THE BRITISH UNIVERSITY IN DUBAI

MSC INFORMATION TECHNOLOGY

ALNER: ARABIC LOCATION NAMED ENTITIES RECOGNITION

تحديد أسماء الأماكن بالنص العربي

MASTERS THESIS

SUBMITTED BY

HAITHAM MOHAMAD KADDOURA

OCTOBER 2010

SUPERVISOR:

DR. KHALED SHAALAN

THE BRITISH UNIVERSITY IN DUBAI

Abstract

This dissertation describes a rule based approach carried out to determine Location Named Entities in Arabic. ALNER, an Arabic Location Named Entities Recognition system, implements the rule based approach and is introduced in this thesis. This research is the first of its type to specialize in Location NER as a stand-alone system from other named entity types. Such dedication on one named entities helps in investigating the performance of comprehensive NER systems.

The Named Entity Recognition (NER) task has great influence on various Natural Language Processing (NLP) applications (e.g. Information Retrieval, Question Answering, etc.). Various research works conducted toward building language independent NER systems that will work on any language but very limited work has been done for NER systems to work with Arabic language.

It is known that Arabic language has complex morphology as a language which makes the NER task more difficult. Readers will find an overview about the Arabic language morphology and how it is different from other languages. We also highlighted the key challenges in Arabic language for the NER task. In addition, overall presentation about previous work toward Arabic NER is presented.

ALNER system using rule-based approach was evaluated and achieved accuracy of 87.27% and further investigation was conducted to study per module effectiveness and contribution.

نبذة

تهدف هذه الأطروحة إلى بناء نظام قائم على قواعد محذوفة مسبقة التعريف لتحديد أسماء الأماكن بالنص العربي، ولهذه العملية تأثير كبير على تطبيقات المعالجة الطبيعية للّنصوص العربية مثل استرجاع المعلومات ونظام إجابة الأسئلة، وكثيرة هي الأبحاث المماثلة التي تم إجراؤوها لبناء أنظمة تعالج نصوصا بلغات مختلفة ولكن الأبحاث في بناء أنظمة تعالج النص العربي كانت محدودة جدا.

من المعروف أن للّغة العربية قواعد صرف معقدة ما يجعل مهمة النظام أكثر صعوبة، وقدمنا في هذا البحث لمحة عن هذه القواعد التي تجعل اللّغة العربية مختلفة عن غيرها من اللّغات كما سلطنا الضوء على التحديات التي واجهناها أثناء بنائنا لهذا النظام وءشرنا إلى الأبحاث السابقة التي بحثت في نفس الموضوع.

تم تقييم النظام وحقق دقة بنسبة ٨٧.٢٧ %.

Acknowledgements

First, I wish to thank Allah for the full support I gained to be able to submit this dissertation.

Second, I am really grateful to Dr. Khaled Shaalan for help and guidance during the whole period of the project.

Special Thanks to my mother, father, and my family for understanding and continuous support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Named Entity Recognition (NER) task is used to identify and extract Named Entities (NEs) from open and closed domains text. This chapter aims to introduce readers on the problem and research questions that this paper is solving and answering. In addition, readers will find preliminary overview about Arabic language.

In this chapter, description of requirements to understand the thesis is given. In Section 1.1, we present overview in Arabic language along. Definition of the problem that we are solving is giving in Section 1.2. Followed by Section 1.3 that will state our motivation in this research. We highlighted the objective of this thesis in Section 1.4. In Section 1.5, we give a special focus on the research questions. Finally, we overviewed the roadmap of the thesis in Section 1.6.

1.1 Arabic Language Overview

Arabic is considered a Semitic language with more than 280 million speakers as a first language[1]. It has three forms [6]:

1. Classical/Traditional/Quranic Arabic: it can be found usually in religious and old writings.

2. Modern Standard Arabic (MSA): common language of all Arabic speakers and widely used in different media types. Most researches including this thesis uses "Arabic" to refer to MSA.

3. Colloquial Arabic: a simpliefied Arabic form that is used in our daily communication.

Arabic has some special features that differentiate it from other languages. Following is a summary of those features:

- No capital letter for Nouns nor starting of sentences

[1] http://en.wikipedia.org/wiki/Arabic_language

Arabic Text	English Translation	Affix (clitics)
ودبي	And Dubai	و (Waw)
كدبي	Like Dubai	ك (kaf)
القدس	The Jerusalem	ال (Alif-laam)
لدبي	To Dubai	ل (laam)

Table 1.1: Arabic words with affix attached

- The use of diacritics
- Using of Affixes and suffixes
- Nouns are in three forms; Singular, dual, and plural
- It is also an inflectional language: means that a word is composed of: prefix(es) + lemma + suffix(es). Where prefixes can be articles, prepositions or conjunctions as in Table 1.1. Suffixes are objects or personal/possessive anaphora
- Words derived from a root that is composed of consonants/radicals. Thus, Arabic is called a derivational language since all the words (nouns, verbs and adjectives) are derivate in templatic way; Lemma = Root+ Pattern [15, 6]
- Three cases for nouns and adjectives: nominative, accusative, genitive

1.1.1 The importance of Arabic Language

The importance of Arabic Language is increasing day after another. Recently, Yahoo!, a Leading International English Portal acquired Maktoob, the leading online community in the Arab world. The acquisition was due to Yahoo!'s strategy toward investments in the Arab world. Yahoo! now is to deliver new localized products to the Arab internet users[2]. Unfortunatelly, current researches tackling Arabic language are still at their early stages. But with the current international consideration shift toward the Arab region, an increase is expected in quantity and quality of researches conducted in NLP for Arabic language.

1.2 Problem Definition

In any text, names play vital role in a text for detecting, identifying and extracting content [34]. Hence, names recognition has been considered a very crucial and key improver for many applications in Natural Language Processing (NLP). In Arabic language, there is lack in natural language processing systems and this is due to the lack of the key systems (i.e. Named Entity Recognition tools) which is considered the backbone of NLP systems. Hence, in this thesis, we are focusing in building Arabic Named Entity Recognition system that focus on Location Named Entities which can be extended to include other Named Entities.

[2]http://thenextweb.com/me/2009/11/04/ado-yahoo/

1.3 Motivation

The motivation of this thesis is that some Natural Language Processing (NLP) applications use the output text from Named Entity Recognition application to improve their performance. Hence, creating new approach and system with high accuracy in recognizing Arabic Named Entities will have crucial effect on the Arabic NLP applications such as Machine Translation that includes Arabic language, Information Retrieval systems and Question Answering systems.

1.4 Objective

This thesis aims at proofing that Rules bases approaches works better than statistical approaches for Location Named Entities in Arabic which is due to the peculiarities of Arabic language. The research investigates Arabic linguistics approaches to determine Location named entities in Arabic and how we can emulate such approaches and used them in the NLP tools.

This thesis represents a new model of tackling Arabic NER in terms of: 1)providing a dedicated study on Location named entities specifically, we were able to build quite good location database that contains more than 3,500 entities, 2) a new approach of imitating Arabic language linguistics when determining location based entities where we developed and included modules to save knowledge about named entities extracted in every document, 3) our preliminary results show a quite enhancement in our system compared with other published work.

1.5 Research Questions

The aim of this research is to provide answer for the following questions:

- How Arabic Linguistics determines Location Named Entities?
- How we can imitate Arabic Linguistics in determining Location Named Entities.

Arabic linguistics have very high accuracy (almost 100%) in determining Location NE. Hence, trying to understand how they do the recognition job will mark great contribution to the filed. Then applying that will be the second contribution to the field by offering a state-of-the-art system that will have very high accuracy similar to Arabic linguistics.

This thesis represents a new model of tackling Arabic NER in terms of: 1) providing a dedicated study on Location named entities specifically, we were able to build quite good location database that contains more than 3,500 entities, 2) a new approach of imitating Arabic language linguistics when determining location based entities where we developed and included modules to save knowledge about named entities extracted in every document, 3) our preliminary results show a quite enhancement in our system compared with other published work.

1.6 Thesis Overview

In Chapter 2, we go through an overview about NER task in general along with the applications that are dependent on this task. In addition, approaches to tackle

the subject were presented with the evaluation criteria.

Chapter 3 give a literature review about Arabic NER task. It also provides introductory analysis about supporting tools that are usually used with Arabic NER. Moreover, it goes through difficulties faced by researchers to tackle and improve efficiency of their systems.

Chapter 4 introduce the Arabic Location Named Entity Recognizer (ALNER). The research work done to formulate a Location name extraction system. Detailed description of approach with the different modules used is presented next. Moreover, our efforts to build additional tools such as text converter are explained as well.

The evaluation methodology and evaluation corpus along with the results presentation and analysis are presented in Chapter 5.

In Chapter 6 we draw our conclusions and future directions for this research.

Chapter 2

Named Entity Recognition Task

Named Entity Recognition (NER) is considered as a subtask of information extraction. It aims to locate, identify and tag Named Entities (NE) to some predefined classes/types such as person, location and organization names.

In this chapter, we will give historical background about NER in Section 2.1. Then, in Section 2.2, we will overview some hot applications that use NER. After that in Section 2.3, we will present the main approaches that are commonly used in NER and the difference between them. Finally, we will end the chapter by providing how NER systems are evaluated in Section 2.4

2.1 Overview

Here you are talking about it in general so you may refer to other languages. A general idea about the NER tack including Conferences contribution on the field. What are the named entities (location, person, etc.).

The word "Named Entity" was formulated for the Sixth Message Understanding Conference (MUC-6) [19]. The term now is widely used in Natural Language Processing researches. The computational research aim at detecting, identifying and classifying named entities in text automatically. The research in its early stages (between 1991 to 1995) was in its early stages and published papers where relatively low and devoted to English language. It started to accelerate in 1996 since the first major event with dedicated task on the subject [19].

Many researches have been tackling English language but most of them were language independent and multilingual. In addition, some work was tackling specific language such the German in Conference on Computational Natural Language Learning (CONLL-2003), Spanish in CONLL-2002, Japanese in MUC-6 and Arabic which has started to receive a lot of attention after Automatic Content Extraction (ACE).

At the beginning, the problem was to recognize “proper names” in general [12]. But in later years, MUC-6 competition introduced “ENAMEX” type to include “Persons”, “Locations” and “Organizations” named entities. Some named entities such “Locations” can be divided to include subtypes such as city, state, country, etc [18]. In addition to “ENAMEX”, “TIMEX” type includes “date” and “time”. While “NUMEX” types include “money” and “percent”. All these types have been introduced by MUC. Moreover, some domain specific works such as bioinformatics leads to create special types related to the domain such as “protein”, “DNA” and “RNA” [41].

2.2 NER Integrated with Natural Language Processing Applications

It could be standalone and be very well integrated with a larger NLP application

In any text, names play vital role in a text for detecting, identifying and extracting content [34]. Hence, names recognition has been considered a very crucial and key improver for many applications in Natural Language Processing (NLP) [32, 43, 21, 39, 40, 17].. Those applications includes: Information Retrieval (IR) systems [27], Machine Translation [3, 44], and Question Answering (QA) systems [13, 1, 6, 4, 5].

2.2.1 Information Retrieval

Information Retrieval (IR) is the process of retrieving relevant objects including; documents, Web pages, relational databases and text, based on user query. IR science is interdisciplinary of computer science, mathematics, information architecture, linguistic and/or statistical[1]. Unlike Information Extraction (IE), IR returns multiple relevant objects based on some relevancy criteria defined. A good example of IR application is search engines. Users submit queries to search engines which then return of list of related results [11].

Some research works discussed and investigated about the relation between named entities in topics and the performance or retrieval systems [28]. The founding was that there is a strong relation between named entities and retrieval systems. The research was tackling and discussing English, German and Spanish languages. Further investigations are required to other languages to provide more information on the relation and how we can correlate it.

2.2.2 Question Answering

Users who are looking for exact answer for their queries would require trying Question Answering (QA) systems. Question Answering systems are systems that retrieve exact and accurate answer for different types of questions. Unlike IR applications, QA is the task of retrieving a single answer whether it is a word, sentence or paragraph. The retrieved answer is considered the most related result in data repository. QA may utilize structured database and/or NLP documents.

[1] http://en.wikipedia.org/wiki/Information_retrieval

Questions such as “who”, “where”, “what”, “when” and “how much” seek answers of type person, location, organization, time, prices names. Hence, detecting such names will result in accurate answer retrieval.

For instance; the question “where is the tallest tower/building in the world?”, would have an answer of type location. Therefore, answer retrieved from the following text: “Burj Khalifa is a skyscraper in Dubai, United Arab Emirates, and the tallest man-made structure ever built”[2] would be “Dubai, United Arab Emirates” since this is a location named entity.

To get precise result, NER task is used in QA to improve the quality of data repository. Because of the fact that names represent very high percentage of search quires [34, 1], identifying these names by NER systems will proportionally affect the accuracy of QA system.

Researchers identified four types of questions where each type has its own complexity and challenges. Those types are:

- Reason
  - Why
- Procedure
  - How
- Purpose/Objective
  - What

[2] http://en.wikipedia.org/wiki/Burj_Khalifa

- General
  - Factoid
  - Definition

Another classification has been stated in [6] as: “Casual questioner”,“Template questioner”,“Cub reporter” and “Professional information analyst”.

Good achievement in this field is to detect and recognize NE in the questions and in the indexed/available corpus. As stated earlier, enhancing NER will dramatically enhance the QA system. It was stated from best achieved researchers in TREC 2007. Authors of “Rose” QA system [31], which achieved an overall result of 48.4%, reported that the improvement on NER was the key factor to get higher results.

2.2.3 Machine Translation

Machine translation systems can be improved by utilizing the NER task. For instance; if we have the following text: “Unique Technologies FZC” which is an organization name and its correct Arabic translation is: “يونيك تيكنولوجيز ش . م . ح” (Bulkwalter Transliteration[3]: “ywnyk tyknwlwjyz $.m.H”). But if it was not considered as an organizational name, then the translation would be: “التكنولوجيا الفريدة فزس” (Bulkwalter Transliteration: “AltknwlwjyA Alfrydp fzs”), which is wrong. Hence, not being able to detect the Organization NE will lead to a different translation.

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

Many papers In the field describe how effective is the use of Name Entity Recognizer. In [23], authors presented how importance is Named Entities in machine translation. They provided measurement mechanism to show how effective is the named entity translation to the machine translation approaches. The research work has been done over English and Spanish translation.

2.3 Main Approaches

Researchers highlighted two approaches to achieve NER; one is linguistic grammar/rule based approach and second is statistical models. In early stages, rule based approach were the dominant approach. Most recent works are utilizing the statistical approaches to automatically induce rule-based systems from a collection of test data or corpus. [37] showed that if there is lack of huge training corpus, the rule based approach remains the preferred approach.

Following are more detailed description of each approach.

2.3.1 Rule Based

Rule based approach is the approach where linguistic is needed to define rules for every NE [1]. Rules vary from language to another. And rules that work for English might not work for Arabic. So for each language and for each entity type such as location, we require linguistic effort. This is the major concern for rules based approach. On the other side, researches showed that rule based approaches achieve near-human performance [1, 38, 39, 40].

2.3.2 Statistical Modeling

The idea in statistical approach is to study the features of positive and negative NE classes in a huge training corpus. The requirement of large annotated corpus is considered the main shortcoming of statistical approach [32]. Statistical modeling includes Decision Trees [36], Support Vector Machines (SVM) [2], Maximum Entropy (ME) [8, 14, 6], Conditional Random Fields (CRF) [29] and Hidden Markov Models (HMM) [7].

2.4 Evaluation of NER

NER systems are usually evaluated based on their output compared with linguistic output on the same text. Many techniques were proposed to evaluate NER systems based on their capability to imitate language linguistics in annotating texts. In this section we are going to present main scoring techniques that are used in well-known conferences.

In MUC conferences [20, 10], systems are evaluated in two ways: correct type detection ability (TYPE), and exact text finding ability (TEXT). TYPE is counted if the entity is assigned the correct type, regardless of boundaries given that there is an overlap. On the other side, TEXT is counted if entity boundaries are correct regardless of the type. In both ways TYPE and TEXT, there are three key measures: the number of correct answers (COR), the number of actual system finding (ACT) and the number of possible entities in the solution (POS). the final MUC score is the f-measure (MAF) which is calculated by

$$F_{\beta=1} = \frac{(\beta^2+1)*precision*recall}{\beta^2*(precision+recall)}$$

Where precision is calculated by COR/ACT or the percentage of correct NEs found by the system. It can be expressed as:

$$precision = \frac{Numberofcorrectnamedentitiesfoundbythesystem}{Numberofnamedentitiesfoundbythesystem}$$

and recall is calculated by COR/POS or the percentage of NEs existing in the corpus and which were found by the system. It can be expressed as:

$$recall = \frac{Numberofnamedentitiesfoundbythesystem}{TotalnumberofNEs}$$

In ACE conference, the evaluation has more complex procedure. This is because of the fact that in ACE, it was introduced more named entity levels such as "subtypes", "class" and "entity mentions".

In ACE evaluation, each entity has a parameterized weight which contributes to the maximal proportion (MAXVAL) of the final score. The weight differs from entity to entity based on ACE parameters. Moreover, customizable costs (COST) are used for missed entities, false alarms and type errors. The final score is called Entity Detection and Recognition Value (EDR) which is 100% minus the penalties (COST). ACE evaluation is considered the most powerful evaluation scheme because of the cost of error and the full coverage of the problem. But on the other hand, it is problematic since it considers all the parameters are fixed.

In the Conference on Computational Natural Language Learning (CoNLL) classes were taken into consideration were the following: PER, LOC, ORG, and MISC.

Haitham	B-PERS
Kaddoura	I-PERS
Went	O
to	O
Dubai	B-LOC

Table 2.1: IOB Tagging

It is worth to mention that the evaluations of the data were annotated using the IOB2 schema which is a variant of the IOB schema introduced by [33]. This tagging schema rules are as following:

- The words which are Outside NEs are tagged as "O".
- The tag "B-TYPE" is used for the first word (Beginning) of an NE of class TYPE.
- Words which are part of an NE of class TYPE but are not the first word are tagged as "I-TYPE" (Inside)

Table 2.1 is a sample output of the IOB tagging schema.

Chapter 3

Arabic NER task: A review

Our aim in this Master thesis is to investigate approaches followed to overcome the NER in a reliable and efficient way by comparing other researchers' contribution with the results obtained. We'll investigate and analyze results based on our understanding.

3.1 Challenges Tackled by Arabic NER task

Arabic is very high morphological complex language. It has the following challenges:

1. Lack of capitalization; unlike English and many Latin languages that have capitalization as a sign for NE. Arabic, on the other hand, doesn't have such feature which makes the detection of NE very hard to find.

2. Great sparseness (complete morphology): sparseness is defined as the combination of root and affixes. We require lots of pre-processing modules to tackle data sparseness. To solve data sparsness problem, there are two possible solutions:

   (a) Light-stemming: deleting and removing the affixes and keeping only the stem. Authors of [25] reported a comparative study between different techniques to lead for best results.

   (b) Tokenization: affixes are not removed but only separated by space character.

3. No standardization for transliterated words since we have sound letters (Alf, Wow, and Ya'a) that will sound like diacritics (Fatha, Dhamma, Kasra).

4. Ambiguity: For instance: Ahmad Abad can be Location name and Person name.

Type of NE	Percentages
Person	29.5%
Location	26%
Organization	20.5%
Miscellaneous	19.3%

Table 3.1: Percentage of questions in CLEF 2004 and 2005 containing NE per type

3.2 Related work

In this section, we will go through some important systems that were built for the NER in general and have identified Location NE results separately.

To our knowledge no published work has been conducted to specialize in Location Named Entities neither in Arabic nor in any other language. Instead, most of the published researches have investigated the location NE along with other named entities. Hence, our overview about the current researches took into consideration the researches that are for Arabic language and they show results for Location NE.

3.2.1 Rule Based Approach

TAGARAB [26] is an Arabic NER which utilizes the morphological and POS-tagging modules and then pattern-matching. For testing, they used fourteen articles from Al-Hayat CD-ROM as their corpus and used MUC-style scoring program to compare the TAGARAB output with the hand-tagged version. The Loca-

tion NE results were 85.3%, 94.5% and 89.7% for precision, recall and f-measure, respectively.

Abuleil [1] presents a fully heuristic rules based technique. The goal was to extract proper names from text to build a database of names to be used for Question Answering system. The system has been tested on 500 articles from the Al-Raya newspaper, published in Qatar. The corpus includes 97 location names. System identified 97 names using sub-class measure. On the other hand, it identified 96 names and miss-classified 1 name when using the major-class measure. The f-measure stated was 93% for Location NE. But the names are very few and it requires large scale testing. Moreover, no details were given about corpus annotation.

In their paper [35], Samy et al. developed a mechanism to make use of paralle corpora in Spanish and Arabic, where the NE tagger transliterate the Arabic text to Spanish and the NE tagger will identify the tags in Spanish which will guide in tagging the names in Arabic corpus. The implementation was based on pattern matching, lexical, orthographic and phonetic criteria. The corpus used was 1200 of paired-sentences. The evaluation was to compare the results from the tagger with manually annotated gold standard set. Their approach uses a filter to the Arabic words, which omitted the Stop Words from the possible transliterated candidates improved the precision. The results obtained reports high precision (84% improved to 90%) and recall (97.5%). The only consideration, which any researcher who is interested to implement the same or build on it, is to have a parallel corpus.

[39] presents a rule based approach system where they developed a whitelist

dictionary for names and a regular expression for name recognition. A filtration mechanism was also used. The Location based performance in terms of precision, recall and f-measure are 77.4%, 96.8% and 85.9% , respectively

In [22], authors focus on translation from Arabic into French of NE. They are following the rule based approach and focus on sports domain where they balance between grammar and lexical resources. They collected a corpus that contains a hundred text of which 200 are sports venues NE. They achieved precision, recall and F-measure as follows: 69%, 67% and 68%, respectively.

3.2.2 Statistical Approach

In [6], authors experimented using a statistical approach towards NER (Person, location & organization) using probabilistic models; Maximum Entropy and then further Conditional Random Fields (CRF). Authors built their own corpus, called ANERcorp, in order to train and test the CRF model. ANERcorp is composed of a training corpus and a testing corpus annotated especially for the NER task. The Location based performance combining all features in terms of precision, recall, and f-measure are 91.69%, 82.23%, and 86.71%, respectively.

Authors in [46] conducted a statistical approach for named entity detection and recognition. In their research, a mention can be either name, nominal or pronominal. Any entity is a cumulative of all the mentions in all the levels that refer to one conceptual entity. The system is trained and tested on the Arabic ACE 2003 and part of the 2004 data. The testing corpus contains 178 documents from Arabic Treebank, broadcast and newswire documents. The aim is to check the effectiveness of n-gram stemming feature in mention detection system. The

stemming n-gram features showed interesting results in terms of precision (64.2% vs. 64.4%), recall (55.3% vs 55.7%) and f-measure (59.4% vs. 59.7%).

3.3 Tools supporting the Arabic NER task

In this section, we are going to present some researches and techniques that are used to support NER applications.

3.3.1 Tagging

There are several tagging strategies that are implemented in Information Extraction applications. We will overview these tagging strategies and evaluate them.

The trivial (Triv) strategy uses single class for each slot type and additional "O" for other classes. The shortcoming of this strategy is when there is consecutive entities of the same class. In such case, both entities will be under one class since it lack the beginning and ending of each entity.

Another strategy is the IOB and its variant IOB2 tagging. In IOB2, there is a use of "B" for beginning of the class, and "I" for inside/inner of the class. So IOB2 tagging will appear as "B-type" and "I-type". On the other side, IOB (also called IOB1) uses "B-type" when it is appropriate and necessary only.

BIE tagging is another tagging strategy that differs from IOB in that it uses E-type to tag and ending entity in the class. Moreover, it utilizes BE-type for entities of single class.

In [42], author presented a comparison between those strategies and suggested

a new strategy called BIA. The new strategy is similar to IOB2 except for the use of A-type to mark the first token after a slot filler. In the comparison, it was proved that IOB2 strategy works the best in all tagging strategies.

3.3.2 POS

AMIRA-2.0 system[1] described in [16]. AMIRA-2.0 is based on supervised learning with no explicit dependence on explicit modeling or knowledge of deep morphology. The technology employs Support Vector Machines in a sequence modeling framework using YAMCHA toolkit[2]. The accuracy achieved is over 96%.

Authors of [30], presented two approaches to provide a POS tagging system. Their first approach was to use complex words that describe full words without the need of any word segmentation. Their second approach was segmentation based approaches. The results was better for word-based POS tagging compared with the segmentation based tagging (93.93% vs. 93.41%). But they reported that the word-based tagging was performing better for known words while segmentation tagging performs better for unknown words. In [45], authors presented a combination technique that will use Hidden Markov Model (HMM) with morphological analyizer. The aim of HMM is to represent the Arabic sentence structure so that they can consider the logical linguistic sequencing. The results of their system were 96%.

[1] http://www1.ccls.columbia.edu/~ryan/AMIRA/AMIRA-2.0.tgz

[2] http://chasen.org/~taku/software/yamcha/

3.3.3 Gazetteer

Currently most researches in the NER field depend fully or partially on Gazetteer. So how good is the gazetteer will result in good NER systems. Recently researchers were conducting the effectiveness of gazetteers and how to build them automatically.

Author of [24] proposed and implemented a pattern validation search where they can build gazetteers automatically from unlabeled corpus. Author compared the results with and without the automatically generated gazetteer where the f-measure with the gazetteer where higher.

Another research which presents as a sequence model that will use general features to achieve higher accuracy was conducted in [9]. Authors compound learning of sequence model with a gazetteer driven labeling algorithm to label tokens in unlabeled data. Authors claimed that their method will be easier to implement than Conditional Random Fields with same performance.

3.3.4 Morphological Analysis

In some languages, a typical method used for Named Entity Recognition is the morphological analysis.

In [2], authors presented a character-based chunking method for Japanese language. They analyze the input sentence to produce multiple answers. Then, each character is annotated with its character types along with its possible POS tags of the top n-best answers. Finally, they used a support vector machine chunker to select portions of the input sentence as Named Entity.

Another state-of-the-art tool was called YamCha[3] the is a generic, customizable and open source text chunker that uses Support Vector Machines (SVMs). The system performed the best in the CoNLL2000 Shared Task

3.4 Conclusion

To sum up this section; we have concluded the following results:

- NER in Arabic has been investigated very less. Published works about Arabic NER highlighted the high complexity of this task for Arabic language because of its complex morphology. It requires additional modules to overcome the morphology complexity.
- Statistical Approaches requires lots of data while Rule based approach requires linguistic team.
- Best results were achieved in Rule based because of its standardization which will allow to work in open domain, while statistical approaches highly dependent on the corpus/training set domain.

[3]http://chasen.org/ taku/software/yamcha/

Chapter 4

ALNER System

This chapter presents our proposed system. Then, it will give a detailed description of the supported modules and the benefit for each of them. In addition, it will present some supporting tools that were built to assist the main system with some sample output from the system.

In our system we used rules based approach instead of statistical approaches. The reason was that due to the lack of huge Arabic corpus that is available for researchers. Since Arabic lacks the capital letters which helps in detecting boundaries of NE (as explained earlier), Arabic on the other hand, is based on rules on how to detect NE with the understanding of sentence context. We either require a linguistic to define these rules and we implement them or we need huge corpus with identification of NE, which is not yet available. Unfortunately, most researchers are either building their own corpus and they build the system in a way where they test their systems on part of the corpus (which is small) and then they apply it on the whole corpus. Through our research, we didn't find any identical research that is testing based on others corpora.

4.1 Problems solved By ALNER

Our research is proposing a solution for the challenges stated in Section 3.1 with the focus on Location named entities only. Currently there is no state-of-the-art Named Entity Recognizer tool that is very accurate. In this case, we are following the divide-and-conquer approach to solve and come up with rigid NER system. Instead of working on all named entities at once, we believed that each named entity can be solved separately. We are starting with the location named entity recognition task, then, as we will discuss in Section 6.2, we will work on other named entities.

Location Named Entity type is placed second, in terms of appearance, in between other named entities. Hence, solving and detecting such entities will help

us in solving a big portion of the problem.

4.2 Overall Architecture of the proposed ALNER

Our work presented in this thesis is specifically designed for Location Named Entities Recognition in Arabic. In our system, we have adopted the rule-based approach using Arabic linguistic grammar-based techniques.

Figure 4.1 represents the system architecture. It demonstrates the participating modules along with the recognition process flowchart. The participating modules are:

1. Part of Speech (POS) tagging: used for detecting named entities, sparseness of data and name boundaries.
2. chunk creator: used for creating named entities of different sized
3. location indicator: used for sniffing data to find keywords that represents location entities and solves the transliteration challenge.
4. dictionary and gazetteer: used for known location named entities and solves the ambiguity challenge.
5. expert learning: used for imitating human memory in learning new location named entities

Following sections are detailed description of each module.

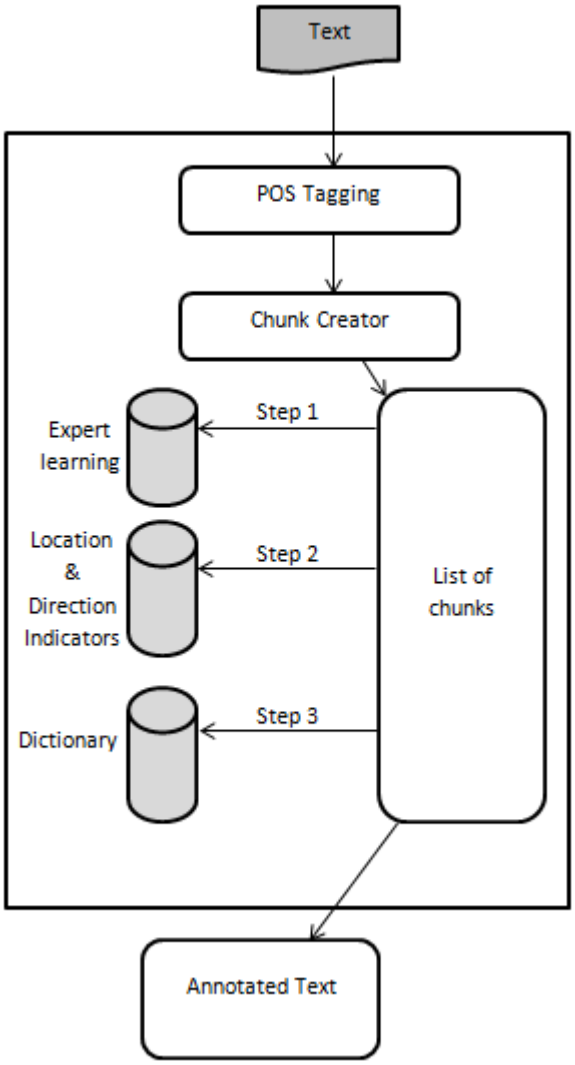

Figure 4.1: ALNER Architecture

4.3 Rules in ALNER

The set of rules were derived from translating written Arabic rules into computer friendly regular expression;

4.3.1 Rule 1

((Location_indication + ws)? + Location_name)

Where we defined a list of around 20 words for Location indication such as: "مدينة" (city), "دولة" (country).

This rule detects location named entities by checking nouns that are listed in the location indication list. Such rule detects Locations NE as in the following example:

..مدينة أبوظبي وهي عاصمة دولة الإمارات العربية المتحدة..

.. Abu Dhabi city is the capital of United Arab Emirates..

In this example, Abu Dhabi has been identified by the location indicator word "مدينة".

Enhancing and enlarging the location indication list will proportionally improve the accuracy of the system. Moreover, introducing new locations such as rivers, mountains, streets, etc will be done by including such indicators in the list.

4.3.2 Rule 2

((Direction_indication + ws)? + Location_name)

We created a list of 20 words as direction indication such as "غرب" and "جنوب"

..الشاطيء الجنوبي للّخليج العربي..

.. the southern shore of the Arabian Gulf..

In this example, "للّخليج العربي" (Arabian Gulf) has been identified by the direction indicator word "الجنوبي"

4.3.3 Rule 3

(Location_name (+ ws + direction)?)

..وتقع أبوظبي غرب مدينة دبي..

.. and Abu Dhabi is located to the west of Dubai..

In this example, "أبوظبي" (Abu Dhabi) has been followed by a direction indicator word "غرب"

4.3.4 Rule 4

من X حتّى Y

We call this rule: FROM-X-TILL-Y. This rule has two location entities “X” and “Y”. This is critical rule, because both entities must have the same type. But the same rule can also be used for Time and Location NE. Hence, Succeeding in detecting one of the entities by the other rules will guide us to detect the other NE. For example:

..من قاعدة دولة قطر غربا وحتّى رأس مسندم شرقا ..

.. From the base of the State of Qatar to the west and east to Ras Musandam

In the above example, we were able to detect Qatar as a location name using Rule 1 (the location indicator is “دولة” (country)). Hence, the named entities “رأس مسندم” (Ras Musandam) will be identified as Location NE.

4.3.5 Rule 5

من X إلى Y

Similar to Rule 4, the FROM-X-TO-Y rule has two location entities “X” and “Y”. Both entities must have the same type. But the same rule can also be used for Person and Location NE. Hence, Succeeding in detecting one of the entities by the other rules will guide us to detect the other NE. For example:

..وسافر رئيس الدولة من أبوظبي إلى فينّا..

..and His Highness Country President traveled from Abu Dhabi to Vienna..

In the above example, it is known that Vienna is a location name using Rule 7. Hence, the named entities "أبوظبي" (Abu Dhabi) will also be identified as Location NE.

4.3.6 Rule 6

$$X \text{ و } Y$$

This rule takes care of the Arabic grammar rule called X-AND-Y

..تكثر في إمارتي الفجيرة ورأس الخيمة نظرا

.. abound in the emirates of Fujairah and Ras Al Khaimah as..

Similar to Rules 4 and 5, both X and Y must have the same type. Hence, determining one of the variables will help us in determining the other.

4.4 Part of Speech (POS) Module

ALNER utilizes AMIRA 2.0 POS tagger that was described in Section 3.3.2. We have not yet compared results between AMIRA and other POS taggers. Due to its description and experimental results that used AMIRA, AMIRA were achieving the best results among others.

We used the POS module to help out identifying named entities from others to be tagged in ALNER. For instance, the following text:

شركات صناعة السيارات في ألمانيا تواجه عاما صعبا

The output of the POS is as follows:

شركات@@@**NNSFP** صناعة@@@**NNFS** ال#@@@**DET** سيارات@@@**NNSFP**
في@@@**IN** ألمانيا@@@**NNP** تواجه@@@**VBPFS3+S** عاما@@@**NN**
صعبا@@@**JJ**

In the output of the POS tagger, we can see that how it is differentiating the named entities with the class NN such as " شركات@@@NNSFP". ALNER uses the output to only tag such names and ignore the other classes since they are not have been identified as names.

4.5 Chunk Creator Module

This module will take as an input the tagged text from POS Module described in Section 4.4. It selects all consecutive words which were tagged as Named Entities. Along with the names classes, it is considering also the prefix entities that are identified by the system with "DET" class such as "ال#@@@DET ". Such entities are attached with the beginning of the next entity.

The aim of this module is to create a set of chunks from nouns that appeared in a consecutive order. This set of chunks will be examined by other modules to detect Location named entities. For instance, if we have four consecutive names, we will create a maximum of 10 chunks. The number of chunks is determined by:

$$\text{number of chunks} = \sum_{n=1}^{4} n$$

Chunk Set	Counter
Word1 word2 word3 word4	(1)
Word1 word2 word3	(2)
Word1 word2	(3)
Word1	(4)
word2 word3 word4	(5)
word2 word3	(6)
word2	(7)
word3 word4	(8)
word3	(9)
word4	(10)

Table 4.1: Chunk Creator process

The order of chunks will be evaluated as in Table 4.1.

The identification will start from chunk (1) and will apply Modules 4.6,4.7 and 4.8 respectively to detect whether the chunk is a location NE or not. If it is a location NE, the whole chunk will be reported as Location NE in the following manner:

Word1	B-LOC
Word2	I-LOC
Word3	I-LOC
Word4	I-LOC

And will start with another chunking process, starting from word5. If chunk (1) was not successfully tagged as Location NE, then we move to chunk (2) and we do the same process as above till we reach chunk (4). At this stage, Word1 failed to contribute and was never assigned as Location NE, hence, it will be tagged as O and the process will continue from chunk(5).

We stated earlier that DET entities will be considered along with Nouns. The reason is that such entities can appear between location named entities and we cannot set it as boundary ending. The following example will clarify this point more:

دولة الإمارات

The POS tagging will be:

NNSFP@@@إمارات DET@@@#ال NNFS@@@دولة

In this example, the location is the whole word "الإمارات" which has been tagged in the POS in two words: "ال#@@@DET " and "إمارات@@@NNSFP".

The later example shows another advantage of using the POS along with chunker modules is to provide hint for named entities boundaries. This feature helps in detecting compound location entities with more than one word. For instance the country name: "جنوب أفريقيا " which is tagged as "جنوب@@@**NN** أفريقيا@@@**NNP**"; two consecutive nouns.

To clarify the whole chunking process, let's assume the following POS tagged text:

IN@@@من **NNP**@@@أفريقيا **NN**@@@جنوب **IN**@@@ب **NNFS**@@@جولة

First, the word "جولة" has been tagged as "NNFS" followed by the entity "ب" with the tag "IN" so the sentence that will be chunked contains one word; "جولة" which will be examined using the other modules. In the seconds chunking process will go to the word "جنوب" which is tagged "NN", then "أفريقيا" which is tagged "NNP" and stop which it reach "من" because it is tagged "IN". The sentence that will be chunked is "جنوب أفريقيا". Since the sentence has more than one word, the chunking will be examined as follows: "جنوب أفريقيا" will be tested first, if it fails, then "جنوب" will be examined. If it fails, then the work will be tagged as "O". Then the chunking will continue with "أفريقيا" and do the same. In the case the whole sentence succeeded to be a location as in "جنوب أفريقيا", then the tagging will be as follows: "جنوب B-LOC" and "أفريقيا I-LOC".

4.6 Expert Learning

In Arabic, when there is any city name that is not well-known, we define it in details at the first time. Then we just mention the city name as if it is well known. In our system, we define a learning mechanism that keeps in memory any defined words and search for them for future appearance in the text. For instance,

..قاضي القضاة في مدينة مليلية المحتلة..

..Magistrate judges of occupied city of Melilla..

..احتل جل القبائل القريبة من مليلية و وصل إلى قلب تمسمان..

..Occupied the bulk of the tribes near Melilla, and reached into the heart of Tomsoman..

So even if the word Melilla does not exist in the dictionary and gazetteer, but we know that it is a city name as per the rules-based approach. For any further occurrence of the city name, it will not be detected as location NE, unless we apply this module. We have created a threshold for the expiry period of keeping detected location named entities in memory. The reason was that some words that are detected as location in some context should not be detected in another context. The best practice we found during our research is to set the threshold to expire after exiting the article. In other words, the context learning module is applicable only for each article separately.

4.7 Location Indication

This is a linguistic approach for understanding whether the NE is a location or not. In Arabic, location Names can be identified by different ways. Sometimes, there is a use of location indication then the location name then more detailed about the location. For instance:

..مدينة أبوظبي وهي عاصمة دولة الإمارات العربية المتحدة

.. Abu Dhabi city is the capital of United Arab Emirates..

..وتقع أبوظبي غرب مدينة دبي

.. and Abu Dhabi is located to the west of Dubai..

..وسافر رئيس الدولة من أبوظبي إلى فينّا

..and His Highness Country President traveled from Abu Dhabi to Vienna..

In the first example, the city name has been placed directly after the location indicator which is "مدينة" (city). We identified the city name as it came directly after the location indicator word. In the second example, the city name came before a direction indicator of a city which means the word is a city. In our third example, we are following the "from X to Y" rule where, both X and Y must be of similar type. In our case, if we identified either X or Y as a location NE, then we'll be able to identify the other word as a location NE.

Where we defined a list of around 20 words for Location indication such as: "مدينة", "دولة" and we created 20 words as direction such as "غرب" and "جنوب"

4.8 Dictionary & Gazetteer

ALNER dictionary and gazetteer module uses NERgazet[1] as it is base. NERgazet consists of three gazetteers; location, person and organization gazetteers. Since ALNER focuses on location named entities, it will consider only the location gazetteer from NERgazet. The location NERgazet contains 1,950 names of continents, countries, cities, rivers and mountains that were collected from the Arabic version of Wikipedia[2].

Our contribution in the dictionary and gazetteer was to clean and enhance NERgazet as a first step. Then we added additional location named entities to reach more than 3,500 records. Moreover, we migrated the text-based gazetteer to use MySQL database server[3]. Two reasons were behind this migration step; 1) MySQL database server record processing performs much faster than text file processing and 2) The organization of data is more efficient and effective than the text file processing. In addition, for every record in the database, there is a negative field. The advantage of this is to provide additional context understanding feature to solve ambiguity. For instance, the word "السلع" (Al Sila'a) has two implications; as 1) A city name, 2) Goods for trading. Hence, the negative words that can be added are "البضائع التجارية" (Commercial Products) since we are interested in this word only if it is a location name and adding such negative words will exclude the Goods for trading meaning. This feature adds context meaning to words, imitates human behavior in understanding sentences and eliminates ambiguity of the two meanings and selects the location named entity only.

[1] http://users.dsic.upv.es/grupos/nle/?file=kop4.php

[2] http://ar.wikipedia.org

[3] http://www.mysql.com

As we stated earlier, ALNER gazetteer was built based on the Location Gazetteer of NERGazet. Instead of using text file format, we deployed the gazetteer on MySQL database server and cleaned the records to end up with approximately 1500 location names as our first step process.

Second, we manually added All countries list and their capitals. Then, we started to build the cities per country records. Our initial efforts was to build the gulf countries city database and then globalized it to include Levant countries and after that for All Arab countries and the world. We achieved of building a Location Named Entity corpus with over 3,500 records.

One of our supporting tools was to allow us to automatically add group of cities at once without creating duplicates value. We are currently working on this tool to make it available online where users can add and download the Location Corpus in their appropriate format.

4.9 Conversion Tools

During our research, we found the need to build some additional supportive tools. First, we based our work on Arabic language and Arabic characters. Our developed main application is using Microsoft Visual Studio with .NET Framework 3.5 which must run on a Windows machines only. On the other hand, the POS tagger module which was described in Section 4.4 uses Yamcha software which works on Linux machines only. Moreover, the POS module works only with Bulkwalter text. These lead us to build a technique to allow us switch between Windows with Arabic text to Linux with Bulkwalter text.

Our main application uses Microsoft Visual Studio with .NET Framework 3.5 along with MySQL database. Aside from our main application functionality, we have extended the software to make some data preparation tasks such as converting corpus into text and vice versa.

In addition, we built some web pages to allow the communication between Windows and Linux workstations.

Moreover, we couldn't find any online tool that will make transliteration between Arabic and Bulkwalter formats rather than giving us the option to select whether we need the conversion to be XML friendly or not. We build these pages and implemented all those features. We are now working on hosting them online. Moreover, the output of POS is of three forms: tokenization, POS and Base Phrase Chunker (BPC). Our online conversion tool was able to convert these formats from Bulkwalter to Arabic format. The benefit of doing so is to allow us working with Arabic POS tagged text.

4.10 Sample of output

In Figure 4.2, we show sample of annotated text by ALNER. The text provided is from Wikipedia[4]. In the system output, each record has two columns, one for the word and the second for the tag.

[4]http://ar.wikipedia.org/wiki/

الإمارات	B-LOC	حدود	O	إلى	O
العربيّة	I-LOC	بحرية	O	الإمارات	B-LOC
المتّحدة	I-LOC	مشتركة	O	السبع	O
هي	O	من	O	التي	O
دولة	O	الشمال	O	شكّلت	O
عربية	O	الغربي	O	اتحادا	O
اتحادية	O	مع	O	فيما	O
تقع	O	دولة	O	بينها	O
في	O	قطر	B-LOC	وهي	O
شرق	O	ومن	O	إمارة	O
شبه	B-LOC	الغرب	O	أبوظبي	B-LOC
الجزيرة	I-LOC	مع	O	وإمارة	O
العربية	I-LOC	المملكة	B-LOC	دبي	B-LOC
في	O	العربية		وإمارة	O
جنوب	O	السعودية		الشارقة	B-LOC
غرب	O	ومن	O	وإمارة	O
قارة	O	الجنوب	O	رأس	B-LOC
آسيا	B-LOC	الشرقي	O	الخيمة	I-LOC
مطلة	O	مع	O	وإمارة	O
على	O	سلطنة	B-LOC	عجمان	B-LOC
الشاطئ	O	عُمان	I-LOC	وإمارة	O
الجنوبي	O	تأتي	O	أم	B-LOC
للخليج	B-LOC	تسمية	O	القيوين	I-LOC
العربي	I-LOC	الإمارات	B-LOC	وإمارة	O
لها	O	نسبة	O	الفجيرة	B-LOC

Figure 4.2: System Output

Chapter 5

Results and Evaluation

In this chapter, overview of the evaluation measures that we took into consideration to evaluate the system accuracy will be presented. Then, we will overview the corpus used in testing. After that, results and their analysis will be explained in more details.

5.1 Evaluation methodology

ALNER performance was measured by precision, recall and f-measure that are describe in Section 2.4. These measures are considered standard measures for NER systems[43]:

$$Precision = \frac{correctentitiesrecognized}{totalentitiesrecognized}$$

It is the total number of correctly recognized entities - some researchers called it (*true positive*) with respect to total entities recognized- (*true positive + false positive*).

$$Recall = \frac{correctentitiesrecognized}{totalcorrectentities}$$

On the other side, recall is the total number of correctly recognized entities- (*true positive*) with respect to total correct entities in the corpus- (*true positive + false negative*)

$$F_{\beta=1} = \frac{2*precision*recall}{precision+recall}$$

While the F-measure is considered to be the tradeoff of between precision and recall. As we stated in Chapter 4, the system is dependent on modules behavior, and we are considering all pre-requisite module (POS module) is working perfect without any error rate.

5.2 The Evaluation Corpus

We build our own corpus for the preliminary testing. The corpus is created from 100 articles from different sections such as national, international, sport news of Al Bayan daily newspaper, published in United Arab Emirates[1]. The corpus contains 2100 manually annotated location named entities. The total size of the corpus is 260KB and approximately 44000 words. The corpus contains location entities of various sizes such as 1, 2 and 3 words.

The corpus we built is used for getting some preliminary results. The lacks of time to submit this dissertation and create a result output where we can compare with the ANERCorpus are the main reasons to get into this step.

5.3 Results and discussions

We have conducted a preliminary testing using ALNER system with corpus described in Section 5.2. The results are encouraging. ALNER achieved precision, recall and F-measure: 82.76%, 92.31% and 87.27% respectively. We only take into consideration the Location Named Entities.

In Table 5.1, we took our results into more details to show the performance per module. This detailed analysis shows us the performance of each module, hence will give us hints where to improve. As the Table 5.1 shows in the correct (true positive) row, the percentage of dictionary, rule bases and expert are as follows: 16.5%, 8.5% and 75%, respectively. These percentages give us an indication on

[1] http://www.albayan.ae

	Dictionary	Rule Based	Expert
Correct (true positive)	16.5%	8.5%	75%
Mistake in labeling (false positive)	30%	10%	60%

Table 5.1: Comparison between ALNER Modules

how news articles are formatted. The foundation we figured is that 75% of the location entities are repeated in the same article of the corpus we created.

Apart from the Expert module which achieved 75%, we need to look at the performance of Dictionary and Rule Based modules. If we ignore the Expert Module, the percentages of Dictionary and Rules based will be: 66% for the Dictionary and 34% for the Rule Based approach. The 66% of the location entities was detected correctly by Dictionary module is due to the fact that the corpus is contains news about the United Arab Emirates and the dictionary contains most of the location entities in the Arabian Gulf countries. Moreover, the local newspapers rarely introduce the city names by location indicators which allow the rule based approach to detect. On the other side, rule based module detected 34% of the location entities most of them from the international news.

On the second row of Table 5.1, Mistake in labeling (false positive), Expert module detected 60% of the false location entities. This percentage is due to the false detection of Dictionary and Rule based modules. So, If we ignore the Expert Module, the percentages of Dictionary and Rules based for the false positive row will be: 75% for the Dictionary and 25% for the Rule Based approach. The 75% of the location entities was detected incorrectly by Dictionary module is due to not understanding the context where the location name can be used in more than

one context with different meaning. On the other side, rule based module detected incorrectly 25% of the total incorrect location entities. This is due to the rules that are used in detecting location and other named entities such as Rule 4, Rule 4 and Rule 6.

Moreover, the Expert learning module will work great with different articles rather than the entire corpus will be as one file. The more the data is separated, the higher accuracy we will get. The reason is that a Location named entity could be added to the Expert Learning in a context that no negative word was there. But in another article, the same named entity recognized could have negative word and should not be recognized as Location NE. Moreover, the expert learning module is more into context understanding not a general understanding.

Our presented approach will not work for other languages but only for those who have same as Arabic Rules and Arabic morphology.

Chapter 6

Conclusion and Future work

This chapter will overview and sum up the thesis contribution and overview the chapters we presented in this dissertation. After that, research directions which might be taken to achieve higher accuracy will be presented.

6.1 Conclusion

In this thesis, we have presented our achievements in the Arabic Location Named Entity Recognition task. Our aim is to identify Location Named Entity types within an open-domain Arabic text. We overviewed the various applications that benefit from NER and how the improvement on accuracy will improve other applications in the NLP field. An overview about the two approaches used to tackle NER challenges was presented. We described in details the challenges that face researchers in the Arabic NER task. Summary of researches conducted to contribute in Arabic NER task. Then, we overviewed our approach in details. We explored the various modules with their contribution to our ALNER project along with the supporting tools. We showed how our approach achieved 87.27% F-measure. To improve the system more, the accuracy per module has been conducted and highlighted the weaknesses of our system and how it can work better.

6.2 Future work

Our future plans are to extend the location dictionary to include streets, mountains, rivers, etc. Making the dictionary available online and allowing users to interact with the application to add more locations and download it in their preferred format. We will be extending the work to include other named entities such as Person and Organization names.

Bibliography

[1] Saleem Abuleil. Extracting names from arabic text for question-answering systems. In Christian Fluhr, Gregory Grefenstette, and W. Bruce Croft, editors, *RIAO*, pages 638–647. CID, 2004.

[2] Masayuki Asahara and Yuji Matsumoto. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 8–15, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[3] Bogdan Babych and Anthony Hartley. Improving machine translation quality with automatic named entity recognition. In *EAMT '03: Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools*, pages 1–8, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[4] Michele Banko, Eric Brill, Susan Dumais, and Jimmy Lin. Askmsr: Ques-

tion answering using the worldwide web. In *In Proceedings of 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 7–9, 2002.

[5] Yassine Benajiba, Mona T. Diab, and Paolo Rosso. Arabic named entity recognition using optimized feature sets. In *EMNLP*, pages 284–293. ACL, 2008.

[6] Yassine Benajiba and Paolo Rosso. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT&NLP within the Arabic World, LREC'08*, pages 26–31, 2008.

[7] Daniel M. Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. Nymble: a high-performance learning name-finder. In *ANLC '97: Proceedings of the fifth conference on Applied natural language processing*, pages 194–201, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[8] Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. Nyu: Description of the mene named entity system as used in muc-7. In *In Proceedings of the Seventh Message Understanding Conference (MUC-7*, 1998.

[9] Andrew Carlson, Scott Gaffney, and Flavian Vasile. Learning a named entity tagger from gazetteers with the partial perceptron. In *Proceedings of the 2009 AAAI Spring Symposium on Learning by Reading and Learning to Read*, 2009.

[10] Nancy Chinchor. Overview of muc-7/met-2. In *In Proc. Message Understanding Conference MUC-7*, 1999.

[11] Prabhakar Raghavan Christopher D. Manning and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[12] Sam Coates-Stephens. The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 26:441–456, 1992. 10.1007/BF00136985.

[13] Jim Cowie and Wendy Lehnert. Information extraction. *Commun. ACM*, 39(1):80–91, 1996.

[14] James R. Curran and Stephen Clark. Language independent ner using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 164–167. Association for Computational Linguistics, 2003.

[15] Ezra Daya, Dan Roth, and Shuly Wintner. Learning to identify semitic roots. In Nancy Ide, Jean Véronis, Harald Baayen, Kenneth W. Church, Judith Klavans, David T. Barnard, Dan Tufis, Joaquim Llisterri, Stig Johansson, Joseph Mariani, Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann, editors, *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, pages 143–158. Springer Netherlands, 2007. 10.1007/978-1-4020-6046-5_8.

[16] Mona Diab. Second generation tools (amira 2.0): Fast and robust tokenization, pos tagging, and base phrase chunking. In Khalid Choukri and

Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009. The MEDAR Consortium.

[17] Doug Downey, Matthew Broadhead, and Oren Etzioni. Locating complex named entities in web text. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, pages 2733–2739, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[18] Michael Fleischman. Automated subcategorization of named entities. In *ACL (Companion Volume)*, pages 25–30, 2001.

[19] Ralph Grishman and Beth Sundheim. Message understanding conference - 6: A brief history. In *Proceedings of the International Conference on Computational Linguistics*, 1996.

[20] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. In *COLING '96: Proceedings of the 16th conference on Computational linguistics*, pages 466–471, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

[21] Abdelmajid Ben Hamadou, Odile Piton, and Héla Fehri. Recognition and translation arabic-french of named entities: case of the sport places. *CoRR*, abs/1002.0481, 2010.

[22] Abdelmajid Ben Hamadou, Odile Piton, and Héla Fehri. Recognition and translation arabic-french of named entities: case of the sport places. Technical Report arXiv:1002.0481, Feb 2010.

[23] Lynette Hirschman, Florence Reeder, John D. Burger, and Keith Miller. Name translation as a machine translation evaluation task. Technical report, Proceedings of LREC'2000. ISO-8601 ftp://ftp.qsl.net/pub/g1smd/8601v03.pdf, 1997.

[24] Zornitsa Kozareva. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *EACL*. The Association for Computer Linguistics, 2006.

[25] Leah Larkey, Lisa Ballesteros, and Margaret Connell. Light stemming for arabic information retrieval. In Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann, editors, *Arabic Computational Morphology*, volume 38 of *Text, Speech and Language Technology*, pages 221–243. Springer Netherlands, 2007.

[26] John Maloney and Michael Niv. Tagarab: a fast, accurate arabic name recognizer using high-precision morphological analysis. In *Semitic '98: Proceedings of the Workshop on Computational Approaches to Semitic Languages*, pages 8–15, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[27] Thomas Mandl and Christa Womser-Hacker. The effect of named entities on effectiveness in cross-language information retrieval evaluation. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 1059–1064, New York, NY, USA, 2005. ACM.

[28] Thomas Mandl and Christa Womser-Hacker. How do named entities contribute to retrieval effectiveness? In Carol Peters, Paul Clough, Julio Gon-

zalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images*, volume 3491 of *Lecture Notes in Computer Science*, pages 833–842. Springer Berlin / Heidelberg, 2005.

[29] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *CONLL '03: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 188–191, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[30] Emad Mohamed and Sandra Kübler. Arabic part of speech tagging. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association, 2010.

[31] Dan I. Moldovan, Christine Clark, and Moldovan Bowden. Lymba's poweranswer 4 in trec 2007. In Ellen M. Voorhees and Lori P. Buckland, editors, *TREC*, volume Special Publication 500-274. National Institute of Standards and Technology (NIST), 2007.

[32] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January 2007. Publisher: John Benjamins Publishing Company.

[33] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In David Yarovsky and Kenneth Church, editors, *Proceed-*

ings of the Third Workshop on Very Large Corpora, pages 82–94, Somerset, New Jersey, 1995. Association for Computational Linguistics.

[34] L.F. Rau. Extracting company names from text. In *Proceedings Seventh IEEE Conference on Artificial Intelligence Applications, 1991. Proceedings.,*, pages 29–32, 1991.

[35] Doaa Samy, Moreno-Sandoval, Antonio, and José M. Guirao. A proposal for an arabic named entity tagger leveraging a parallel corpus (spanish-arabic). In *Proceedings of Recent Advances in Natural Language Processing RANLP 2005*, pages 459–465, 2007.

[36] Satoshi Sekine. Nyu: Description of the japanese ne system used for met-2. In *Proc. of the Seventh Message Understanding Conference (MUC-7*, 1998.

[37] Satoshi Sekine and C Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In *In Proc. Conference on Language Resources and Evaluation*, 2004.

[38] Khaled Shaalan and Hafsa Raza. Person name entity recognition for arabic. In *Semitic '07: Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages*, pages 17–24, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[39] Khaled Shaalan and Hafsa Raza. Nera: Named entity recognition for arabic. *J. Am. Soc. Inf. Sci. Technol.*, 60(8):1652–1663, 2009.

[40] Khaled F. Shaalan and Hafsa Raza. Arabic named entity recognition from diverse text types. In Bengt Nordström and Aarne Ranta, editors, *GoTAL*, vol-

ume 5221 of *Lecture Notes in Computer Science*, pages 440–451. Springer, 2008.

[41] Dan Shen, Jie Zhang, Guodong Zhou, Jian Su, and Chew lim Tan. Effective adaptation of a hidden markov model-based named entity recognizer for biomedical domain. In *In: Proceedings of NLP in Biomedicine, ACL*, pages 49–56, 2003.

[42] Christian Siefkes. A comparison of tagging strategies for statistical information extraction. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *HLT-NAACL*. The Association for Computational Linguistics, 2006.

[43] An De Sitter, Toon Calders, and Walter Daelemans. A formal framework vor evaluation of information extraction. Technical Report TR 2004-0, University of Antwerp, Dept. of Mathematics and Computer Science, 2004.

[44] Rohini Srihari and Erik Peterson. Named entity recognition for improving retrieval and translation of chinese documents. In George Buchanan, Masood Masoodian, and Sally Cunningham, editors, *Digital Libraries: Universal and Ubiquitous Access to Information*, volume 5362 of *Lecture Notes in Computer Science*, pages 404–405. Springer Berlin / Heidelberg, 2008.

[45] Imad Abdulrahman Al-Sughayeir Yahya Ould Mohamed El Hadj and Abdullah Mahdi Al-Ansari. Arabic part-of-speech tagging using the sentence structure. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009. The MEDAR Consortium.

[46] Imed Zitouni, Jeff Sorensen, Xiaoqiang Luo, and Radu Florian. The impact of morphological stemming on arabic mention detection and coreference resolution. In *Semitic '05: Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 63–70, Morristown, NJ, USA, 2005. Association for Computational Linguistics.