Journal of Computer Science

Original Research Paper

A Comparison between Conditional Random Field and Structured Support Vector Machine for Arabic Named Entity Recognition

Marwa Muhammad, Muhammad Rohaim, Alaa Hamouda and Salah Abdel-Mageid

Department of Computer and System Engineering, Al-Azhar University, Cairo, Egypt

Article history
Received: 02-01-2019
Revised: 30-12-2019
Accepted: 28-01-2020

Corresponding Author:
Marwa Muhammad
Department of Computer and System Engineering, Al-Azhar University, Cairo, Egypt
Email:
marwa.mohammad.matar@gmail.com

Abstract: The Named Entity Recognition (NER) is an integrated task in many NLP applications such as machine translation, Information extraction and question answering. Arabic is one of the authorised spoken languages in the united nation. Currently, there is much Arabic information on the internet, so, nowadays the need for tools which process this information becomes significant. In this study, we have examined the impact of the conditional random field and the structured support vector machine in the task of Arabic NER. The structured support vector machine is the first time to be applied in the Arabic name entity recognition. Our proposed system has three stages: Preprocessing, extracting features and building model. We have used simple features like the bag of words in the [-1,1] window, the bag of part of speech tag in the [-1,1] window to enable our system to detect the multi-words entities. Also, we have tried to enhance the Stanford part of speech tagger to enhance the tagger output tags, which enabled our system to differentiate between the name entities from the non-entities. In addition, we have employed the binary features of: Is a person, is a prename, is a pre-location, is a location and is an organization. Our system has been trained and tested on part of ANER Crop. The results have proved that the conditional random field-based Arabic NER system outperforms the structured support vector machine-based Arabic NER using the same features set.

Keywords: Natural Language Processing, Arabic Named Entity Recognition, Stanford Part of Speech Tagger Training, Conditional Random Field, Structured Support Vector Machine

Introduction

The Named Entity Recognition means to detect the named entities and then classify them into a person, location and organization. It is a vital task in Natural Language Processing (NLP) (Nadeau and Sekine, 2007).

Information Extraction (IE)

It is an application which relies on recognizing name entities and then retrieving related documents according to the input query. NER is involved in identifying name entities firstly in the input query and then in relevant documents. So NER is an urgent task in Information retrieval applications (Ahmad *et al*., 2016).

Question Answering

It is another application which utilizes the NER task. This application takes a question as input and produces a definite answer. It utilizes NER in question analysis to discover the name entities in the question. So NER is regarded as a primary task in question answering applications (Shaheen and Ezzeldin, 2014).

Machine Translation (MT)

Translation systems are inadequate in returning reliable translations of NEs. For example "حركة تمرد" is translated into the Insurgency Movement. The primary entity has two parts, the name part "تمرد," and this word should be translated as it "Tamarod" and it shouldn't be translated as normal English word since it is the name of the entity. Then, the kind part "حركة " should be translated into "Movement." So the NER is a significant task in machine translation applications such as (Alqudsi *et al*., 2012) since NER task helps to define the name entities in the sentence and according to that, the translation should work.

Search Results Cluster

This system could employ the NER by ranking the clusters supported the magnitude relation of entities

every cluster involves (Benajiba *et al*., 2009). Applying NER enhances the method of examining the type of every cluster and also improves the cluster approach concerning chosen options.

In this study, we have described our Arabic NER system. We have utilized the definition within the shared task of the Conferences on Computational Natural Language Learning (**CoNLL**). Within the sixth and the seventh editions of the Conference on Computational Natural Language Learning (CoNLL 2002[1] and CoNLL 2003[2]), the NER task was outlined to verify the correct names existing within the text and to classify them into the subsequent three classes: person names, location names and Organization names. For example in the sentence "Former Tunisian President Moncef Marzouki announced his resignation from the presidency of the Harak Tunisia Party", "Moncef Marzouki" is considered as a person name and "Harak Tunisia Party" is considered an organization.

We have built our classifier based on the Conditional Random Field (CRF) (Sutton and Mccallum, 2007) and the Structured Support Vector Machine (SSVM) (Tsochantaridis and Hofmann, 2005). Both algorithms are considered a state of art algorithms in the task of Arabic NER. The SSVM has not been used in the Arabic NER up till now. CRF is a discriminative undirected graphical model, while SSVM is a discriminative model based on large margin theory (Tsochantaridis and Hofmann, 2005).

Our system aims to improve the Arabic NER task by the following:

1. We have enhanced the Stanford Part Of Speech tagger (POS) accuracy to tag our dataset. So, our system could discriminate between entities and nonentities thanks to the Noun, Proper singular tag (NNP)
2. The binary features of is the prename and is a pre-location are employed to enable the system to discover the person and location entities that are not covered by the binary features of is Person and is Location
3. The bag of words in [-1,1] window (The Unigram and Bi-gram word feature) is used to enable our system to recognize the single word and multi-words entities such as جامعة الدول العربية, which is translated as (The Arab league). جامعة is considered as B-ORG, الدول is considered as I-ORG and العربية is considered as I-ORG. So, our system combines between every two consecutive words and two consecutive part of speech feature to be like جامعة/ الدول, الدول/ العربية, جامعة/ NNP/ NNP and NNP/NNP
4. Our Gazetteers cover many types of texts such as political, sport, art and religion domains to enable our proposed system to recognize entities in these domains

[1]https://www.clips.uantwerpen.be/conll2002/ner/
[2]https://www.clips.uantwerpen.be/conll2003/ner/

Arabic Challenges in the Context of NER

Arabic and its Variations

Arabic has variations or dialects in the Arab world. It has varieties of formal and informal languages. The informal style is the used language between families. Classical Arabic is the expression of the Quran (Abdel Monem *et al*., 2008). So students who analyze the Quran or Hadith should follow classical Arabic. Modern Standard Arabic (MSA) follows the recent expression which is used in most universities. Universities in all Arab world concentrate on MSA in their studies. Also, MSA is used in newspapers, magazines and letters. Colloquial Arabic is known as lah-Jaa. It is the language which is practiced by Arabs in their regular life. It is a language of several regions in the same country (Farber *et al*., 2008; Shaalan, 2014).

To be able to solve this problem in our research, our system operates only on Standard dialect.

Lack of Capital Letters

Arabic as a language doesn't look like the languages that use the Latin script since the Arabic lacks the capital letters at the beginning of the Name entities. So, it causes the Arabic ambiguity (Farber *et al*., 2008).

To solve this problem in our system, we used the POS feature, so, the system can distinguish between the entities and non-entities.

Agglutination

An Arabic word may combine one or more prefixes, stem and one or more suffixes in different ways. Clitics are separated in English but not separated in Arabic. They involve conjunctions like و (Waw and) and ف (if, then), or two clitics combinations such as وب (Waw - baa), for example, وبالقاهرة, which means (and by Cairo), القاهرة, which means (Cairo) is a location name, وب is a combination of two clitics و and ب. So, segmentation is a significant step in Arabic (Oudah and Shaalan, 2012).

To solve this problem in our system, we used the FARASA tool[3] to segment the dataset to its separated words, prefixes and suffixes.

Optional Diacritics

Arabic includes diacritics which may change the phonetic representation and give another meaning to the same lexical form. Now Arabic is written without diacritics like many of texts appear in media. Also, this led to ambiguity (Alkharashi, 2009) for a computational system. For example (مصر) which means (Egypt as a location name) and (مصر) which means (insisted on), (أحمد) which means (Ahmed as a person entity) and (أحمد) which means (I thank).

To solve this problem, we normalized the text in the preprocessing stage to get rid of the diacritics and the

[3]http://qatsdemo.cloudapp.net/farasa/demo.html

system should define the type of entity according to the other features and from the context itself since our system works on [-1,1] window.

The Ambiguity between Name Entity Types

In Arabic, the same word spelling can be one or more name entity types. For example, الأحمدي (Al Ahmadi) may be a location or person which should be known from the general context. (Shaalan, 2014).

Arabic Writing Styles not Uniformity

An Arabic word may have many writing styles (Shaalan and Raza, 2007) because of the lack of standardization. For example, سورية and سوريا are two styles for the same word Syria location name.

Spelling Mistakes

Arabic writers make some mistakes. For example, ة tied (ta), which is used to indicate the feminine. It can be written ه (Ha) instead of Ta (Shaalan, 2014).

To solve this problem, the Arabic spelling checker can be used. There are many available Arabic spelling checkers online such as FARASA Arabic checker[3].

Related Work

There are four approaches have been used as follows:

Rule-Bases Approach

This approach is based on strong linguistic knowledge since it relies on grammatical rules. The rule-based approach is based on developing rules to recognize and classify entities into various tags. However, if there is any update, it will be time-wasting. Also, it will be a crucial problem if knowledge of Arabic rules is unavailable.

Shaalan and Raza (2007) have suggested PERA. It has three components, a gazetteer, a grammar and a filtration mechanism. The gazetteer checks the input text to define the matching names. Then, the input text is provided to the grammar that is made as regular expressions, to identify the name of the person entity. Finally, the filtration mechanism is used to eliminate ambiguous and invalid NEs. PERA obtained satisfactory results on ACE and the Treebank Arabic datasets.

As a continuation of PERA work, NERA was introduced by (Shaalan and Raza, 2008) and Shaalan and Raza (2009). NERA is a rule-based system to recognize NEs of 10 types: person, location, organization, date, time, ISBN, price, measure, phone numbers and filenames. The system implementation was within the FAST ESP framework. Also, NERA has three parts, such as the PERA system with the same functionalities but to include the 10 NE types. According to their results, the system obtained satisfactory accuracy.

Zaghouani (2012) has suggested a rule-based Arabic NER system (RENAR) to obtain person, location and organization NEs. The system has three stages: (1) Morphological pre-processing, (2) examining NEs and (3) applying linguistic rules to obtain unknown NEs. RENAR outperforms ANERsys 1.0 (Benajiba *et al*., 2007), ANERsys2.0 (Benajiba and Rosso, 2007) in extracting Location NEs when applied to ANERcorp dataset.

Aboaoga and Aziz (2013) have suggested a rule-based Arabic NER system that obtains person entities. The system includes sports, politics and political economy domains. The authors have made four main linguistic rules. The system goes through three steps for Arabic person names recognition: (1) Pre-Processing (tokenization, data cleanup and sentence splitting), (2) Automatic NE tagging and (3) Using the grammar rules to extract person names that don't exist within the inherent dictionaries. The dataset was gathered from entirely different online Arabic newspapers. The results illustrated that the system obtained the best accuracy within the sports domain compared to other domains.

Machine Learning Approach

This approach depends on large tagged data sets for training and testing. It requires a set of features obtained from datasets to build statistical models for the entity. It is preferred since it can be maintained at any time with insignificant effort and time if there is enough tagged data.

Benajiba *et al*. (2007) have built ANERsys. They created the ANERsys based on maximum entropy. Gazetteers were created to examine the effect of external resources on ANERsys. Hence, they have created their corpora which we have used to train and evaluate our proposed system. Their system recognized four types of NEs: Person, location, organization and miscellaneous. The ANERsys 1.0 system had some problems in detecting multi-word NEs. So, they created ANERsys 2.0 (Benajiba and Rosso, 2007) which consists of a two-step mechanism for NER: (1) Boundary detection of NEs and (2) NE types classification.

Benajiba and Rosso (2008) have used CRF to enhance performance. Each of the ANERsys 2.0 and CRF-based system has the same set of features including POS tags, base phrase chunks, gazetteers and nationality. The CRF-based system has obtained higher accuracy results.

Then (Benajiba *et al*., 2008) have created another NER system using SVM. The features are contextual, lexical, morphological features, gazetteers, POS tags, BPC, nationality and English capitalization. The system has been evaluated by utilizing ACE Corpora and ANERcorp. They obtained the most effective results when they applied all the features introduced earlier in the system.

Deep Learning (DL) Approach

Currently, Dl-based NER models are the dominant method and it obtains a state of art results. Deep learning is helpful than feature-based approaches in specifying hidden features automatically. DL is a field of machine learning

that is consisted of many processing layers to discover representations of data with various levels of abstraction. The DL has two basic operations: feed-forward and back-propagation. The forward pass calculates a weighted sum of their inputs from the preceding layer and transfers the result into a non-linear function. The backward pass is used to calculate the gradient of an objective function concerning the weights of a multilayer stack of modules through the chain rule of derivatives (Li *et al*., 2018).

Applying deep learning techniques to NER has three benefits:

1. DL generates non-linear mappings from input to output and this is considered more beneficial for the NER. Compared with linear models (e.g., log-linear HMM and linear-chain CRF), DL models can discover complex features from data through non-linear activation functions
2. DL saves significant work in designing NER features. The feature-based approaches require engineering skills and experience
3. DL NER models can be trained by gradient descent and this enables us to design possibly complex NER systems

Mohammed and Omar (2012) have developed an ML-based system, which uses the Artificial Neural Networks approach. The system has three main stages: (1) Pre-processing which includes (data cleaning, text segmentation and tagging), (2) Arabic Romanization and (3) Using the Artificial Neural Networks classifier for the text. The system could recognize four types of NEs: Person, location, organization and miscellaneous. They have used ANERcorp dataset for evaluation. The experimental results demonstrate that the Artificial Neural Networks approach obtained better accuracy than the decision tree approach.

Gridach (2018) has developed an Arabic NER system based on DL which learns automatically features from data. His result illustrates that his approach outperforms the model based on Conditional Random Fields by 12.36 points in F-measure. The important characteristic in his system that it can be easily updated and extended to extend other named entities without any additional features or rules.

Attia *et al*. (2018) have developed a Deep Neural Network system that combines the word and character-based representations in convolutional and recurrent networks with a CRF layer. They have combined the Modern Standard Arabic and Arabic Egyptian dialect. Their system is ranked the second among those participating in the shared task achieving FB1 70.09%.

Hybrid Approach

This approach combines both of the first and second.

Oudah and Shaalan (2012) examined adding of the rule-based approach and the ML-based approach for Arabic NER. The rule-based element is a re-implementation of NERA (Shaalan and Raza, 2008), but the ML-based element uses the decision tree, Support Vector Machine and logistic regression classifiers to evaluate the performance of the system. The system recognizes 11 types of entities. The results of the hybrid approach exceed both results of the rule-based and the machine learning-based approaches.

The Proposed Approach

The proposed system for Arabic named entity recognition (ANER) has three stages (Fig. 1): (1) Pre-possessing, (2) Extracting features and (3) Building Arabic NER model. The following entity types from ANERcrop are used in this research: B-PERS: The beginning of person name, I-PERS: The Inside of person name, B-LOC: The beginning of location name, I-LOC: The inside of location name, B-ORG: The beginning of the organization name, I-ORG: The inside of the organization name, O: Other words that are not proper nouns or digits.

Data Set

In our proposed system, the ANERCrop[4] has been used (Benajiba and Rosso, 2007). Benajiba and Rosso have been built the ANERCrop considering the same classes in CoNLL. They tagged all corpus tokens manually. It contains 150,286 tokens. The NEs are 11% of the corpus and their distribution along all entity types is illustrated in Table 1. We have used 70,000 for NER task and 80,000 for Stanford POS training task.

Pre-processing

It has the following phases:

- Segmentation is used to break the words into a prefix (es), stem and suffix (es). For example, "فأنزلنا"; which is translated into and we send down; consists of a prefix "ف," a stem "أنزل," and possessive pronoun "نا". It is a vital task because of the plenty of Arabic morphology (Benajiba and Rosso, 2008). It enhanced their F-measure from 67.76% to 70.67%. In our research, FARASA (Abdelali *et al*., 2016) tool is used for corpus segmentation (Fig. 2 and 3)
- Data Cleaning is used to convert the shape of some words into a suitable shape to be processed easily in our system Fig. 4
- Normalization is used to convert a list of letters in the words into their famous form. So, in this step, the letters of " أ," " آ," and " إ" would be converted into "ا". Also, the letter of "ة" into"ه". This step is so important since the model may be trained on "احمد" and then it can't recognize "أحمد" because of a different way of writing the same word. Also, it may that "أحمد" can't be recognized by person gazetteer. So all gazetteers, lists and dataset have been normalized

[4]http://users.dsic.upv.es/~ybenajiba/downloads.html

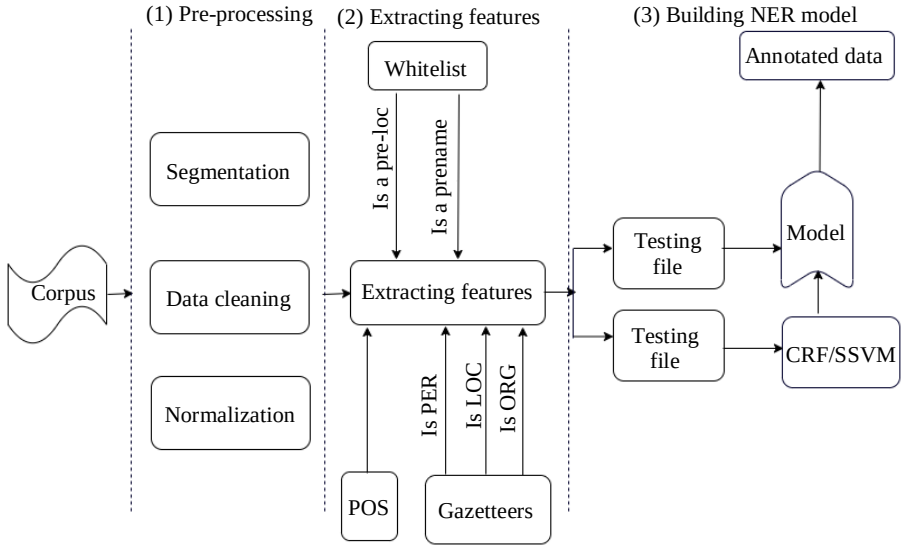

Fig. 1: ANER model

والیوم ، یتابع الولید بن طلال أعماله التجاریة عقب أن وحد استثماراته في مجموعة تتابع إمبراطوریته الاستثماریة الضخمة تحت اسم " شركة المملكة القابضة " والتي تتخذ من العاصمة الریاض مقرا لها ، وتنطوي تحت هذه الشركة عدة شركات عالمیة یمتلكها الولید أو یمتلك حصصا فیها ، فیما تستقر المكاتب الرئیسیة لالشركة في مبنى برج المملكة الذي یعد معلما بارزا في العاصمة السعودیة نظیر تصمیمه الفرید الذي حصل ، عام 2003 ، على جائزة أجمل تصمیم لمبنى برجي في العالم.

Fig. 2: The non-tokenized input text to FARASA

و+الیوم ، یتابع الولید بن طلال أعمال+ه التجاریة عقب أن وحد استثمارات+ه في مجموعة تتابع إمبراطوریت+ه الاستثماریة الضخمة تحت اسم " شركة المملكة القابضة " و+التي تتخذ من العاصمة الریاض مقرا ل+ها ، و+تنطوي تحت هذه الشركة عدة شركات عالمیة یمتلك+ها الولید أو یمتلك حصصا في+ها ، فیما تستقر المكاتب الرئیسیة ل+الشركة في مبنى برج المملكة الذي یعد معلما بارزا في العاصمة السعودیة نظیر تصمیم+ه الفرید الذي حصل ، عام 2003 ، على جائزة أجمل تصمیم ل+مبنى برجي في العالم.

Fig. 3: The output text of FARASA Segmenter

The letter	Before cleaning Data	After Cleaning Data
ال.	ال+مسلم.	المسلم.
ات، ون ، ین ، ان.	المسلم+ون ، المسلم+ات، مسلم+ین، مسلم+ان.	المسلمون، المسلمات ، مسلمین، مسلمان.
ت.	ألحق+ت.	ألحقت.
ة.	جدید+ة.	جدیدة.
ف ، ب، ل، و.	ف+إن، و+ب+حسب، ل+یعلم	ف إن، و ب حسب، ل یعلم.
ك، كما، كم، ه، ها، هما، هم، هن، نا، ي.	ان+ك، ل+كما، على+كم، على+ه، ان+ها، كتاب+هما، كتب+هم، كتب+هن، وطن+نا، وطن+ي.	ان ك، ل كما، علي كما، علي ه، انها، كتا بهما، كتب هم، كتب هن، وطن نا، وطن ي.

Fig.4: Data cleaning rules

Table 1: Ratio of NEs per class.

Class type	Ratio
PERS	39%
LOC	30.4%
ORG	20.6%
Miscellaneous class	10%

Extracting Features

We have used the following features for each token.

Word Itself

The token or the word is considered a feature in our proposed system.

Bag of Words

This bag includes the word itself, the word preceding the word itself and the word following the word itself. So, it includes unigrams and bigrams of the token in the window of [-1, 1].

Bag of POS Tag

Each token is associated with its unigram, bigram of POS tag in window of [-1, 1]. It is a significant feature because NNP tag discriminates between the entity and the non-entity; it is done in this research by training Stanford POS Tagger on our manual tagged corpus to get better tags - Our POS corpus is trained on the part of ANERcrop-. We have tagged the POS tagger training manually according to the Penn Treebank II tag set, which is available on Computational Linguistics and Psycholinguistics research centre (CLIPS)[5]. We have succeeded to tag about 80,000 tokens manually.

[5] https://www.clips.uantwerpen.be/pages/mbsp-tags

After Stanford Arabic POS tagger training, it has been tested and the results show that our Stanford POS Arabic tagger has obtained 92% total right tags which exceed the Stanford Arabic POS Tagger that has obtained 85% of Stanford Arabic POS Tagger.

Is a Person Binary Feature

Our person (Per) Gazetteer defines this feature value, if the word was in the Per gazetteer, then its feature value should be assigned to one; Otherwise, it should be assigned a zero value. It includes 3709 complete names of people. The Per gazetteer covers the domains of sport, politic, art, and religion domains. It includes the following names: The names of Allah, Allah's messengers, preachers, messenger's companions from men and women. It also includes countries presidents, countries prime ministers, foreign affairs minister, defence ministers, the famous names of footballers, handball players, motor racing players, TV presenters, Political people, beauty queens, writers, journalists, actors and actress.

It is a binary feature. Hence, if this token is included in the person (Pers) gazetteer, then it would be assigned the value of one. Otherwise, it would be assigned the value of Zero. For instance, if the word "عمرو" is in the Pers gazetteer, then in the following sentence, " لقد ألقى السيد عمرو موسى الكلمة الافتتاحية اليوم بالجامعة العربية", which means Today, Mr.Amr Mussa delivered the open speech of Arab League, the column of Is a person binary feature for the row of word "عمرو" will take value of one.

Is a Pre-Name Binary Feature

Our pre-name whitelist defines this feature value, so, if the pre-name is in the pre-name whitelist, then its feature value should be assigned to one; otherwise, it should be assigned to zero. It contains 123 of unrepeated prename tokens.

It includes the people titles (president, professor, etc.), nationalities and some adjectives which precede the person entities such as الممثل عادل امام, which means (the actor, Adel Imam). The sentence includes the adjective الممثل, which means (the actor) which may introduce person entities.

Is a Location Binary Feature

Our location (Loc) Gazetteer defines this feature value; if the word is in the location gazetteer, then it should be assigned to one; otherwise, it should be assigned to zero. It includes 1950 complete names of real locations in the world. The loc gazetteer includes the names of world countries, capitals of world countries, governorates, states, areas, cities, continents, villages, seas, oceans, streets, rivers and squares.

For instance, if the word "القاهرة", which is translated to Cairo, is included in the Loc gazetteer. Then the column of this feature for the row of "القاهرة" would be assigned the value of one in the sentence of: " وصل اليوم الى القاهرة المستشارة الألمانية أنجيلا ميركل", which is translated into Arrived in Cairo today, German Chancellor Angela Merkel.

Is a Pre-Location Binary Feature

Our pre-location (Pre-Loc) whitelist defines this feature value; if the word is in the pre-loc whitelist, then it should be assigned to one; otherwise, it should be assigned to zero. It includes 45 complete pre-locations names. The pre-loc whitelist includes some names and adjectives that precede the location names such as مخيم اليرموك which means (Yarmouk Camp). The name of ,مخيم which means (Camp) that preceded some locations.

Is an Organization Binary Feature

Our organization (Org) Gazetteer defines this feature value; if the word is in the organization gazetteer, then it should be assigned to one; otherwise, it should be assigned to zero. It includes 424 complete names of famous organizations around the world. The organization gazetteer includes the names of governmental organizations and non-governmental organizations. Governmental organizations include the names of ministries, universities, schools, armies, etc. The non-governmental organizations include the name of mosques, churches, clubs, sports teams, political parties, companies and charity organizations**.**

For instance, if the word "الجزيرة", which is translated to Aljazeera, is included in the gazetteer. Then the column of this feature for the row of "الجزيرة" would be assigned the value of one in the sentence of " قال الدكتور فارس السقاف في تصريح للجزيرة نت", which is translated into Dr. Faris al-Saqqaf said in a statement to Al-Jazeera Net.

Building NER Model

In this research, we investigated two machine learning algorithms CRF and SSVM. We have used both Unigram and Bi-gram features in the window [-1, 1] to build Arabic NER model.

CRF is a discriminative undirected probabilistic graphical model. It is considered a sequence labelling algorithm. In training, the conditional probability is maximized. It is used in Arabic NER successfully. CRF has a few implementations (Sutton and Mccallum, 2007). We have used the CRFSharp tool[6] for our experiment. CRFSharp provides a template file, training file and testing file. The template file has been used to specify our unigram and Bi-gram features. The Training file is considered as a row for each word and column for each feature; the last column contains the entity type. The test file looks like the training file except that the last column is empty.

[6]https://github.com/zhongkaifu/CRFSharp

The SSVM combines the advantage of maximum margin classifier, kernels and efficiency of HMM. SSVM is a large margin discriminative algorithm. It is used for structural data like sequences, trees and bipartite graph. It is suitable for problems like NER (Hofmann and Joachims, 2004). We have used the SVM^{hmm}[7] tool in our experiment. SVM^{hmm} provides only training and testing files. The training file has the same formatting as the testing file. The file is divided into columns and rows. Each word has a row. Each feature has a column. All features should be binary features.

Experiments and Evaluation

In this research, the Conlleval tool[8] is applied to evaluate our work; 7-fold cross-validation has been utilized. For each fold, the CONLL evaluation standard metrics of precision, recall and F-measure is applied. The following equations can express precision, recall and F-measure:

$$Precision = \left(\frac{true\ positive}{true\ positive + false\ positive} \right) \quad (1)$$

$$Recall = \left(\frac{true\ positive}{true\ positive + false\ negative} \right) \quad (2)$$

$$F - Measure = 2 * \frac{pracision * recall}{pracision + recall} \quad (3)$$

Then, the average precision, recall and F-measure for each tag are calculated.

Discussion

It was expected that the SSVM based Arabic NER system would outperform the CRF based Arabic NER system because of the high recall of SSVM. Other studies on NER tasks show that SSVM based NER task required less time and achieved better performance than the CRF based NER task for clinical entity recognition when using the same features (Tang *et al*., 2013; 2012). But in our proposed system we found that CRF based Arabic NER task outperforms the SSVM based Arabic NER task when using the same features. This may be due to the differences in the nature of data and the complexities of the Arabic language itself.

[7]www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html
[8]https://www.clips.uantwerpen.be/conll2002/ner/bin/conlleval.txt

Table 2: CRF

F-measure	Recall	Precision	All
88.41	86.48	90.67	Location
74.65	67.53	83.91	Organization
81.75	79.03	84.77	Person
82.76	79.38	86.86	Overall

Table 3: SSVM

F-measure	Recall	Precision	All
87.36	85.72	89.22	Location
72.65	67.38	79.00	Organization
79.98	77.91	82.35	Person
81.16	78.54	84.23	Overall

Table 4: Bengiba and Rosso 2010 results

F-measure	Recall	Precision	All
89.74	86.67	93.03	Location
65.76	53.94	84.23	Organization
73.35	67.42	80.41	Person
79.21	72.77	86.9	Overall

Table 5: Abdul-Hamid and Darwish results

F-measure	Recall	Precision	All
88	83	93	Location
73	64	84	Organization
82	75	90	Person
81	74	89	Overall

Our proposed system was trained and tested on ANERCORP. The results of applying the above all features on both CRF and SSVM are shown in Table 2 and 3. As demonstrated in Table 2 and 3, the CRF Approach outperforms the SSVM Approach since the CRF tends to model the ORG, LOC and PER better than the SSVM with the same features. The CRF approach precision, recall and F-measure are higher than the SSVM approach precision, recall and F-measure by 2.6%, 0.7% and 1.5%.

On the other hand, if we compared the SSVM approach with Benajiba (Table 4) and Abdul-Hamid (Table 5) approaches (Abdul-Hamid and Darwish, 2010; Benajiba and Rosso, 2008), we noticed that our SSVM approach outperforms in the recall both Benajiba approach by more than 5% and Abdul-Hamid approach by more than 4%. But, our SSVM approach precision is less than Benajiba and Abdul-Hamid approaches. So our SSVM F-measure approach may be equal to the F-measure of Abdul-Hamid approach but it is higher than the F-measure of Benajiba by about 2%.

Conclusion and Future Work

In this research, we have proposed the machine learning-based system the NER task where the CRF and SSVM were utilized as machine learning classifiers. We investigated the FARASA segmenter tool usage in the preprocessing phase which enabled us to solve some

orthographic and morphological complexities of Arabic. Also, we investigated the effect of using a bag of POS, person, location, organization gazetteers, pre-name pre-location whitelist and a bag of words. Our experiments on the corpus showed that the CRF outperforms the SSVM when applying the same features in the task of Arabic NER by 1.5%.

Our future work will concentrate on enhancing the Name entity recognizer system using DL approach. The deep learning succeeded to tackle many challenges in NER task because of its non-linear behavior (Li *et al*., 2018). In addition, the DL doesn't depend on feature engineering or rules like the machine learning and rule-based approach so, this will ease the system maintenance if there are any updates in data. Also, we will try to enlarge our dataset to get better results.

Acknowledgement

I have to thank Dr. Mohammed Abdul-Kareem Modhaffer for the unlimited time that he spent to support me in preparing the training corpus and testing the Stanford part of speech tagger.

Author's Contributions

All authors contributed to this research equally.

Ethics

The manuscript is original and all our material are unpublished.

References

Abdel Monem, A., K. Shaalan, A. Rafea and H. Baraka, 2008. Generating Arabic text in multilingual speech-to-speech machine translation framework. Machine Translat., 22: 205-258. DOI: 10.1007/s10590-009-9054-9

Abdelali, A., K. Darwish, N. Durrani and H. Mubarak, 2016. FARASA: A fast and furious segmenter for Arabic. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Jun. 12-17, ACL, San Diego, California, pp: 11-16. DOI: 10.18653/v1/N16-3003

Abdul-Hamid, A. and K. Darwish, 2010. Simplified feature set for Arabic named entity recognition. Proceedings of the Named Entities Workshop, Jul. 16-16, Uppsala, Sweden, pp: 110-115.

Aboaoga, M. and M.J.A. Aziz, 2013. Arabic person names recognition by using a rule based approach. J. Comput. Sci., 9: 922-927. DOI: 10.3844/jcssp.2013.922.927

Ahmad, A., L.U. Joan and Q. Xu, 2016. Arabic information retrieval: A relevancy assessment survey. Proceedings of the 25th International Conference on Information Systems Development, (ISD' 16), Katowice, Poland, pp: 345-357.

Alkharashi, I., 2009. Person named entity generation and recognition for Arabic language. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools, (LRT '09), Cairo, Egypt, pp: 205-208.

Alqudsi, A., N. Omar and K. Shaker, 2012. Arabic machine translation: A survey. Artificial Intell. Rev., 42: 549-572. DOI: 10.1007/s10462-012-9351-1

Attia, M., Y. Samih and W. Maier, 2018. GHHT at CALCS 2018: Named entity recognition for dialectal Arabic using neural networks. Proceedings of the 3rd Workshop on Computational Approaches to Code-Switching, (CAC '18), Melbourne, Australia, pp: 98-102. DOI: 10.18653/v1/W18-3212

Benajiba, Y. and P. Rosso, 2007. ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. Proceedings of the 3rd Indian International Conference on Artificial Intelligence Workshop on Natural Language-Independent Engineering, (LIE'07), Mumbay, pp: 1814-1823.

Benajiba, Y. and P. Rosso, 2008. Arabic Named entity recognition using conditional random fields. Proceedings of the Workshop on HLT and NLP within the 6th International Conference on Language Resources and Evaluation, (LRE '08), Marrakech, pp: 143-153.

Benajiba, Y., M. Diab and P. Rosso, 2008. Arabic named entity recognition: An SVM-based approach. Proceedings of the Arab International Conference on Information Technology, (CIT '08), Hammamet, pp: 16-18. DOI: 10.3115/1613715.1613755

Benajiba, Y., P. Rosso and J.M. Benedي, 2007. ANERsys: An Arabic Named entity recognition system based on maximum entropy. Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, (ITP' 07), Springer, Berlin, Heidelberg, pp:143-153. DOI: 10.1007/978-3-540-70939-8_13

Benajiba, Y., P. Rosso and M. Diab, 2009. Arabic Named entity recognition: A feature-driven study. IEEE Tran. Audio Speech Language Proc., 17: 926-934. DOI: 10.1109/TASL.2009.2019927

Farber, B., D. Freitag, N. Habash and O. Rambow, 2008. Improving NER in Arabic using a morphological tagger. Proceedings of the 6th International Conference on Language Resources and Evaluation, May 26-Jun. 1, Marrakech, Morocco, pp: 2509-2514.

Gridach, M., 2018. Deep Learning Approach for Arabic Named Entity Recognition. In: Computational Linguistics and Intelligent Text Processing, Gelbukh, A. (Ed.), Springer, ISBN-13: 978-3-319-75476-5, pp: 439-451.

Hofmann, T. and T. Joachims, 2004. Support Vector Machine Learning for Interdependent and Structured Output Spaces. Proceedings of the 21th International Conference on Machine Learning, (CML' 04), Banff, Canada, pp: 1-8. DOI: 10.1145/1015330.1015341

Li, J., A. Sun, J. Han and C. Li, 2018. A survey on deep learning for named entity recognition.

Mohammed, N.F. and N. Omar, 2012. Arabic named entity recognition using artificial neural network. J. Comput. Sci., 8: 1285-1293. DOI: 10.3844/jcssp.2012.1285.1293

Nadeau, D. and S. Sekine, 2007. A survey of named entity recognition and classification. Lingvisticae Investigat., 30: 3-26. DOI: 10.1075/li.30.1.03nad

Oudah, M. and K. Shaalan, 2012. A pipeline Arabic named entity recognition using a hybrid approach. Proceedings of the Coling Organizing Committee, (COC' 12), Mumbai, India, pp: 2159-2176.

Shaalan, K. and H. Raza, 2007. Person name entity recognition for Arabic. Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, (CIR' 07), Prague, Czech Republic, pp: 17-24. DOI: 10.3115/1654576.1654581

Shaalan, K. and H. Raza, 2008. Arabic Named Entity Recognition from Diverse Text Types. In: Advances in Natural Language Processing, Nordstr□m, B. and A. Ranta (Eds.), Springer, Berlin, Heidelberg, pp: 440-451.

Shaalan, K. and H. Raza, 2009. NERA: Named entity recognition for Arabic. J. Am. Soc. Inform Sci. Technol., 60: 1652-1663. DOI: 10.1002/asi.21090

Shaalan, K., 2014. A survey of Arabic named entity recognition and classification. Comput. Linguist., 40: 469-510. DOI: 10.1162/COLI_a_00178

Shaheen, M. and A.M. Ezzeldin, 2014. Arabic question answering: Systems, resources, tools and future trends. Arabian J. Sci. Eng., 39: 4541-4564. DOI: 10.1007/s13369-014-1062-2

Sutton, C. and A. Mccallum, 2007. An Introduction to Conditional Random Fields for Relational Learning. In: Introduction to Statistical Relational Learning, Getoor, L. and B. Taskar (Eds.), MIT Press.

Tang, B., H. Cao, Y. Wu, M Jiang and X. Hua, 2012. Clinical entity recognition using structural support vector machines with rich features. Proceedings of the ACM 6th International Workshop on Data and Text Mining in Biomedical Informatics, (MBI' 12), ACM, New York, pp: 13-20. DOI: 10.1145/2390068.2390073

Tang, B., H. Cao, Y. Wu, M. Jiang and X. Hua, 2013. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. BMC Med. Inform. Dec. Mak.

Tsochantaridis, I. and T. Hofmann, 2005. Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res., 6: 1453-1484.

Zaghouani, W., 2012. RENAR: A rule-based Arabic named entity recognition system. ACM Tran. Asian Language Inform., 11: 1-13. DOI: 10.1145/2090176.2090178