ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/319934973

Statistical Arabic Name Entity Recognition Approaches: A Survey

Article *in* Procedia Computer Science · September 2017

DOI: 10.1016/j.procs.2017.08.288

CITATION
1

READS
149

3 authors:

Wael Etaiwi
Princess Sumaya University for Technology
18 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Arafat Awajan
Princess Sumaya University for Technology
70 PUBLICATIONS **123** CITATIONS
SEE PROFILE

Dima Suleiman
University of Jordan
27 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

NIDS for IoT View project

Secure RFID Access Control System View project

All content following this page was uploaded by Dima Suleiman on 23 September 2017.

The user has requested enhancement of the downloaded file.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 57–64

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)

Statistical Arabic Name Entity Recognition Approaches: A Survey

Wael Etaiwi*, Arafat Awajan and Dima Suleiman

Princess Sumaya University for Technology, Amman. Jordan

Abstract

With the increase of Arabic textual information via internet websites and services, a tools for processing Arabic text information are needed to extract knowledge from them. Name Entity recognition aims to extract name entities such as: person names, locations and organizations from a given text. Name Entity recognition approaches were classified into two main approaches: rule-based approach and statistical approach. Although the literature on Name Entity recognition is quit extensive, few researches to extract Name Entities in the Arabic language could be found. This paper provides a comprehensive survey about statistical approaches of Arabic Name Entity extraction.

© 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.

Keywords: Name Entity Recognition, Machine-learning approaches, NLP, Statistical, Arabic Language, Neural Network, HMM.

1. Introduction

Name Entity (NE) is an expression that refers to proper names such as persons, locations, and organizations. For example: Arafat Awajan is a full professor at Princess Sumaya University for Technology in Jordan, then Arafat Awajan, Princess Sumaya University for Technology, and Jordan would be identified as reference to person, organization, and location, respectively. The task that attempts to locate, extract, and automatically classify named entities into predefined classes or types in open-domain and unstructured texts, such as newspaper articles, was called Name Entity Recognition (NER)[1].

* Corresponding author. Tel.: +962795744288.
E-mail address: w.etaiwi@psut.edu.jo

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.288

1.1. NE applications

1.1.1 Information Retrieval: This is the task of retrieving data or documents according to a search input query, this task requires identifying NEs in the input query and identifying NEs within the search data or documents, in order to retrieve the relevant document. For example: the word الهلال (alhilal) can be recognized as an organization name such as (Saudi football club) or as noun such as (moon). The correct identification of NEs will facilitate retrieving the correct document. A study by[2] has indicated that about 60% of the queries in search engines contain NEs.

1.1.2 Question Answering: The task of giving an answer for a given question is called a Question Answering application. NER can be used to analyze questions that will help in identifying the correct domain and constructing a relevant answers. Moreover, the answer of many questions contains NEs, For Example: the answer of questions that begin with who (من) usually involves persons or locations, and the answer of questions that start with where (أين) usually involves locations.[1]

1.1.3 Machine Translation: The task of automatically translating a given text from one language to another called Machine Translation. In this task, NER systems play a key role in the overall quality of machine learning applications, it is very important in order to determine which part of NE should be meaning-translated, and which part should by transliterated, such as personal names. For example: جامعه الأميرة سمية للتكنولوجيا is translated to Princess Sumaya University for technology, in this example, the word سميه is transliterated to Sumaya, while the other words translated normally.

1.1.4 Navigation Systems: Using digital maps to provide directions and information about nearby places is the main task of navigation systems. In such systems, all places and locations stored in the system database with their geographical coordinates are NEs.

1.2. Arabic language aspects and challenges

"Arabic is a language of rich morphology and complex syntax"[3]. It is classified into three main types: Classical Arabic; which is the language of Islam that used for over 1500 years. Modern Standard Arabic; which is one of the six official languages of United Nations, and most of Arabic NLP researches are focused on Colloquial Arabic; which is the spoken Arabic language. It is irregular and differs among countries and regions.

The task of Arabic NER is challenging due to the following Arabic language features:

- Lack of capitalization: Capitalization feature is not existing in Arabic language script, unlike other languages such as English, in which capital letter used to recognize NEs. The absence of this feature makes Arabic NER more difficult by the fact that most of Arabic NEs are indistinguishable from common nouns and adjectives. For example: the Arabic location word الزرقاء (city in Jordan) can be used as an adjective (refers to something with blue color). This type of ambiguity commonly resolved by analyzing the context surrounding the NE.[1]
- Complicated morphology: Each word in Arabic language may consist of one or more prefixes, a stem or root, and one or more suffixes, resulting in a complicated morphology. Moreover, clitics may be attached to the NE including conjunctions, prepositions, or a combination of them.
- Optional Short Vowels: Arabic words contains diacritics (small marks placed above or under an Arabic letter) representing most of vowels that give different meaning to the same lexical form. Most Arabic text appears in letters, magazines, or other media are undiacritized for simplification, which led to lexical types of ambiguity. For example: the word مؤسسة could be recognized as location NE when it is diacritized as مُؤَسَّسَة (foundation or corporation) or as a person NE when it is diacritized as مُؤَسِسَة (a founder)[1].
- Ambiguity in Named Entities: two or more NEs could be ambiguous and refers to many different NEs types. For example: the word أمنيه could be considered as person NE, or could be considered as organization NE (a telecommunication company in Jordan).
- Lack of Uniformity in Writing Styles: this ambiguity occurs when transliterating a NE from other language to Arabic language, this happens because Arabic has more speech sounds than other languages. For example: transliterating English NE such as Gallery Mall (a location in Jordan) into Arabic NE could produce many variants such as: جاليري مول ، غاليري مول.
- Lack of Resources: Arabic language has limited number of available resources to be used in NER systems. Corpora (tagged documents) and gazetteers (list of types NE) are used to implement and test the performance of Arabic NER systems. Researchers in Arabic NLP relays on their own human annotated corpora, some of them are published and become available freely to others, whereas others are available under paid licensed agreements.[1]

2. Extraction approaches

Many Arabic NER approaches proposed using statistical NER (also called Machine-Learning Approach), this section briefly surveys the statistical Arabic NER approaches.

2.1 Hidden Markov Model: Hidden Markov Model is a probabilistic Markov function process that computes the probability distribution over possible labels on a sequence of units. The observation of output can be generated according to associated probability distribution of input sequence of units, in which the probability of the next unit depends only on the previous k units.

HMM could be used to disambiguate NEs as proposed by[4], the proposed approach consists of two main modules: the first module model all states and observations found on the dataset; while the second module disambiguate NEs by decoding an input observation sequence into the most probable sequence of states using HMM, the authors used DBpedia ontology to extract dataset from Wikipedia dump taken in February 2012. The authors evaluated their work using Micro and macro accuracy evaluation measures, and their results exceed some other approaches such as: Cucerzan and Kulkarni, and did not exceed other approaches such as: Hoffart.

Hidden Markov Model (HMM) needs annotated corpus in order to provide good results, the authors of [5] used a corpus of 200,000 words in training phase and then created their own corpus extracted from Gigaword, the corpus was collected from different resources such as Aljazeera, sport news Alhayat, international and economic news. In order to determine the class of a certain word; features, number of occurrences and class of previous word were considered, this is called a bigram model that predicts the current word using the features of single previous word. The precision of the proposed system was 77% while recall was 73%. The model mainly depended on the stem of the word.

A stream of three words was taken to recognize Name Entity: a target word and two previous words, this called trigram or second order HMM which was used by[6]. There are three classes of Name Entity that were considered in their research: person, organization and location. The precision and recall of proposed method was 83% and 82% respectively. The testing corpus was collected from different resources, and consists of about 20000 words from Aljazeera, Alhayat and Alarabiya and other news sites.

2.2 Conditional Random Field (CRF): It is a statistical modeling approach applied in machine learning used for structured prediction, segmentation and labeling sequence of data. It is a generalization of HMM. It aims to find y that maximize $p(y|x)$ for the sequence x; where y is the labeled sequence, $p(y|x)$ is the probability of y given x, considering the previous and the succeeding elements.

In order to solve problems related to recognizing Arabic Name Entity, a group of researchers[7] integrated two machine learning techniques into one model: bootstrapping and Conditional Random Field (CRF) classifier. Ten NE classes were used, includes: Cell Phone, Currency, Person, Location, Device, Organization, Job, Car, Date, and Time classes. The experimental tests conducted using 6-fold cross-validation. In order to test and train CFR, ANERcorp dataset was used. The semantic field features and patterns were effective since CRF was boosted.

CRF was used in[8], in which the input of CRF were internal and external features. Internal features were related to linguistics such as words, POS, Gazetteers features, Gram character features and indicator features, while external features were represented by taking three named entities and find the semantic relationships between them such as Class, Instance and Relation features in order to enhance accuracy. ALETEC corpus was created using Arabic Language Technology center; this corpus is not free and it consists of 288737 tokens. Moreover, Gazetteers consists of Arabic Name Entity Recognizer (ANER), ANERcorp dataset and Wikipedia gazetteers were used to train and test the proposed approach. F-measure result of applying ANERcrop dataset was 87.86%, and 84.72% for ALTEC dataset.

The authors of[9] presented a set of features for Arabic NER, they trained CRF sequence labeling model on features that used n-gram of leading and trailing letters in word and n-gram words, the presented features set improved results of NER.

CRF could be combined with another approaches to improve the Arabic NER, such as the approach proposed by[7], in which, the authors combined CRF with another approach called "bootstrapping semi-supervised pattern recognition", the authors Used 6-fold cross-validation experimental tests in order to evaluate their approach, and the experiment results showed that the proposed approach overcame CRF in term of precision, recall and F-measure.

A CRF-base NER systems proposed by[10], the proposed system, called Noor, extracts person names from Arabic religious texts. A tagged corpus from ancient Arabic religious text were created, called NoorCorp, it consists of

Prophet Mohammed's Hadith, historical and jurisprudence books. The proposed system consists of four main steps, as illustrated in figure 1: first step, the preprocessing step, which included tokenization and transliteration. The second one was part of speech tagging which used AMIRA software. The third step was proper name candidate injection, in which POS tags enriched to improve results. And the final step was CRF training.

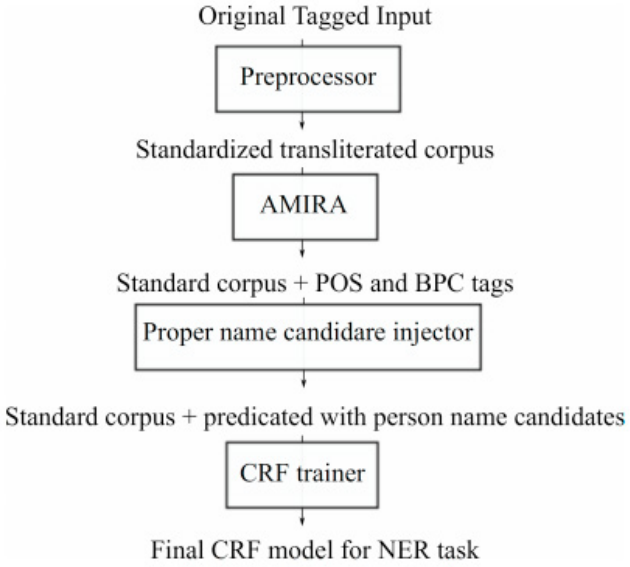

Figure 1: Structure of Noor NER system[10]

2.3 Naïve Bayes (NB): is a probabilistic multiclass classification algorithm based on the Bayes theorem[11] it aims to compute the conditional probability distribution of each feature. NB assumes that features are independent of each other and can be represented by a vector v= (x_1, x_2, x_3, … x_n), where x_1, x_2, x_3, …, x_n are n features. In order to classify a vector into class, the conditional probability will be used, such that, in order to classify vector v to class C, then the value of conditional probability will be equal to the multiplication of conditional probability of each feature in the vector being in the class C

$$P(C_k|x) = P(C_k|x_1)\, P(C_k|x_2)\, P(C_k|x_3) \ldots P(C_k|x_n) \quad (1)$$

Each individual probability of a feature to be in a given class C (in the above equation) is calculated as shown in formula 2.

$$P(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (2)$$

In multinomial NB, p(x|Ck) = count(x,Ck)/count(Ck). When a feature is not included in the training set, the probability of it p(x) will be zero, which will affect the other probabilities. To solve this issue, Laplace Smoothing technique is used by adding λ parameter to equation (2) as can be seen in equation (3):

$$p(x|C_k) = \frac{\text{count}(x,C_k)+\lambda}{\text{count}(C_k)+\lambda|X|} \quad (3)$$

A novel model for NER and classification based on multinomial naïve Bayes classifier was proposed by[12], the authors used term frequency, inverse document frequency and fit them to a tf-idf-vectorizer to extract features from training dataset. The proposed approach worked with Precision, Recall and F1-measure of 83%, 79% and 81% respectively.

Another approach based on Bayesian Belief Network (BBN), called NAMERAMA, proposed by[13], the proposed approach used in medical domain to extract disease names, symptoms, treatment methods, and diagnosis methods from Arabic text. The system consist of four main steps: the first step was the preprocessing step; in which text tokenization occurs and POS tagging process carried out using AMIRA tool. The second step was data analysis step; in which the optimal features set was extracted after the analysis of data, frequency, collocation and

concordance. The third step was feature extraction; in which POS tagging was supported using many different kinds of features (such as: Gazetteer features, lexical markers features, pattern features) and a list of stop words. Finally, the classification step which was made using (BBN). The experimental results show that the overall F-measure was 71.05% after applying the proposed approach on dataset obtained from the King Abdullah Bin Abdulaziz Arabic Health Encyclopedia (KAAHE) website.

A Comparison research was proposed by[14] to study the impact of feature representation, feature set and statistical modeling. The authors used four different classification algorithms; NB, Multinomial NB, SVM and Stochastic Gradient Descent to map Wikipedia articles into predefined set of NEs classes. The authors concluded that using language-dependent features did not lead to better performance, however the classification could be corrected by selecting the correct feature sets and representation.

2.4 Neural Networks: Neural Network Classifier is a classification approach[15] based on the principles of artificial neural networks (NN). The backpropagation gradient neural network (the most frequently used network) consists of three or more neuron (a neural network unit) layers: one input layer, one output layer and at least one hidden layer, as shown in figure 2.

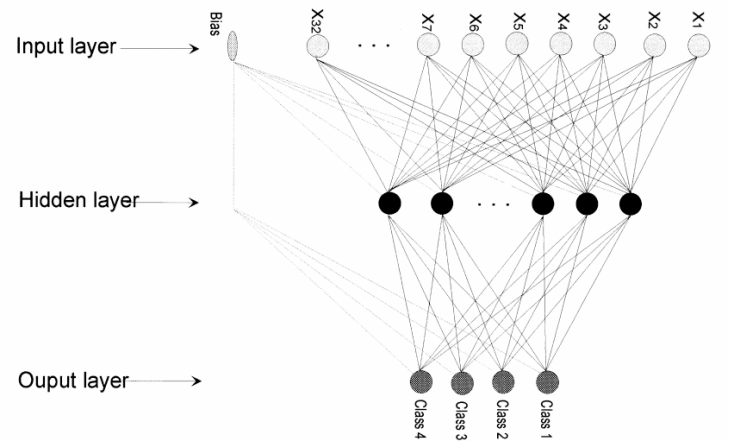

Figure 2: backpropagation gradient neural network[15]

Each neuron of the neural network consists of a set of inputs, weights and outputs, Figure 3.

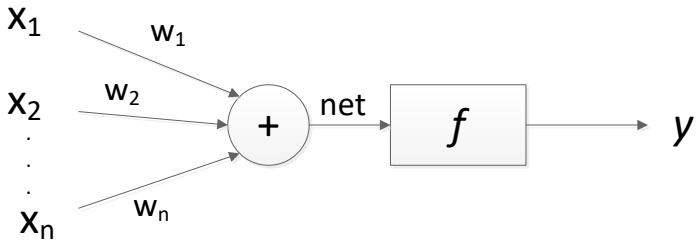

Figure 3: Model of artificial neuron

Where (x_1, x_2,…, x_n) are neuron input, (w_1, w_2, …, w_n) are input's weights, f: is the activation function, y is the output, and net is a weighted sum calculates as in equation (4):

$$net = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \quad (4)$$

An Arabic NER system based on neural networks proposed by[16], the proposed system consists of three main steps: the first step was preprocessing step which aims to cleans the collected data and to prepare it to be used in the system, in addition, this step also included tokenization, removing foreign words and punctuation marks. The second step was transliteration (converting Arabic letters to Roman alphabets), and the third step was to apply neural network to classify the collected data. The authors compared their system with decision tree approach, and concluded that neural networks overcome decision tree in term of accuracy with 92%.

2.5 Entropy: It is a general technique for estimating probability distributions from data, it prefers the uniformity or maximum entropy if no data is observed, and in this case, the distribution should be as uniform as possible, on other words, it represents the average amount of information we get to expect the output of some input value x:

$$H(x)= \sum_i pi \log \frac{1}{pi} \quad (5)$$

Where p_i is discrete probability distribution on the countable set $\{x_1, x_2, \ldots, x_i\}$.

An Arabic NE recognition approach (ANERsys) based on Maximum Entropy (ME) was presented by[17], the proposed approach consists of two main steps to determine the boundaries of ANEs: in the first step, ME was used to compute weights for each feature in the training phase, which will be used later in the testing phase. The second step was related to using Mona Diab's Arabic POS tagger and considers only noun phrases items (as illustrated in figure 4). The authors build their own corpora (ANERcorp) and gazetters (ANERgazet) to train, test and evaluate their approach. The results showed that the proposed approach was 7 points accurate more than Siraj (a commercial NE recognition system) and F-measure was 55.23% using person names, location, organization, and "miscellaneous".

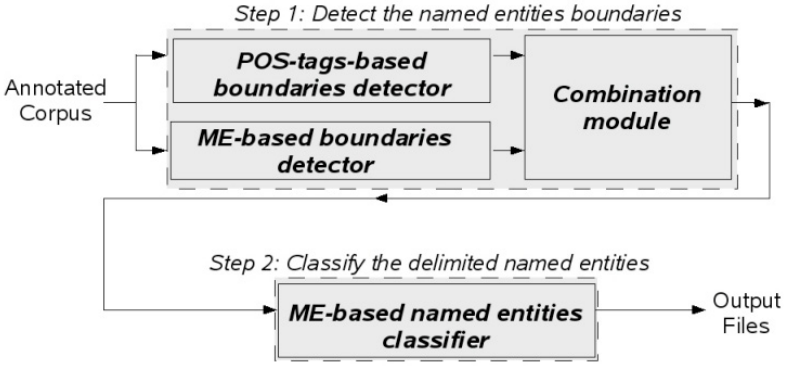

Figure 4: ANERsys Architecture[17]

2.6 SVM: SVM is a binary machine learning classification algorithm that classifies all items to only one out of two classes. The main task of SVM classification model is to find the maximum margin hyperplane that classify the group of features vectors among two classes (0 or 1), which is the maximum distance between the hyperplane and the nearest x from either classes, figure 5 illustrates the simplest SVM[14].

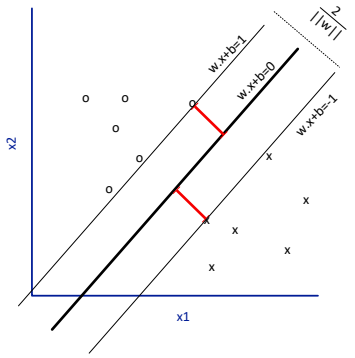

Figure 5: Linear Support Vector Machine

Many researchers investigated the impact of feature sets used to extract Arabic NEs using SVM classifier[18,19], they explored morphological, lexical and contextual features and nine different datasets. These features was a combination of Arabic language dependent features and other language independent features. The authors concluded that using a combination of all features achieved the highest performance measure, F1=82.71.

An ANE hybrid approach proposed by[20] , that combined rule-based approach with Machine-learning approach into one model in order to extract Arabic NE, the authors used three different statistical approaches: decision tree, SVM, and Logistic Regression, to evaluate their hybrid model over three different datasets. The experimental results showed that the proposed hybrid approach leads to highest performance in term of F-measure, with 94.4% for Person, 90.1% for Location, and 88.2% for Organization.

In order to extract the relationship between Arabic NEs; a supervised process, called RelANE, proposed by[21], the authors studied various features of the word in the sentence and applied several supervised algorithms (such as: SVM, NB, Decision tree) in order to predict which term (feature) can explicit a relationship between NEs. The authors concluded that SVM and Adaboost techniques were more efficient for NE relation extraction task.

3. Conclusion

This paper presented a comprehensive survey of statistical-based Arabic NER. Arabic NER was categorized into six main approaches: CRF, NB, HMM, ME, SVM and Neural network. After analyzing the surveyed articles, it is clear that the most common NE classes are: persons, locations and organizations. Moreover, there is a lack of Arabic resources to be used in Arabic NER tasks, where ANERcorp dataset is the most common dataset used to extract NE. Most of the surveyed articles used F-measure as an evaluation metric of the proposed approaches. As listed in table 1, SVM and CRF are the most common approaches in which the researchers focused on and investigated. The main task researched in the NE area was the recognition task, as illustrated in table 2, and the NE disambiguation was the least topic investigated by Arabic NE researchers. Table 4 shows that most of researchers who proposed researches to study NER were used many different datasets to evaluate their works rather than using only one dataset, because the Arabic resources are limited and not universally adapted. In table 5, we note that precision, recall and F-measure were the most common evaluation metrics used by researchers, who mostly investigated the recognition of more than three different NE classes.

Many enhancements could be done in the field of Arabic NER, such as using domain-oriented data sources, and social media data sources, especially from different Arabic dialects. Another important addition is to provide enterprise tools that exploit the increase of using social media and Arabic documents to extract name entities.

Table 1: Number of surveyed articles by extraction approach

Extraction Approach	HMM	CRF	NB	NN	ME	SVM	Others
Count	3	4	3	1	1	4	3

Table 2: Number of surveyed articles by main task

Main Task	Feature Selection	NER	NE Disambiguation
Count	3	13	1

Table 3: Number of surveyed articles by dataset used

Dataset Name	“Multiple Datasets”	Own datasets	AIDA dataset	NoorCorp	Wikipedia	ANERcorp dataset
Count	5	5	1	1	1	3

Table 4: Number of surveyed articles by evaluation metrics

Evaluation Metrics	Precision, Recall and F-measure	F-measure	Accuracy
Count	12	3	1

Table 5: Number of surveyed articles by total number of classified NE classes

Total number of classified NE classes	One or two class	Three classes (persons, locations and organizations	More than three classes	Not determined
Count	2	4	9	1

References

1. Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. Comput. Linguist. 40, 2 (June 2014), 469–510. DOI:https://doi.org/10.1162/COLI_a_00178
2. Dayong WU, Yu ZHANG, and Ting LIU. 2011. Analysis of Named Entity Queries in Web Search Logs. J. Comput. Inf. Syst. 7, 16 (2011), 5837–5844.
3. Imad A. Al-Sughaiyer and Ibrahim A. Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. J. Am. Soc. Inf. Sci. Technol. 55, 3 (February 2004), 189–213. DOI:https://doi.org/10.1002/asi.10368
4. Ayman Alhelbawy and Robert Gaizauskas. 2013. Named entity disambiguation using hmms. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03. IEEE Computer Society, 159–162.
5. Fadl Dahan, Ameur Touir, and Hassan Mathkour. 2015. First Order Hidden Markov Model for Automatic Arabic Name Entity Recognition. Int. J. Comput. Appl. 123, 7 (2015).
6. Fadl Dahan, Ameur Touir, and Hassan Mathkour. 2015a. Arabic Name Entity Recognition Using Second Order Hidden Markov Model. (2015).
7. AbdelRahman, Samir, Mohamed Elarnaoty, Marwa Magdy, and Aly Fahmy. 2010. Integrated machine learning techniques for Arabic named entity recognition. Int. J. Comput. Sci. Issues 7, 4 (July 2010).
8. Hamzah A. Alsayadi and Abeer M. ElKorany. 2016. Integrating Semantic Features for Enhancing Arabic Named Entity Recognition. Int. J. Adv. Comput. Sci. Appl. 7, 3 (2016), 128–136.
9. Ahmed Abdul-Hamid and Kareem Darwish. 2010. Simplified Feature Set for Arabic Named Entity Recognition. In Proceedings of the 2010 Named Entities Workshop. NEWS '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 110–115.
10. Majidi Bidhend, Behrouz Minaei-Bidgoli, and Hosein Jouzi. 2012. Extracting person names from ancient Islamic Arabic texts. In Proceedings of Language Resources and Evaluation for Religious Texts (LRE-Rel) Workshop Programme, Eight International Conference on Language Resources and Evaluation (LREC 2012). 1–6.
11. Moshe Ben-Bassat, Karin L. Klove, and Max H. Weil. 1980. Sensitivity analysis in Bayesian classification models: Multiplicative deviations. IEEE Trans. Pattern Anal. Mach. Intell. , 3 (1980), 261–266.
12. S. Amarappa and S.V. Sathyanarayana. 2015. Kannada named entity recognition and classification (nerc) based on multinomial na"ive bayes (mnb) classifier. Int. J. Nat. Lang. Comput. 4, 4 (August 2015), 39–52. DOI:https://doi.org/10.5121/ijnlc.2015.4404
13. Saad Alanazi, Bernadette SHARP, and Clare STANIER. 2015. A Named Entity Recognition System Applied to Arabic Text in the Medical Domain. IJCSI Int. J. Comput. Sci. Issues 12, 3 (2015).
14. Anon. The Nature of Statistical Learning Theory | Vladimir Vapnik | Springer,
15. David Reby, Sovan Lek, Ioannis Dimopoulos, Jean Joachim, Jacques Lauga, and Stéphane Aulagnier. 1997. Artificial neural networks as a classification method in the behavioural sciences. Behav. Processes 40, 1 (April 1997), 35–43. DOI:https://doi.org/10.1016/S0376-6357(96)00766-8
16. Naji F. Mohammed and Nazlia Omar. 2012. Arabic Named Entity Recognition Using Artificial Neural Network. J. Comput. Sci. 8, 8 (July 2012), 1285–1293. DOI:https://doi.org/10.3844/jcssp.2012.1285.1293
17. Y. Benajiba and P. Rosso. 2007. ANERsys 2.0 : Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information. In India, Pune.
18. Anon. 2009. IAJIT - Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognitio. (2009). Retrieved February 12, 2017 from http://ccis2k.org/iajit/?option=com_content&task=view&id=428
19. Yassine Benajiba, Mona Diab, Paolo Rosso, and others. 2008. Arabic named entity recognition: An svm-based approach. In Proceedings of 2008 Arab International Conference on Information Technology (ACIT). 16–18.
20. Mai Oudah and Khaled F. Shaalan. 2012. A Pipeline Arabic Named Entity Recognition using a Hybrid Approach. In Coling. 2159–2176.
21. Ines Boujelben, Salma Jamoussi, and Abdelmajid Ben Hamadou. 2014. RelANE: Discovering Relations between Arabic Named Entities. In Petr Sojka, Aleš Horák, Ivan Kopeček, & Karel Pala, eds. Text, Speech and Dialogue. Lecture Notes in Computer Science. Springer International Publishing, 233–239.

View publication stats