
http://www.aljazeera.net	34.8%
Other newspapers and magazines	17.8%
http://www.raya.com	15.5%
http://ar.wikipedia.org	6.6%
http://www.alalam.ma	5.4%
http://www.ahram.eg.org	5.4%
http://www.alittihad.ae	3.5%
http://www.bbc.co.uk/arabic/	3.5%
http://arabic.cnn.com	2.8%
http://www.addustour.com	2.8%
http://kassioun.org	1.9%

In addition to the ANERCorp, we have also built a feature set, specifically for the following named entities: PERSON (PERS), ORGANIZATION (ORG), and LOCATION (LOC).

3.2.3. Approach

3.2.3.1. Stage 1 – Building RenA

For the initial implementation of the NER, the knowledge base is used to classify the entities of the provided texts. There is a need to build a dictionary to map the words in the knowledge base to their entity type (PERS, ORG, and LOC). Once the collection of words has been mapped to the appropriate entities, we can use the populated dictionary and a chunker (tokenizer) to classify a collection of text and determine the words' entities. The chunker will tokenize based on whitespaces. For each word, the chunker will identify possible results of the tags; for example, the word, "Washington" can be added to the dictionary with the tag (PERS) to indicate a person; we can also add "Washington" as a (LOC) to indicate a location. Once applied, every occurrence of the word "Washington" will return two different tags, [PERS, LOC].

The results produced are reasonable. Some of the keywords are tagged, while others are not. However, some of the words are tagged incorrectly. The most obvious problems in this stage are related to the stopwords and Harakat. These problems are addressed in stage 2.

3.2.3.2. Stage 2 – Improving RenA

In this stage, the main focus is to properly filter out stopwords and normalize words. Removing stopwords is considered a simple approach as a list of stopwords can be used to filter out words of little significance. Normalization helps in reducing the words' dimensionality and preventing duplicate results when chunking.

At this point, results are improved by filtering out stopwords and by normalization to reduce the varieties of the words. However, this still raises the issues of some words being improperly tagged and some words missing tags. Indeed, results for organization named entities lead to inaccurate results due to the complexity of organization naming. Often, it introduces ambiguities with person and location entities. Some Arabic NERs report low accuracy result for organizations [2, 10]. Surprisingly, this issue appears to be a continuing problem for Arabic NER. However, the persons and locations produce reasonable results, with only minor defects that can be addressed easily.

In order to improve the precision of each entity based on the knowledge base, we must train the corpus. For this collection, the use of an inclusion and an exclusion lists yield an exceptional increase in precision. In our experiment, we have selected a random sampling of news articles that contain a substantial number of keywords to help improve our training by enriching it with names of persons, organizations, and locations. Table 5 shows the list of resources and their regions, which reflect sources used to enrich our knowledge base.

Table 5: News articles used to train the NER collection via inclusion/exclusion

Sites	**Region**
http://www.aljazeera.net	Saudi Arabia
http://alwatan.com	Qatar
http://www.ahram.org.eg	Egypt
http://www.alarabalyawm.net	Jordan
http://www.al-akhbar.com	Lebanon
http://www.alanwar.com/	Lebanon
http://www.albayan.ae	UAE
http://www.alquds.co.uk	London (Universal)

With the use of inclusion and exclusion lists, our knowledge base starts to improve by the addition/removal of words to their appropriate entity type classes. For each training (enhancing) phase, we are relying on the current state of our NER and enriching the knowledge base by adding new tagged or removing wrong tagged entities for each phase. One of the interesting note to take into consideration is that once the NER is trained on an article, it usually does better on the next set of articles.

Figure 6 below summarizes the steps used to build our NER (RenA). The inner part of RenA deals with building the knowledge base (dictionary) by importing the ANERCorp knowledge base, after normalizing it, then adding our inclusion/exclusion list, all together will build our NER dictionary. We used the normalizer and stopword removal to preprocess our article then the chunker will act as a tokenizer and mapper. The chunker tokenize the preprocessed article based on the articles whitespaces and for each token it will attempt to check if an entity exists in the dictionary based on the token. Finally, the entities extraction function will produce the results from the mapped tokens.

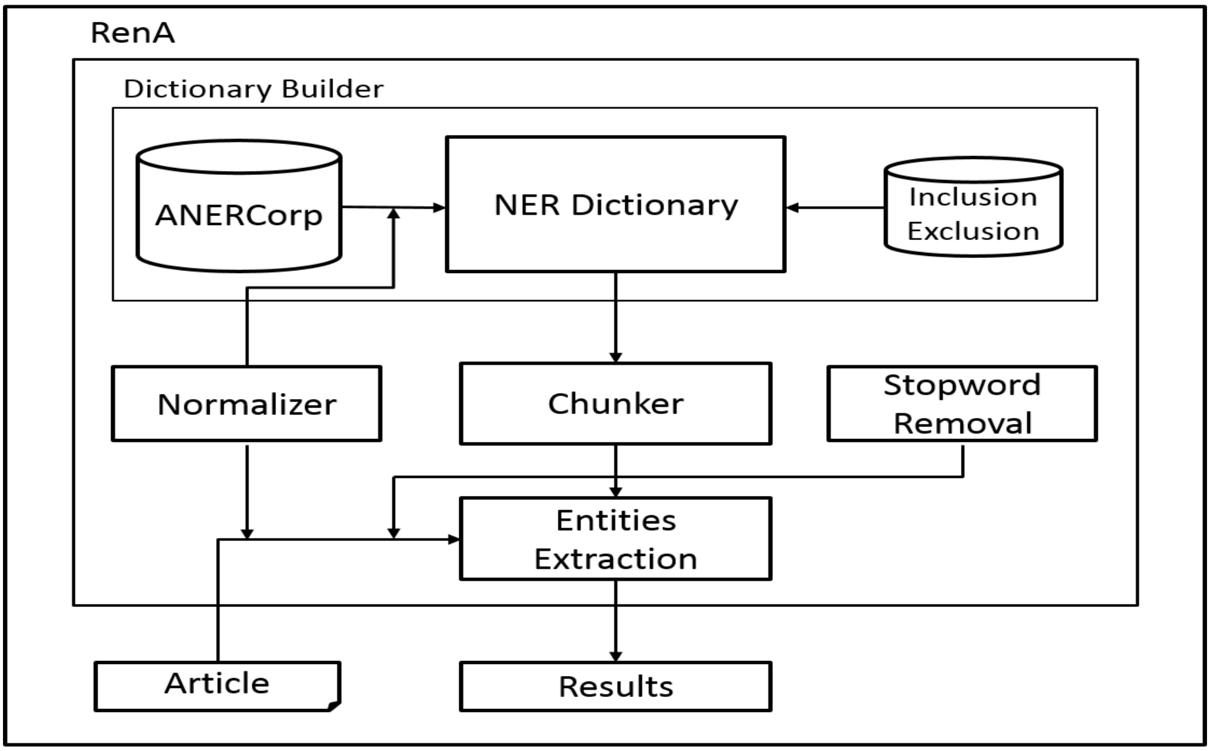
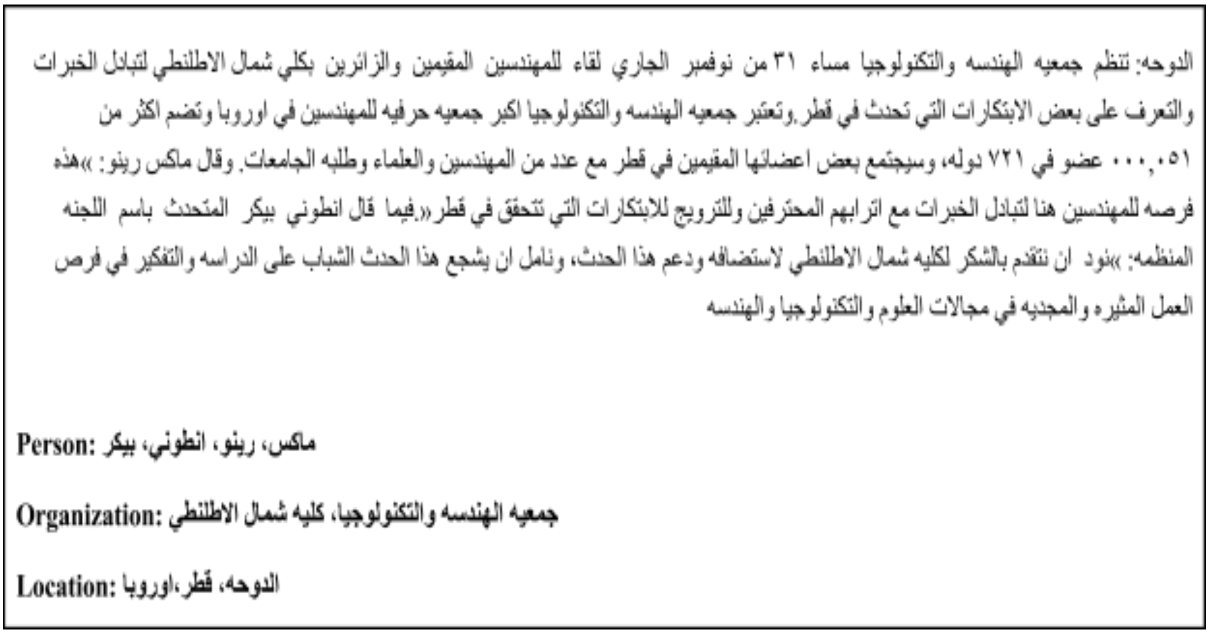


Figure 6: RenA Architecture

3.2.4. Result

Using the same articles from dataset examples in Section 2.3.1.1, we show results after extracting the named entities in the article. The results of the named entity extraction are shown in bold in the figures below. See Figure 7 for the Arabic example.



الدوحه: تنظم جمعيه الهندسه والتكنولوجيا مساء ٣١ من نوفمبر الجاري لقاء للمهندسين المقيمين والزائرين بكلي شمال الاطلنطي لتبادل الخبرات والتعرف على بعض الابتكارات التي تحدث في قطر.وتعتبر جمعيه الهندسه والتكنولوجيا اكبر جمعيه حرفيه للمهندسين في اوروبا وتضم اكثر من ٠٠٠,٠٥١ عضو في ٧٢١ دوله، وسيجتمع بعض اعضائها المقيمين في قطر مع عدد من المهندسين والعلماء وطلبه الجامعات. وقال ماكس رينو: »هذه فرصه للمهندسين هنا لتبادل الخبرات مع اترابهم المحترفين وللترويج للابتكارات التي تتحقق في قطر«.فيما قال انطوني بيكر المتحدث باسم اللجنه المنظمه: »نود ان نتقدم بالشكر لكليه شمال الاطلنطي لاستضافه ودعم هذا الحدث، ونامل ان يشجع هذا الحدث الشباب على الدراسه والتفكير في فرص العمل المثيره والمجديه في مجالات العلوم والتكنولوجيا والهندسه

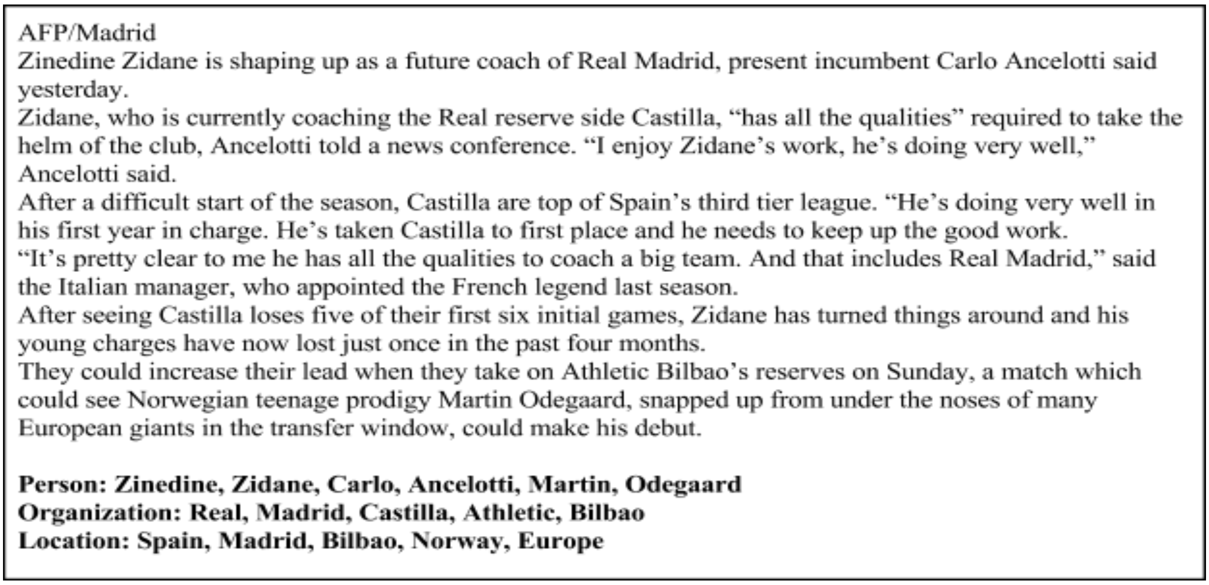
Person: ماكس، رينو، انطوني، بيكر

Organization: جمعيه الهندسه والتكنولوجيا، كليه شمال الاطلنطي

Location: الدوحه، قطر،اوروبا

Figure 7: Arabic News Article and Extracted Named Entities

For more illustration, an English example is also included to explain how our process works. Figure 8 shows an example of an English news article and its named entities.



AFP/Madrid
Zinedine Zidane is shaping up as a future coach of Real Madrid, present incumbent Carlo Ancelotti said yesterday.
Zidane, who is currently coaching the Real reserve side Castilla, "has all the qualities" required to take the helm of the club, Ancelotti told a news conference. "I enjoy Zidane's work, he's doing very well," Ancelotti said.
After a difficult start of the season, Castilla are top of Spain's third tier league. "He's doing very well in his first year in charge. He's taken Castilla to first place and he needs to keep up the good work.
"It's pretty clear to me he has all the qualities to coach a big team. And that includes Real Madrid," said the Italian manager, who appointed the French legend last season.
After seeing Castilla loses five of their first six initial games, Zidane has turned things around and his young charges have now lost just once in the past four months.
They could increase their lead when they take on Athletic Bilbao's reserves on Sunday, a match which could see Norwegian teenage prodigy Martin Odegaard, snapped up from under the noses of many European giants in the transfer window, could make his debut.

Person: Zinedine, Zidane, Carlo, Ancelotti, Martin, Odegaard
Organization: Real, Madrid, Castilla, Athletic, Bilbao
Location: Spain, Madrid, Bilbao, Norway, Europe

Figure 8: English News Article and Extracted Named Entities

Figures 7 and 8 show articles that an NER can be used with to extract named entities (shown in bold). These extracted entities can be used to identify the underlying context of the article and show some of its key elements. For example, in Figure 8, a football fan will be able to recognize that this article is about the Real Madrid soccer team (Organization named entity) in regards to a player called Zidane (Person named entity).

3.2.5. Evaluation

Table 6 shows us the results between RenA and the basic NER from LingPipe [26]. These results compare the differences between recall, precision, and F1 measure.

Table 6: Recall, Precision, and F1 Values for RenA and LingPipe NERs

	RenA NER			**LingPipe Toolkit NER**		
	Recall	**Precision**	**F1**	**Recall**	**Precision**	**F1**
PERSON	0.826	0.497	0.539	0.582	0.371	0.374
ORGANIZATION	0.813	0.421	0.446	0.39	0.377	0.329
LOCATION	0.77	0.558	0.564	0.55	0.338	0.356
Average	0.803	0.492	0.516	0.507	0.362	0.353

In Figure 9 below, the precision values of both NERs are displayed for each entity. It shows that RenA produces higher precision results for each entity.

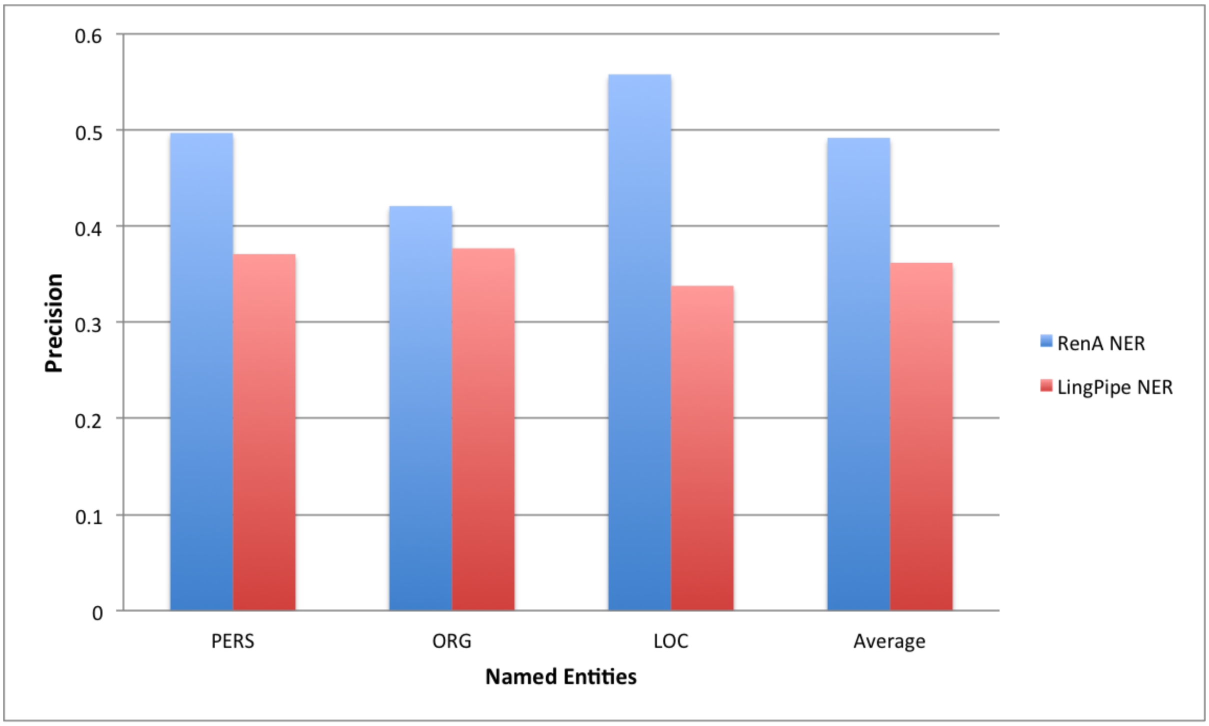


Figure 9: Precision Values for RenA and LingPipe NER

In Figure 10 below, the recall of both NERs is displayed for each entity. RenA showed better results than its counterpart NER in terms of recall.

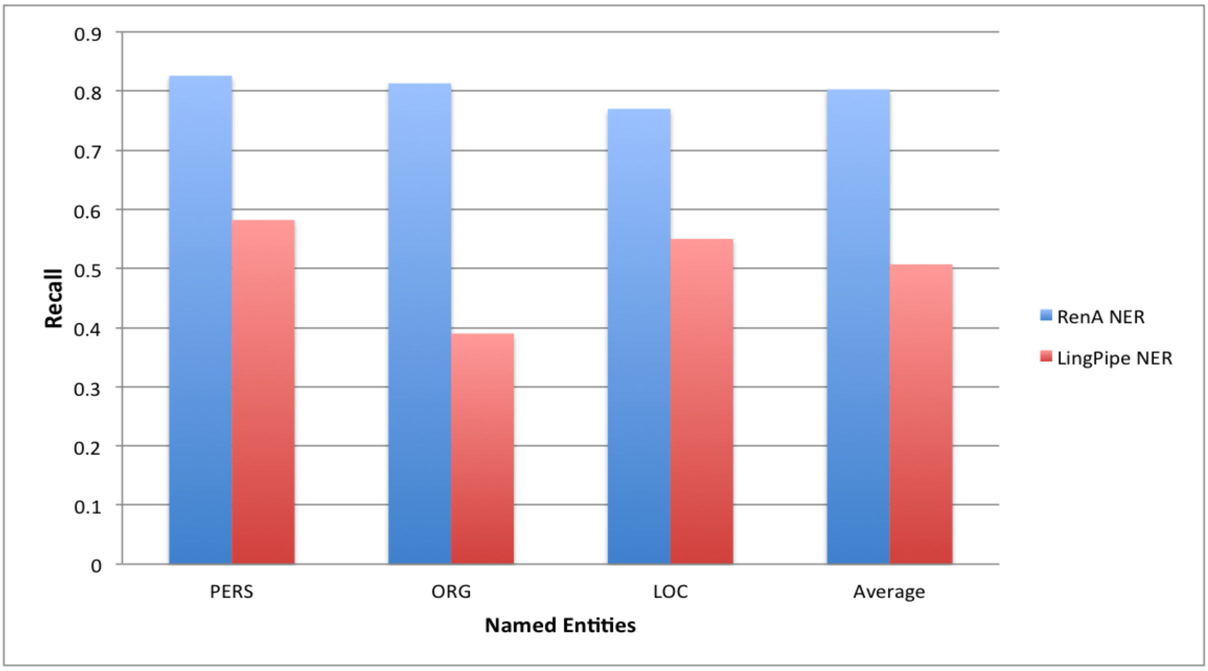


Figure 10: Recall Values for RenA and LingPipe NER

In Figure 11 below, the F1 measure of both NERs is displayed for each entity (the greater the value, the better it is). F1 measure shows that RenA is retrieving better results. Table 6 shows that RenA is on average 15% more effective in retrieving results.

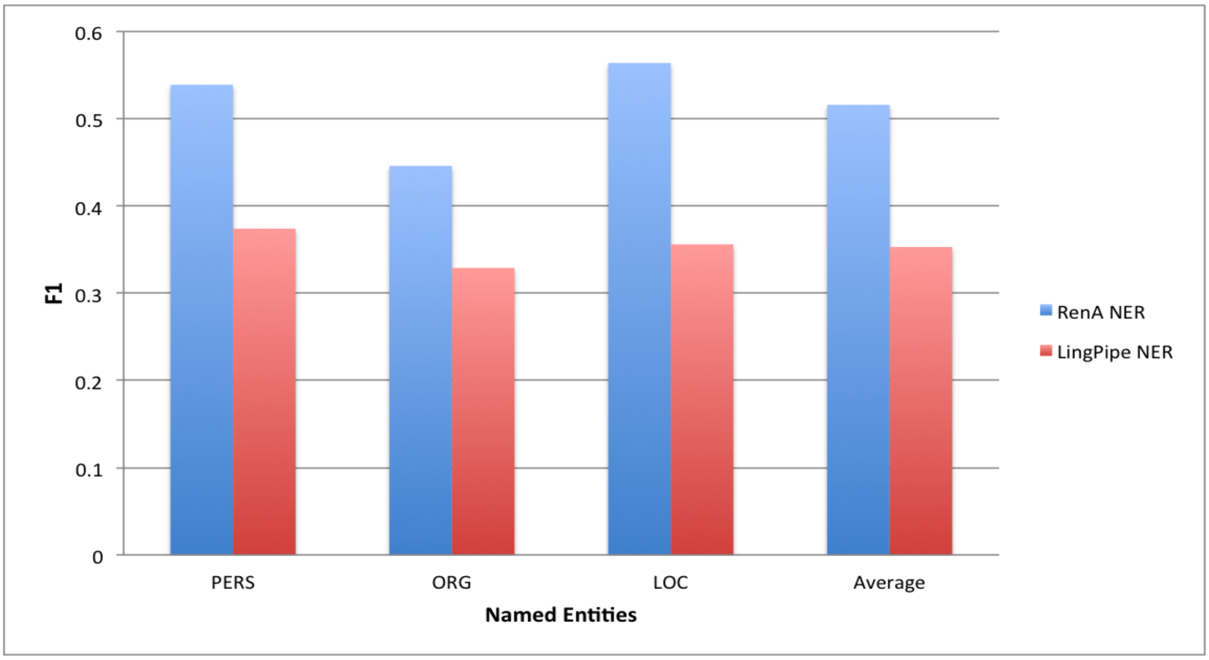


Figure 11: F1 Values for RenA and LingPipe NER

3.3. Arabic Latent Dirichlet Allocation – ALDA

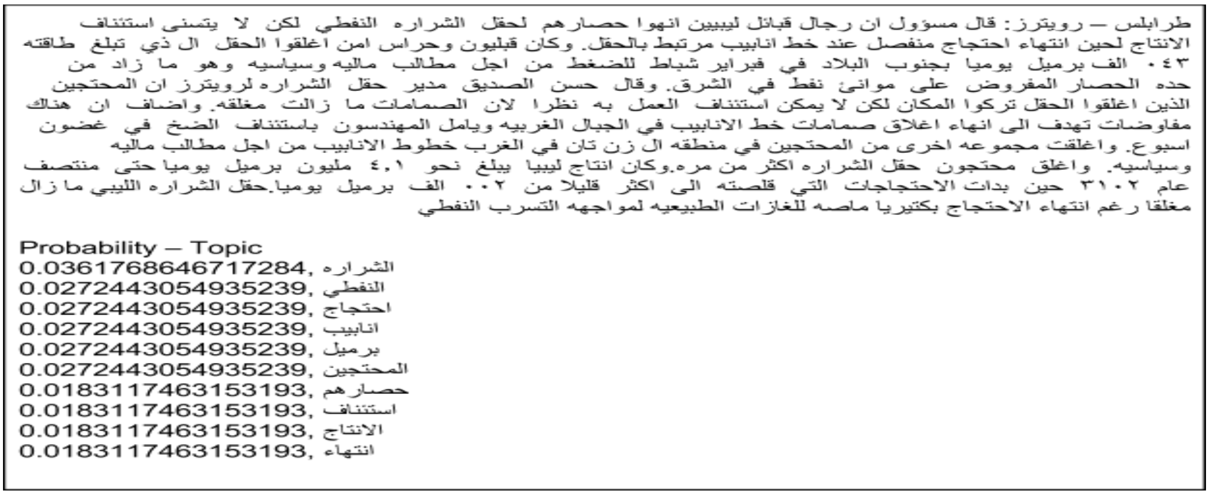
3.3.1. Algorithm

In order for the LDA to understand the different topics – since a corpus may consist of multiple topics and various words, which involves various distributions – Bayesian inference must be applied. The inference techniques that have been used in the LDA algorithm involve collapsed Gibbs sampling in order to marginalize the distribution and probabilistically determine the topics of a corpus. Gibbs sampling allows the LDA algorithm to observe sequences of an approximate joint distribution to understand high probabilistic topics (via histogram) and compute our latent variables (hyper-parameters). By using a Dirichlet Multinomial distribution (also referred to as categorical distribution) in the sampling process, the algorithm can marginalize and determine the topical model. It should be noted that since the LDA is a generative model using Gibbs sampling, iteration is required to generalize the steady-state (or Markov chain) model since the sampling algorithm is random in the initial state, whereas each iteration will further improve the topic choices. One of the characteristics of the LDA's implementation is the modularity, such that preprocessing the data can further enhance the topic selection, for example, stemming and stopword removal.

The open source code we have used consists of two main parts: 1) Core, which involves the LDA generative model and statistics algorithms that have been modified to support the Arabic language, stemming, stopword removal, and normalization; and 2) Viewer, for which we have modified and implemented an interface in order to include extra interactive variables, such as number of topic, words per topic, and number of inference iterations using the LDA model in the Core library. The result of each model will be displayed as a table and also persisted in a CSV file for further evaluation. The following sections show screenshots and examples produced by our Arabic Latent Dirichet Allocation (ALDA).

3.3.2. Result

Figure 12 shows a news article example and the results after applying the ALDA algorithm. It shows the main topic covered by this article as a list of words with their corresponding probability.

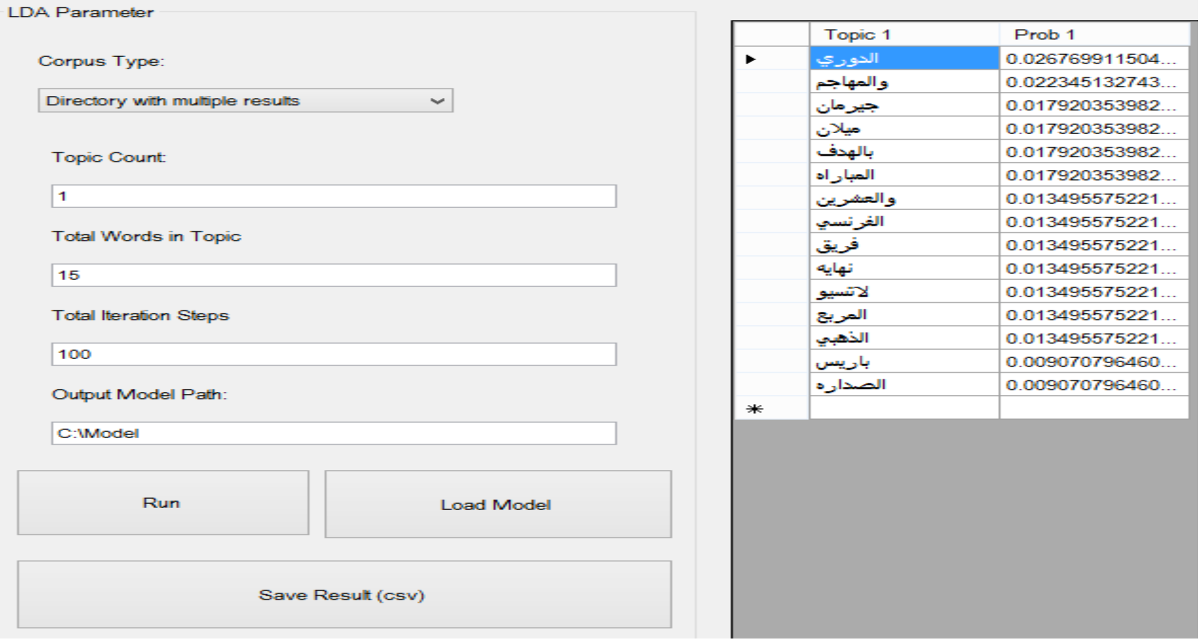


طرابلس – رويترز: قال مسؤول ان رجال قبائل ليبيين انهوا حصارهم لحقل الشراره النفطي لكن لا يتسنى استئناف
الانتاج لحين انتهاء احتجاج منفصل عند خط انابيب مرتبط بالحقل. وكان قبليون وحراس امن اغلقوا الحقل ال ذي تبلغ طاقته
٠٤٣ الف برميل يوميا بجنوب البلاد في فبراير شباط للضغط من اجل مطالب ماليه وسياسيه وهو ما زاد من
حده الحصار المفروض على موانئ نفط في الشرق. وقال حسن الصديق مدير حقل الشراره لرويترز ان المحتجين
الذين اغلقوا الحقل تركوا المكان لكن لا يمكن استئناف العمل به نظرا لان الصمامات ما زالت مغلقه. واضاف ان هناك
مفاوضات تهدف الى انهاء اغلاق صمامات خط الانابيب في الجبال الغربيه ويامل المهندسون باستئناف الضخ في غضون
اسبوع. واغلقت مجموعه اخرى من المحتجين في منطقه ال زن تان في الغرب خطوط الانابيب من اجل مطالب ماليه
وسياسيه. واغلق محتجون حقل الشراره اكثر من مره.وكان انتاج ليبيا يبلغ نحو ٤,١ مليون برميل يوميا حتى منتصف
عام ٣١٠٢ حين بدات الاحتجاجات التي قلصته الى اكثر قليلا من ٠٠٢ الف برميل يوميا.حقل الشراره الليبي ما زال
مغلقا رغم انتهاء الاحتجاج بكتيريا ماصه للغازات الطبيعيه لمواجهه التسرب النفطي

```
Probability – Topic
0.0361768646717284, الشراره
0.0272443054935239, النفطي
0.0272443054935239, احتجاج
0.0272443054935239, انابيب
0.0272443054935239, برميل
0.0272443054935239, المحتجين
0.0183117463153193, حصارهم
0.0183117463153193, استئناف
0.0183117463153193, الانتاج
0.0183117463153193, انتهاء
```

Figure 12: News Article with its Corresponding Topic Using ALDA

Figures 13 and 14 show two screen shots for ALDA and the related results. Different parameters are used for each one of the two shots.



LDA Parameter

Corpus Type: Directory with multiple results

Topic Count: 1

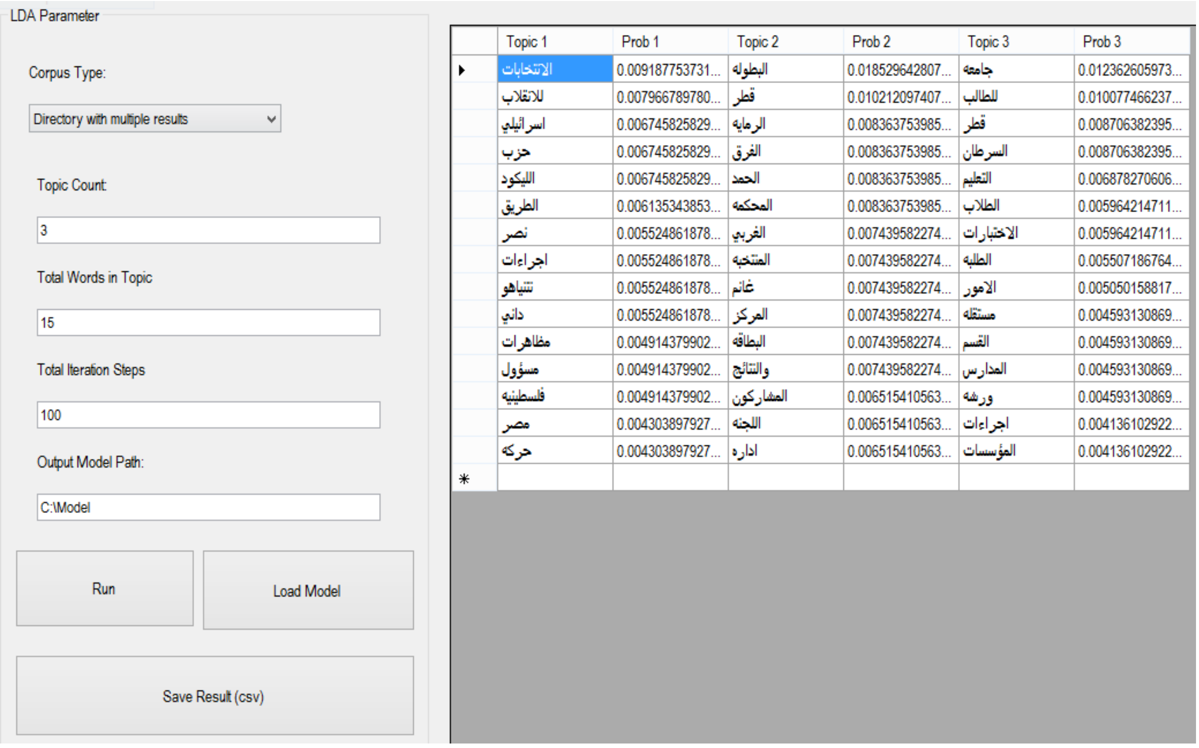
Total Words in Topic: 15

Total Iteration Steps: 100

Output Model Path: C:\Model

	Topic 1	Prob 1
▶	الدوري	0.026769911504...
	والمهاجم	0.022345132743...
	جيرمان	0.017920353982...
	ميلان	0.017920353982...
	بالهدف	0.017920353982...
	المباراه	0.017920353982...
	والعشرين	0.013495575221...
	الفرنسي	0.013495575221...
	فريق	0.013495575221...
	نهايه	0.013495575221...
	لاتسيو	0.013495575221...
	المربع	0.013495575221...
	الذهبي	0.013495575221...
	باريس	0.009070796460...
	الصداره	0.009070796460...
*		

Figure 13: ALDA Screen Shot Showing One Topic



LDA Parameter

Corpus Type: Directory with multiple results

Topic Count: 3

Total Words in Topic: 15

Total Iteration Steps: 100

Output Model Path: C:\Model

	Topic 1	Prob 1	Topic 2	Prob 2	Topic 3	Prob 3
▶	الانتخابات	0.009187753731...	البطوله	0.018529642807...	جامعه	0.012362605973...
	للانقلاب	0.007966789780...	قطر	0.010212097407...	للطالب	0.010077466237...
	اسرائيلي	0.006745825829...	الرمايه	0.008363753985...	قطر	0.008706382395...
	حزب	0.006745825829...	الغرق	0.008363753985...	السرطان	0.008706382395...
	الليكود	0.006745825829...	الحمد	0.008363753985...	التعليم	0.006878270606...
	الطريق	0.006135343853...	المحكمه	0.008363753985...	الطلاب	0.005964214711...
	نصر	0.005524861878...	الغربي	0.007439582274...	الاختبارات	0.005964214711...
	اجراءات	0.005524861878...	المنتخبه	0.007439582274...	الطلبه	0.005507186764...
	نتنياهو	0.005524861878...	غانم	0.007439582274...	الامور	0.005050158817...
	داني	0.005524861878...	المركز	0.007439582274...	مستقله	0.004593130869...
	مظاهرات	0.004914379902...	البطاقه	0.007439582274...	القسم	0.004593130869...
	مسؤول	0.004914379902...	والنتائج	0.007439582274...	المدارس	0.004593130869...
	فلسطينيه	0.004914379902...	المشاركون	0.006515410563...	ورشه	0.004593130869...
	مصر	0.004303897927...	اللجنه	0.006515410563...	اجراءات	0.004136102922...
	حركه	0.004303897927...	اداره	0.006515410563...	المؤسسات	0.004136102922...
*						

Figure 14: ALDA Screen Shot Showing Multiple Topics

3.3.3. Evaluation

For evaluating ALDA, we have recruited graduate students who understand Arabic. This is mainly because user experiences, behaviors, and task performances differ according to the academic level of participants and their familiarity with Arabic reading and writing. Each participant is assigned a set of articles from the randomly sampled one thousand articles we have collected in our baseline dataset. For each article, students are tasked to read the article and its corresponding topic generated by ALDA. After that, students are asked to evaluate the topic of each article using a Likert scale, indicating their view of the relevance between the topic and the article. Each topic/article pair must be assigned a number between 0-10 for relevancy, where 0 means the topic is not relevant to the article and 10 means the topic is highly relevant to the article.

Figure 15 and Table 7 show the evaluation results of one thousand articles. Articles are divided into eleven categories (0-10), based on their rates. We had ten participants each evaluated two hundred random articles; where each of the one thousand articles was evaluated by two different participants. We averaged the evaluation score of each articles and counted the frequency of each scoring. As we explained in the previous paragraph, articles that received high topic scoring would indicate that the topic is more relevant to the article, as shown in Table 7, majority of the articles were scored between 7 and 10. From the figure below, it can be concluded that applying the LDA algorithm over Arabic news articles leads to achieving very good results in terms of generating accurate and relevant topics.

Table 7: Number of Articles for each Score

Rate Value	0	1	2	3	4	5	6	7	8	9	10
Number of Articles	0	0	0	15	49	71	74	125	269	241	156

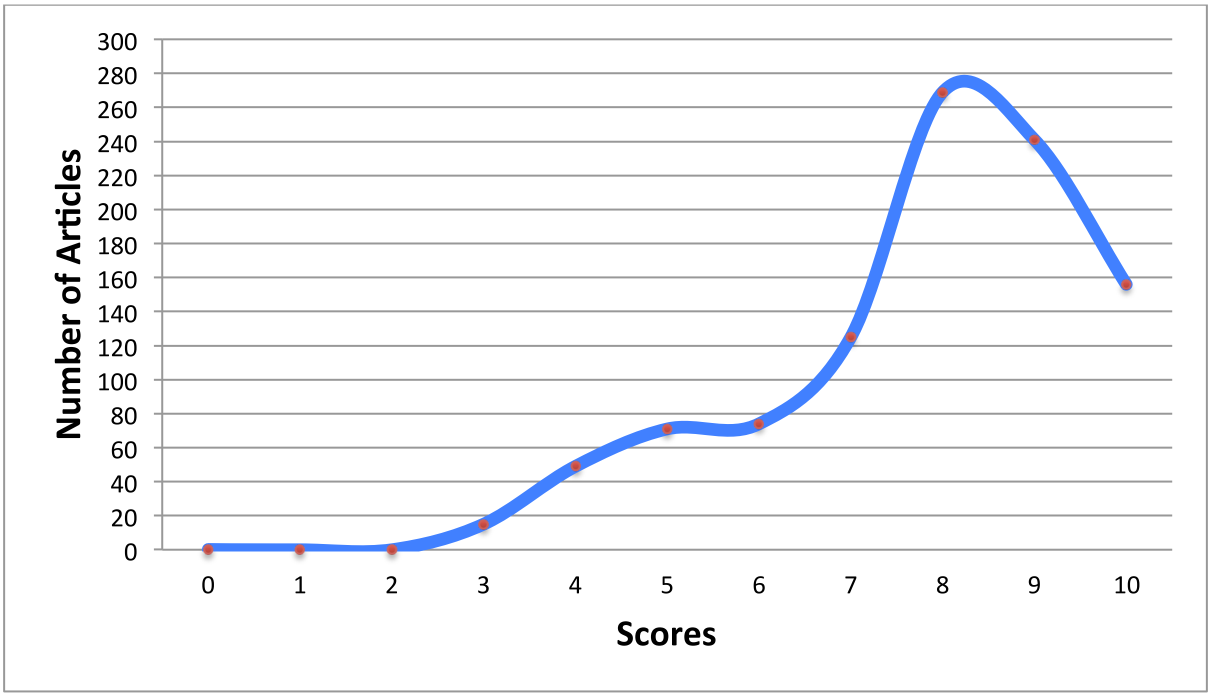


Figure 15: 11 Categories Used to Show ALDA Evaluation Results

4. Conclusion

Natural Language processing research involving the Arabic language is relatively hard, compared to other popular languages, e.g., English. Adding to that, available free NLP tools and resources are very rare. These issues form the source of inspiration for this research. There is no substantial research addressing the extraction of named entities from Arabic news articles. The same applies to generation of topics from articles.

In this study, we aim at developing a Named Entity Recognizer that can extract, with good accuracy, the named entities from Arabic news articles, called RenA. We also modify the popular topic extraction model, LDA, to enable it to handle and generate topics from Arabic news articles, called ALDA. Due to the lack of free resources for a judged news articles corpus, we have resorted to building a corpus, with the help of graduate students fluent with Arabic, to be used with RenA and ALDA evaluations, and later by other researchers.

We use Information Retrieval evaluation measures to evaluate and compare our RenA NER with another NER that is available through the LingPipe toolkit. We considered three types of named entities: Person, Organization, and Location. Our results show that using the proposed RenA enhances the named entity extraction results for the three mentioned types of entities which produced enhanced results than the alternative.

A second experiment, with graduate students who understand Arabic, helped to evaluate our ALDA tool. Using a Likert scale for assessment with our Arabic news article corpus, evaluation results confirmed that our developed tool generates highly relevant topics.

5. Future Work

Our future plan is to expand this research by using the RenA and ALDA results to fill in templates. We also are planning to extract more attributes to fill in templates, towards generating improved Arabic news article summaries. In addition, we aim to use another Arabic stemmer and to compare the ALDA results with the two stemmers.

Acknowledgements

Thanks go to the US National Science Foundation for their support through grants DUE-1141209 and IIS-1319578. We would also like to acknowledge QNRF for their support. This research was made possible by NPRP grant # 4-029-1-007 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the author. We would also like to acknowledge the group of students who volunteered to help building our baseline corpus and evaluated our topics.

References

[1] Habash, Nizar Y. "Introduction to Arabic natural language processing." Synthesis Lectures on Human Language Technologies 3, no. 1 (2010): 1-187.

[2] Yasine Benajiba. “Arabic named entity recognition”, PhD dissertation. Universidad Politécnica de Valencia. Valencia, Spain. 2009.

[3] David M. Blei. *Probabilistic topic models*. Communications of the ACM 55, no. 4 (2012): pp. 77-84. DOI:10.1145/2133806.2133826

[4] Thomas L.Griffiths and Mark Steyvers, “Finding scientific topics”. Proceedings of the National Academy of Sciences 101, no.1 (2004): pp. 5228-5235. DOI:10.1073/pnas.0307752101

[5] GitHub, Hyunjong Lee, “LatentDirichletAllocation- Implementation of Latent Dirichlet Allocation in C# (CSharp”. Feb. 7, 2014. [Cited on 02/15/2015]. https://github.com/hyunjong-lee/LatentDirichletAllocation.

[6] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Vol. 1: Cambridge University Press. Cambridge. 2008.

[7] Saleem Abuleil and Martha Evens. “Extracting Names from Arabic Text for Question-Answering Systems”. In proceedings of RIAO'2004, pp. 638–647, France. 2004.

[8] Ghassan Kanaan, Reyad Al-Shalabi, and Majdi Sawalha. “Fully automatic Arabic text tagging system”. In the proceedings of the International Conference on Information Technology and Natural Sciences, Amman, Jordan. 2003.

[9] Khaled Shaalan and Hafsa Raza. *NERA: Named entity recognition for Arabic*. Journal of the American Society for Information Science and Technology. 60(8): pp. 1652-1663. 2009.

[10] Kareem Darwish. “Named Entity Recognition using Cross-lingual Resources: Arabic as an Example”. In proceedings of the 51st Annual Meeting of the Association of Computational Linguistics (ACL), pp. 1558-1567. Sofia, Bulgaria, August 4-9 2013.

[11] James Allan, Rahul Gupta, and Vikas Khandelwal. “Topic models for summarizing novelty”. In ARDA Workshop on Language Modeling and Information Retrieval. Pittsburgh, Pennsylvania. 2001.

[12] Aurora Pons-Porrata, Rafael Berlanga-Llavori, and Jose Ruiz-Shulcloper. *Topic discovery based on text mining techniques*. Information Processing & Management. 43(3): pp. 752-768. 2007.

[13] Xing Wei and W. Bruce Croft. “LDA-based document models for ad-hoc retrieval”, in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM: Seattle, Washington, USA, pp. 178-185. 2006.

[14] Shixia Liu, Michelle X. Zhou, Shimei Pan, Weihong Qian, Weijia Cai, and Xiaoxiao Lian. “Interactive topic-based visual text summarization and analysis”, in Proceedings of the 18th ACM conference on Information and knowledge management. ACM: Hong Kong, China, pp. 543-552. 2009.

[15] Abderrezak Brahmi, Ahmed Ech-Cherif, and Abdelkader Benyettou. *Arabic texts analysis for topic modeling evaluation*. Journal Information Retrieval, 2012. 15(1): pp. 33-53. Doi:10.1007/s10791-011-9171-y

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent Dirichlet allocation*. Journal of Machine Learning Research, 3, pp.993-1022. 2003.

[17] Douglas Oard and Fredric C. Gey. “The TREC-2002 Arabic/English CLIR track”. In TREC2002 notebook, pp. 81-93. 2002.

[18] Leah S. Larkey, Fangfang Feng, Margaret Connell, and Victor Lavrenko. “Language-specific models in multilingual topic tracking”. In Proceedings of SIGIR 2004, pp. 402-409. Sheffield, UK. 2004.

[19] Thomas Hofmann. “Probabilistic latent semantic analysis”. In Proceedings of the fifteenth conference on uncertainty in artificial intelligence, pp. 289-296. 1999.

[20] Thorsten Brants, Francine Chen, and Ayman Farahat. “Arabic document topic analysis”. LREC-2002 workshop on Arabic language resources and evaluation, Las Palmas, Spain. 2002.

[21] Universite de Neuchatel (UniNE), Jacques Savoy, “IR multilingual Resources-Arabic Stopword List”. [Cited 02/15/2015]. Switzerland. http://members.unine.ch/jacques.savoy/clef/index.html. [This stopword list has been used for research purposes only]

[22] Apache Software Foundation, “Classic Arabic Analyzer”. 2013. [Cited 02/15/2015]. http://lucene.apache.org/core/4_6_0/analyzers-common/org/apache/lucene/analysis/ar/ArabicAnalyzer.html

[23] Google Code, “Stop-Words Project- collection of stopwords in 29 languages”. Feb. 24, 2014. [Cited 02/15/2015]. https://code.google.com/p/stop-words. [This stopword list has been used for research purposes only]

[24] SoruceForge, Al-Khalil Morphological System, “Al-Khalil Arabic Stemmer”. Feb. 21, 2011. [Cited 02/15/2015]. http://alkhalil.sourceforge.net/.

[25] The Center for Computational Learning Systems, Columbia University. Yasine Benajiba, “ANERCorp”. 2010. [Cited 02/15/2015]. http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

[26] LingPipe, a toolkit for processing text using computational linguistics. Alias-i. LingPipe 4.1.0. October 1, 2008. [Cited 02/15/2015]. http://alias-i.com/lingpipe.