# Integrating Semantic Features for Enhancing Arabic Named Entity Recognition

Hamzah A. Alsayadi
Ibb University, Yemen
MSc student in Faculty of Computers and
Information, Cairo University
Cairo, Egypt

Abeer M. ElKorany
Dept. of Computer Science, Faculty of Computers and
Information, Cairo University
Cairo, Egypt

*Abstract*—**Named Entity Recognition (NER) is currently an essential research area that supports many tasks in NLP. Its goal is to find a solution to boost accurately the named entities identification. This paper presents an integrated semantic-based Machine learning (ML) model for Arabic Named Entity Recognition (ANER) problem. The basic idea of that model is to combine several linguistic features and to utilize syntactic dependencies to infer semantic relations between named entities. The proposed model focused on recognizing three types of named entities: person, organization and location. Accordingly, it combines internal features that represented linguistic features as well as external features that represent the semantic of relations between the three named entities to enhance the accuracy of recognizing them using external knowledge source such as Arabic WordNet ontology (ANW). We introduced both features to CRF classifier, which are effective for ANER. Experimental results show that this approach can achieve an overall F-measure around 87.86% and 84.72% for ANERCorp and ALTEC datasets respectively.**

***Keywords—Arabic Named Entity Recognition (ANER); Conditional Random Fields (CRF); Domain Ontology; Semantic Relation Feature (SRF); Arabic WordNet ontology (ANW)***

## I. Introduction

Named Entity Recognition (NER) was introduced in 1990 at the Message Understanding Conferences (MUC-6) 1 [1]. NER is one task of an information extraction to classify proper names from raw texts into types of names [1]. Three major tasks of NER were covered: (person, location, and organization) called ENAMEX, (temporal expressions) called TIMEX, and (some numerical expressions such as monetary amounts and other types of units) called NUMEX [1, 5]. There are other Named Entities (NEs) were defined by NER such as biology domain (like gene, DNA, and RNA NEs), Behavioral Health like (healthy food), and biomedical like (diseases NE) [10 - 12]. In this paper, we deal only with ENAMEX. The goal of Named Entity Recognition (NER) task is the enhancing the accuracy concerning the named entities recognition and extraction [3]. NER task is important for many natural language processing applications such as Search results clustering, Machine Translation, Navigation Systems, enhancing Information Retrieval, and Improving results in Question Answering [1,4].

Due to the political and economic importance of the Arabic language, in the last decades, the NLP researchers started to get interest in research fields in the Arabic language such as Arabic Named Entity Recognition (ANER) [5]. The Arabic language has a rich vocabulary, morphology, and syntax; also, it has a complex morphology [1, 6]. The Arabic language has three styles, Classical Arabic (CA), Modern Standard Arabic (MAS) and Colloquial/Dialects Arabic (DA) [5]. In this work, MSA are dealt. There are challenges for ANER such as high morphological ambiguity, complexity and common noun/words ambiguities. The researchers in ANER tried to collect appropriate data to include all possible language cases having these characteristics and peculiarities such as ANERCorp [2] and ALTEC [3] datasets. Also, they developed tools for that data such as MADA[4], Stanford POS Tagger[5], and AMIRA[6].

The ANER researcher developed system depends on two approaches Ruled Based approach [7, 8, 9, 13, 14] or Machine learning (ML) approach [15 - 18]. The systems were built using Ruled-Based approach, which depends on linguistic rules for recognizing NEs. These rules are usually regular expressions or finite-state transducers. The advantage of the rule-based NER systems is that they are depend on the core of linguistic knowledge. However, any update or maintenance required for these systems is time-consuming and labor-intensive; also, it requires full knowledge of the language [1]. ML approach is to learn NE tagging decisions from annotated texts. The most common approach that is used in ML for NER is Supervised Learning (SL). It represents the NER problem as a classification task that distinguishes between different types of names entities. The advantage of ML-based NER systems are the ease of maintenance, modifications, and adaptation over time. According to [4, 15], CRF and SVM had been proven as the best techniques for ANER. The researchers in [4] proved that CRF is better than other techniques while in [15] they did not state whether CRF is better than SVM or not in Arabic NER.

In this paper, an integrated semantic-based ML is applied for ANER. CRF is used CRF as the classification engine for recognizing three named entity (NE) classes; person, location,

[1] http://cs.nyu.edu/cs/faculty/grishman/muc6.html

[2] http://users.dsic.upv.es/~ybenajiba/

[3] http://www.altec-center.org/Repository_65.html

[4] http://www1.ccls.columbia.edu/~cadim/MADA

[5] http://www-nlp.stanford.edu/software/tagger.shtml

[6] http://www.cs.columbia.edu/~mdiab/

and organization names. This integration is new for ANER, to the best of our knowledge, since it has not been utilized in ANER. This model combines internal features that represented linguistic features as well as external features to represent the semantic of Arabic language. Arabic semantics relate primarily to the semantic correlates of morphological patterns. This correlation is extracted from two different resources each represented the relationships that could exist between the extracted named entities such as ontology (Arabic wordNet ontology) in form of classes, instances, and relations between entity classes, and feed it to CRF classifier as a set of features to enhance the classification process. These semantic features are efficient for ANER, and over performed other CRF that used less number of features with better accuracy.

The paper is organized as follows: Section 2 illustrates some of Arabic Language challenges; Section 3 gives an overview of the domain ontology; Section 4 explain some of the previous systems as related work; Section 5 discusses the components of architecture system; Section 6 show the data that used in this system; Section 7 contains an Evaluation Criteria; Section 8 includes an experimental; Section 9 submits results and discussion Finally; Section 10 present the conclusion and future work.

## II. Arabic Language Challenges

We focus on Arabic NER that has several challenges and characteristics:

*1) Lack of capital letters:* A named entity in Latin languages is usually distinguished by a capital letter at the word beginning. However, Arabic lack the capital letter, so the detection of NE in text based on the letters case more difficult. The lexical triggers used to overcome this problem, which has used that are derived from analyzing the surrounding context of NEs while some others researchers have used the English translation of the NE what is known as the glossing feature produced by the MADA tool [2, 3, 22].

*2) Complex Morphology: the Arabic language has a complex morphology due to the agglutinative nature of language. Agglutinative morphemes have three types: stems, affixes, and clitics. The stem is the primitive form of the word. Affix letters are usually added to the stem, which has three types: prefixes attached to beginning of the stem, suffixes attached to end of the stem, and circumfixes that surround the stem. Clitics are also added to the stem after affixes. Clitics are either proclitic that come before the word or enclitics that come after the word. The conjunction "و" (waw, and) and object pronoun "هن" (hn) are examples of proclitic and enclitics, respectively. A more general example is the word "وسيدرسونها" (and-they-will-study-it) [1].*

*3) Ambiguity: Arabic text has the different meaning for one word (Ambiguity). For example (رجب /Ragab) in Arabic may be used as a person name, and month. The word diacritization is important factor for word meaning, for example, (قطر) which if it is diacritized as قَطر it means country Qatar but if it is diacritized قُطر it means Diameter or territory [1, 2, 22].*

*4) Arabic is a high inflectional language;* often a single word has more than one affix such that it may be expressed as a combination of prefix(s), lemma, and suffix(s) as Word = prefix (es) + lemma + suffix (es). The prefixes are prepositions, conjunctions, or articles. The suffixes are generally personal/possessive or objects anaphora. For example, the Arabic word "وبعروبتنا" is interpreted in English as "and with our Arabism"[2, 3].

*5) Writing Styles Arabic (Spelling variants) has a high level of typographic forms and ambiguity spelling:* An NE can be writing in a many of ways. This multiplicity arises from both different ways of writing the Arabic writers and ambiguous form of transcription schemes. There is no fixed standardization for writing the word like English. For example, the word ' جرام ', jrAm1, 'Gram', can also be written as غرام' ', grAm, with the same meaning, also the word جوجل / Google can be written as غوغل, other example the word سوريا / Syria can be written as سورية [1, 2, 22].

*6)* Systematic Spelling Mistakes Typographical errors were frequently made by Arabic writers according to certain characters. For example, الإسلامي/ The Islamic with (إ) can be written الاسلامي with (ا), and العربية/ the Arabia with (ة) can be written as العربيه with (ه) [1].

*7)* Some foreign persons' names when it was transliterated into Arabic could be identified as pronouns or prepositions such as [Ho, Anna, Ann, and, Lee] their different pronouns or prepositions are [He, I, That, Mine] [22].

*8) Lack of Resources:* Large collections of tagged documents (corpora), gazetteers (predefined lists of typed NEs), and NLP tools, are either rare or not free. This challenge makes collecting and analyzing the data is time-consuming particularly if the NER technique depends on such resources [23, 24]. There are few corpora such as the free ANERCorp, the commercial ACE (2003 – 2005)[7] and ALTEC.

## III. Domain Ontology

Domain ontology, as a formal specification of a shared conceptualization, defines the Knowledge base of the concept, attributes, relations between concepts and properties even relations between properties. Moreover, it describes axioms, individuals and relations between them, and provides sharing knowledge. It has a better capacity of semantic Interpretation. The specific domain wordNet ontology was used.

WordNet is a large lexical database for English by Princeton. It contains information about 147,278 words divided into nouns, adjectives, verbs, and adverbs. Then the words are expand divided into 206,941 senses, with an average of 1.4 senses per word. These senses are grouped by synonymy into 117,659 unorganized sets called synsets. Words in the same synset refer to the same concept and are may be used in many contexts mutually. There are also some semantic relations between the synsets such as the hyponym and hypernym relations. Thus, WordNet is sometimes considered as a lexical ontology. WordNet has realized great

[7] https://catalog.ldc.upenn.edu/LDC2006T06

success and became the dominant English lexicon in NLP applications [25].

The Arabic WordNet (AWN) is a wordNet for Modern Standard Arabic (MSA), it was built depend on the design and contents of WordNet (WN) [26]. Thus, the AWN synsets are linked to WN synsets directly. Up to now, AWN consists of 13,808 non-diacratized Arabic words divided into 23,481 senses that form 11,269 Arabic synsets. All of the synsets are connected to the corresponding English synsets in WordNet. The low AWN/WN ratios suggest low coverage of Arabic words in AWN, which can be easily verified as some of the commonly used Arabic words are missing, such as the noun 'بطولة' (championship), the verb 'تقابل' (meet), and the adjective 'أفريقي' (African) [25].

## IV. Related Work

There are many researchers that applied ML-based for ANER in order to learn NE tagging decisions from annotated texts. There are techniques utilized for ANER are Support Vector Machines (SVM), Conditional Random Fields (CRF), Maximum Entropy (ME), artificial neural network (ANN), and Decision Trees. Each technique need features for NEs identification such as gazetteers features, POS tags and morphology features. Some researchers used CRF [4, 16, 19] depend on their features while in [20] used SVM, other researchers used ME [21]. Finally, in [3], the authors used ANN. Other researchers used CRF and SVM in [15, 17].

In [16] authors introduced a system for improving NER on microblogs, it contain three methods: (1) using large gazetteers from Wikipedia, (2) domain adaptation, and (3) a two-pass semi-supervised method. They used CRF classifier (CRF++[8]). They tagged new training set from Tweeter. The evaluation of system depended on ANERcorp and new training set (which they tagged). They compared their system with other systems; this system shows an improvement of 35.3 F-measure points over other systems.

The authors [17] proposed a simplified feature set system. This system dealt with only some of the Arabic morphological and Arabic orthographic complexities features. They used CRF classifier to identification NEs. They evaluated their work using ANERcorp and ACE2005 dataset. The result of the system proved the effectiveness of simplified feature set for ANER.

In [18] the authors developed system using Cross-lingual Features. They used three Arabic and English Wikipedia cross-language links Cross-lingual Capitalization, Transliteration Mining and using DBpedia. The work used CRF, was evaluated using ANERcorp dataset for training and testing, also used NEWS Test Set and TWEETS Test Set. In this work, the authors showed how cross-lingual Features enhanced ANER.

Semi-supervised learning was used in [27] to develop ASemiNER, a semisupervised algorithm for identifying Named Entities (NEs). The system including Pattern Induction, Instance Extraction, and Instance Ranking/Selection Methodology. ASemiNER does not require any annotated corpora or any gazetteers, but it was compared with ANERcorp and ACE2005 dataset.

In [15] the authors investigated a large of features sets in order to get the optimal feature sets. Multiple classifiers were used in this system to recognize NEs SVM and CRF. They ACE 2003, ACE 2004 and ACE 2005 data sets. The multi-classifier and language independent features outperform the system in [20] that used one classifier by 0.79 F-measure.

In [3] the authors developed the system using Artificial Neural Networks (ANN) approach. The system including three processes preprocessing of the data, transforming the Arabic letters to Text Romanization and applying the ANN classifier to the text. They used ANERcorp dataset and data collected manually from diverse web sources for evaluation. The authors compared the result of the system between decision trees and ANN approaches. The result demonstrated the ANN achieves higher results than that to get from the decision trees approach.

In [2], the authors presented a solution for ANER; this solution is an integration between two machine learning approaches, bootstrapping semi-supervised pattern recognition and CRF classifier as a supervised technique. This system including three modules CRF classifier, pattern recognizer, and the matcher module. In this solution is used RDI-ArabSemanticDB tool and RDIArabMorpho-POS tagger. They used ANERcorp (person, location, and organization) and crawled from the web other NEs for the system evaluation. This integration is designed to increase the CRF F-measure.

In [4], the authors developed their previous works (ANERsys) in [21] to enhance the accuracy of this system using Conditional Random Fields. The performance results achieved on ANERcorp dataset. They identify Person, Location, and Organization classes with F-measure of 73.34%, 89.74%, 65.76%, and 61.47% respectively. They prove that CRF achieves the result better than their previous work in [21] by 12 points in the F-measure average of all classes.

In [22] the authors developed their system in [28]. They used integrated approach: a) name dictionaries and b) name clusters with a statistical model based on extracting patterns that indicate the existence of person's names. They used list of names more than list in [28] was named full_names_19000_list. The result in this system is better than their previous work [20] by 4.09 F-measure.

In [29] the authors used semi-supervised and distance learning techniques, then Bayesian Classifier Combination (BCC) to recognize Arabic NEs. They built Wikipedia-derived corpus (WDC). They used the dataset that built and ANERcorp dataset for evaluation. Previous Systems perform better than this work.

For the best of our knowledge, there is no work used semantic information in ontology such as semantic relations in Arabic Named entity recognition (ANER). The semantic information in ontology was utilized in this work, but there are systems in other languages implemented it with some differences.

[8] https://taku910.github.io/crfpp/#download

In [10] the authors used semantic information in ontology. They used internal features (POS and Word n-gram) and external features from ontology using notebook domain ontology. The CRF classifier was utilized in this system. The system evaluated using ChnSentiCorp corpus.

The ontological features in [31] used for Vietnamese named entity recognition (NER). The authors used CRF classifier and VN-KIM dataset.

In [11] the authors proposed system for Recognize Named Entity in Behavioral Health. They built the manual ontology. The specific domain was used in this system using wordNet ontology.

The three last works showed the advantage of adopting semantic information in the ontology for NER.

## V. The System Architecture

In this work, ML-based ANER is proposed which utilizes two types of features: a) internal features, b) external features. Figure (1) illustrates the Architecture of this system. This system includes four phases: preprocessing (prepare and clean data), features extraction, training, and testing.

### *A. Pre-Processing*

This step includes cleaning the data such as splitting the sentence and tokenization. For preparing and cleaning data, the following processes was applied [31]:

- Remove the Redundant space among the words; remove all characters and symbols that attached to the word from the corpus such as (-, *, +, etc.) [4].
- Omit the prefixes and suffixes that attached as the conjunction ((wa /و) and (ba/ ب))**.**
- Remove the preposition li (lam) (ل).
- Delete all diacritics within the text.

Splitting the sentence is the task of segmenting the text into the sentences. The goal of this step is to define the boundaries of phases in the text according to POS tagger. The tokenization is the process that analyses and splits the input text into tokens such as, word, number, symbol, space. The objective of this step is to divide the sentence into the tokens in order help us to extract the features from ontology. In this step, the white space characters was used to define the tokens in the sentence.

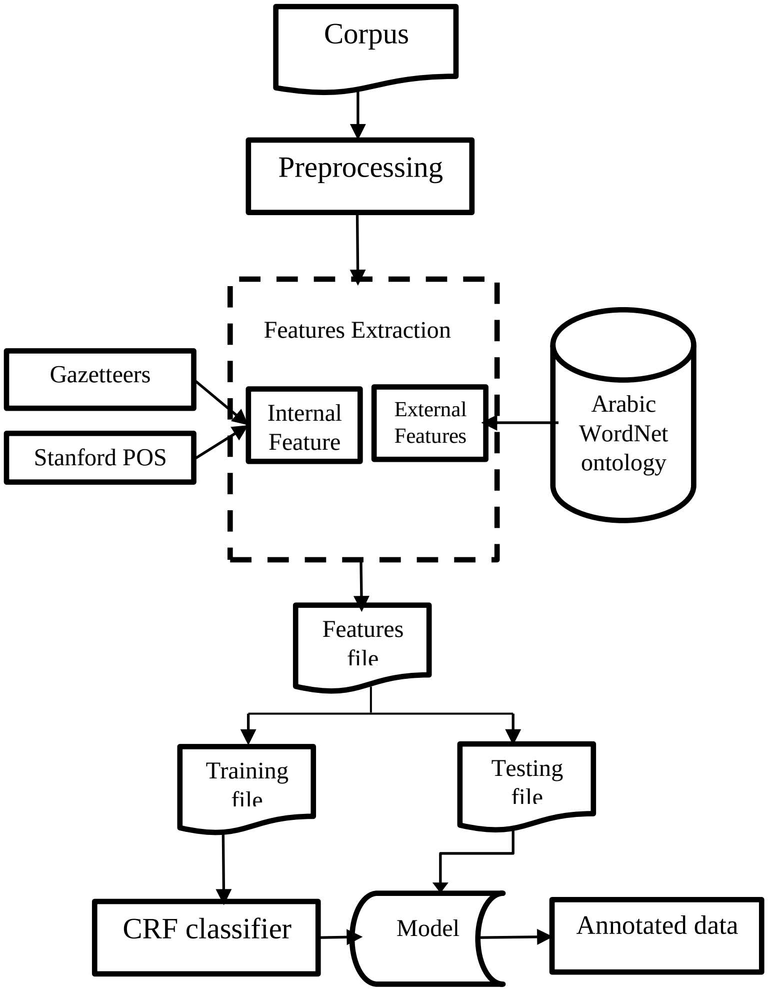

Fig. 1. The system Architecture

### *B. CRF classifier*

In this work, a modified Conditional Random Fields (CRF) is applied. CRF as described in [32] is a probabilistic framework used for segmenting and labeling the sequential data. It is a generalization of Hidden Markov Model in which its undirected graph contains nodes to represent the label sequence *y* corresponding to the sequence *x*. CRF finds the label which maximizes the conditional probability p(y|x) for a sequence x. The following equations represent CRF model:

$$\mathrm{p(c|y)} = \frac{1}{Z(x)} * \exp(\sum_i \lambda_i \, f_i \, (x, c)) \qquad (1)$$

$$\mathrm{Z}(x) = \sum_{c'} \quad \exp(\sum_i \lambda_i \, f_i \, (x, c')) \qquad (2)$$

Where c is the class, x is a context information and $f_i(x,c)$ is the ith feature.

In any ML approach for NER, there are two steps, training, and testing, as shown in figure (1). The first one builds the classifier model by using a set of features. In the second step, the classifier model that was built by the training step is utilized to predict a class for each token (word).

### *C. Features sets*

In the proposed framework, there are two category of features into two types: a) internal features, and b) external features that

*a) **Internal features** some important features for Arabic text are introduced as following:*

*1) **Word (WF):*** is the word itself.

*2) **Part of speech features (POS):*** part of speech tag is useful for ANER for determine the noun. The Stanford POS Tagger[9] was utilized to extract many tags NNP, NN, IN, JJ, NNPDT, NNDT, … etc.

*3) **Gazetteers features (GAZ):*** external resources and classes in Corpus are used, which are mentioned in data collection in section 5 to represent the existence of the word in the gazetteers.

*4) **Indicator features (CF)***

Indicator features are one of the most important features that lead to enhance the accuracy of NER recognition as it support the usage of semantic field feature [2]. They represents a set of words that may be used to identify NE such as preceding indicator words and post indicator. These words are used to recognize some names. For example, (الرئيس|الريس) (the President), (السيد|السيدة) (Mrs. | Miss), and (أبو) (Abu) for person names, (دولة | Country), (مدينة | City), and (شارع | Street) for location names, and (مجموعة | Group), (هيئة | Organization), (نادي | Culp), (شركة | Company), and (بنك | Bank) for Organization names.

*5) **Gram character features (GF)***

These features Presents the first/last two and three letters of the word. This feature is very important for ANER. For example, (عبد | Abd) is very repetitive prefix in Arabic person names.

*b) **External features (ontology or semantic features)***

AWN tool[10] has been modified to be able to analyze texts of wordNet ontology and establish correspondence between syntactic dependencies and semantic relations in order to extract the following features:

*1) **Class feature:*** represents the ontology's concept for the token as person for person names, (city | country or location) for location names, and (company or organization) classes for organization names.

*2) **Instance feature:*** which represent the corresponding instance for NE's token.

*3) **Relation features*** these features represent the relations between each two named entities. Therefore, in this step we aim to identify the trigger words that express the semantic relations between NEs from Arabic text. Based on the probability of relation that could exist between pairs of named entities (person, organization, location), cross multiplication is applied and we extracted all possible combination that may appear in the ontology. Furthermore, since Arabic relations could appear before the first NE, between NEs or after the second NE [33] such as (*__لعب__ أبو تريكة للأهلي* or *أبو تريكة __لعب__ للأهلي*) [Abu Trika **played** for Al Ahli]. In this work we only focus on the relation between a pairs of NEs such as a relation between person's concept and location's concept for example ( *بارك أوباما __رئيس__ امريكا*) [Barack Obama, **the President of** the United States] (Obama) is person name, (United States) is location name and (the President of) the relation between them. The relation between the person's concept and organization's concept for example (*بيل جيتس __مالك__ شركة مايكروسوفت*) [Bill Gates, **owner of** Microsoft Company.]. The relation between location's concept and organization's concept for example (*نادي برشلونة __من__ اسبانيا*) [FC Barcelona **from** Spain]. Finally relation between the location's concept and themselves for example (*القاهرة __عاصمة__ مصر*) [Cairo is **the capital of** Egypt]. Accordingly, two types of relations are identified:

*a)* ***Explicit relations*** which explicitly identified by Arabic wordNet (AWN) and are targeting the following pairs (PERS–PERS, LOC-LOC, PERS–ORG, ORG–LOC, and PERS-LOC)

*b)* ***Semantic Relations*** which are extracted depending on relationship between classes and their properties in AWN. Those types of relations are used to identify the following pairs (PERS–ORG, PERS-LOC, ORG–LOC, and LOC-LOC). The following algorithm shown in figure (2) is developed, to identify those possible relations between pairs of names entities. The algorithm works as follows:

1) For each sentence in the corpus, each two tokens are recognized and their classes are identified.

2) If both tokens are not belonging to the same class, calculate the semantic distance (SD) [11], which is

[9] http://www-nlp.stanford.edu/software/tagger.shtml

[10] http://sourceforge.net/projects/awnbrowser/

considered as the distance in hypernym/hyponym tree between the two classes of tokens.

3) If the semantic distance (SD) less than 3 and greater than 1 there is a relation between two tokens.

4) else if both classes are not found in AWN or both tokens have the same class, then there is no relation

```
Initialize R= O  // Represent the relation between classes
Take two tokens from sentence
Find the classes for two tokens from AWN
   CT1, CT2          // Represent the classes of token1, and token2 respectively
IF CT1 and CT2 not found OR CT1= CT2 Then
   Return O
Else
   SD=Calculate the semantic distance between CT1 and CT2
   IF $1 \leq SD \leq 3$ Then
      R= REL
      Return R
   Else
      Return O
   End IF
End IF
```

Fig. 2. Semantic relation extraction algorithm (SREA)

## VI. Data Collection

In order to train, and test the proposed ANER, necessary linguistic resources of different main categories were used: corpus, gazetteers, dictionaries, and AWN. Two corpuses are used for training, and testing the system. In this section, a description of all linguistic resources is presented.

*1) **ANERcorp** [11] **dataset***, which is freely available for research purposes, is a corpus prepared by Yassine Benajiba in ANER. It has 4901 sentences with 150286 tokens. Each token in this corpus is tagged according to the following classes:

- B-PERS: The Beginning of the person name.
- I-PERS: The Inside of the person name.
- B-LOC: The Beginning of the location name.
- I-LOC: The Inside of the location name.
- B-ORG: The Beginning of the organization name.
- I-ORG: The Inside of the organization name.
- O: The word is not a named entity (Other).

*2) **ALTEC** [12] **dataset*** which is not free, is a corpus prepared by Arabic Language Technology Center, it has 288737 tokens. Each token in this corpus is tagged according to the following classes:

- B-nep: The Beginning of the person name.
- I-nep: The Inside of the person name.
- B-nel: The Beginning of the location name.
- I-nel: The Inside of the location name.
- B-neo: The Beginning of the organization name.
- I-neo: The Inside of the organization name.
- O: The word is not a named entity (Other).

### 3) Gazetteers

Different gazetteers are integrated such as:

**ANERGazet**[13] is prepared by Yassine Benajiba gazetteers contained 2305 person names, 1785 location names and 390 organization names.

**Gate gazetteers** were containing 1883 person names, 403 location names and 215 organization names

**Lists of names** [14] form Wikipedia gazetteers were containing 16037 person names, and 4857 location names.

### 4) Arabic Wordnet[15]

The AWN ontology contains a large amount of location's class instance and a few instance of person class and organization class. We dealt with sub-ontology: person, location, and organization such as Figure (3).

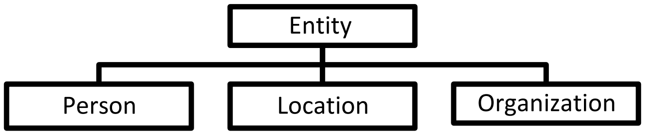

Fig. 3. Sub-ontology of AWN ontology

## VII. Data Evaluation

The CONLL evaluation standard metrics of precision, recall and F-measure are used [34]. Precision and recall can be express as shown in the Equation (3) and (4):

$$\text{Precision} = \frac{true\ positive}{true\ positive\ + false\ positive} \tag{3}$$

$$Recall\ = \frac{true\ positive}{true\ positive\ + false\ negativ} \tag{4}$$

Also the F-measure (F) was used, which is defined as a weighted combination of precision and recall as Equation (5):

$$F - \text{measure}\ =\ 2 * \frac{\text{Precision}\ *\ Recall}{\text{Precision}\ +\ Recall} \tag{5}$$

## VIII. Experiments

For the evaluation of this system, the two corpora datasets that are mentioned in section (6) are used. Since both datasets do not follow the same tagging conventions as in section (6), training and testing were conducted separately for each dataset. The ANERCorp dataset was used to compare our work with previous works. The datasets is divided into 80% as

[11] http://users.dsic.upv.es/~ybenajiba/
[12] http://www.altec-center.org/Repository_65.html

[13] http://users.dsic.upv.es/~ybenajiba/
[14] http://en.wikipedia.org/wiki/List_of_Arabic_names
[15] http://globalwordnet.org/arabic-wordnet/awn-browser/

training dataset and 20% as testing dataset according to Abdul-Hamid and Darwish [17], and Kareem Darwish [18].

New Semantic information features have added into CRF. CRF++ tool[16] is used for training and testing. In training step, the CRF++ tool needs two input template files and training data file. The tool output is the classifier model file. The template file describes which features are used in training and testing. In each template, special macro %x[row , col] will be used to specify a token in the input data. Row specifies the relative position of the current focusing token and col specifies the absolute position of the column. In training data file, each word is represented by a set of features and its actual NE's class in order to produce a CRF classifier. In testing step, the tool needs the output of training step (model file) and testing data file. Output of this step is the predication class for each word.

For semantic information features in ontology, the AMN tool is modified to extract all features from ANW ontology for all words in datasets (mapping between the dataset and AWN ontology). The information was extracted that needed, such as person information, location information and organization information. The semantic information features was introduced in a features file as CRF features. Other feature were added called PART, it represents the classes of two words and a relation between them. For example ( القاهرة عاصمة مصر) [Cairo the capital of Egypt] all sentence is a PART. Contextual window size parameter was used as experimental factor to our feature engineering experiments. Window size significantly effects on NER accuracy. Three type of window size was utilized in this work -1/+1, -2/+2, and -2/+1. Based on the experiments conducted, the window size -2/+1 is best choice for that datasets used in this work.

## IX. Results and Discussion

The system was trained on the data in cumulative additions of features. That said, the system was trained on first two features (WF and POS), then adding GAZ, and so on. The last added is the semantic information features. Table [1] shows the results of ANERcorp and ALTEC datasets obtained from CRF for all feature sets in terms of precision(P), recall(R) and F-measures(F) for Person, Location and Organization. The best results for P, R, and F are **bolded** in the tables. These results show the effect of using cumulative additions of features on training accuracy. There is the most significant impact on performance when adding GAZ. The second feature is semantic information. When all were combined (Table [1]), the resulting precision is (94.44%) which was almost (0.46%) above the best precision obtained, by WF_POS features (93.98%). The recall is (82.13%) which was about (1.04%) above the best recall obtained, by WF_POS_GAZ_CF_GF features (81.09%). In addition, F-measure (87.86%) was most (0.90%) above the best F-measure achieved, by WF_POS_GAZ_CF_GF features (86.96%) when used ANERcorp dataset. While when applied on ALTEC dataset, the recall (79.52%) was (0.95%) point over the best recall got, by WF_POS_GAZ_CF_GF features (78.57%). In addition, F-measure (84.72%) was (1.59%) over the best F-measure acquired, by WF_POS_GAZ_CF_GF features (83.13%).

The results of the experiment illustrate that the ANER with semantic information (ontology features) can achieve better performance. The precision, recall, and F-measure of semantic information (ontology features) are higher than other features, which means that adding semantic information (ontology) can improve the precision, recall, and F-measure of ANER. The reasons may be that ontology is a kind of concept models that could describe the system at the level of semantics and knowledge.

Tables [2] and [3], summaries the best results of this system on ANERCorp and ALTEC datasets respectively.

In comparison to results with previous work, this system outperforms result of other Arabic NER systems when applied on ANERcorp dataset as shown in Table [4]. It also outperforms the previous systems regarding F-measure in extracting Person, Location and Organization NEs from ANERcorp with an overall F-measure= 87.86 %.

We compare our work with previous works done by Benajiba et al. [4], Abdul-Hamid et al. [17], and Darwish [18], which produce better results than their system with less number of features.

## X. Conclusion

This paper presented an integration of features set fort named entity recognition in Arabic. This integration combines internal features that represented linguistic features as well as external features to represent the semantic of Arabic language. The internal features such POS, GAZ, indicator, and Cram character features while the external features is semantic information features were extracted from Arabic wordNet ontology such as classes, instance and relations.

The integration model helped overcome some of the orthographic and morphological complexities of Arabic. Experimental results show F-Measure for ANERCorp and ALTEC around 87.86% and 84.72% respectively. The proposed feature set achieved improved results over those in the literature with as much as 3.56% F-measure improvement for recognizing NE.

In the future, we intend to study the possibility of improving the system performance using other approaches such as Ruled Based approach and Hybrid approach with semantic information features.

[16] https://taku910.github.io/crfpp/#download

TABLE I. RESULTS FOR SUCCESSIVE ADDITION OF FEATURES ON ANERCORP AND ALTEC DATASETS

| | | ANERCorp | | | ALTEC | | |
|---|---|---|---|---|---|---|---|
| **Feature sets** | **Type** | **P** | **R** | **F** | **P** | **R** | **F** |
| WF + POS | PERS | 93.26 | 58.14 | 71.63 | **92.81** | 61.92 | 74.28 |
| | LOC | **95.35** | 71.69 | 81.85 | **93.27** | 80.16 | 86.22 |
| | ORG | **93.33** | 53.85 | 68.29 | **89.66** | 64.11 | 74.76 |
| | Overall | 93.98 | 61.23 | 74.15 | **91.91** | 68.73 | 78.65 |
| WF + POS+ GAZ | PERS | 95.20 | 84.25 | 89.39 | 89.91 | 74.68 | 81.59 |
| | LOC | 94.05 | 84.87 | 89.22 | 86.74 | 85.76 | 86.25 |
| | ORG | 91.32 | 72.25 | 80.68 | 83.23 | 74.80 | 78.79 |
| | Overall | 93.52 | 80.46 | 86.49 | 86.63 | 78.41 | 82.32 |
| WF + POS+ GAZ+ CF | PERS | **96.06** | 81.80 | 88.36 | 90.02 | 75.20 | 81.95 |
| | LOC | 92.57 | 87.87 | 90.16 | 89.43 | 85.33 | 87.29 |
| | ORG | 92.83 | 71.15 | 80.56 | 85.81 | 71.23 | 77.85 |
| | Overall | 93.82 | 80.28 | 86.52 | 88.39 | 77.26 | 82.45 |
| WF + POS+ GAZ+ CF + GF | PERS | 93.90 | 84.28 | 88.83 | 90.11 | 74.04 | 81.29 |
| | LOC | 94.60 | 88.85 | 91.63 | 90.70 | 85.78 | 88.17 |
| | ORG | 92.75 | 70.14 | 79.88 | 83.94 | **75.89** | 79.71 |
| | Overall | 93.75 | 81.09 | 86.96 | 88.25 | 78.57 | 83.13 |
| All | PERS | 95.44 | **85.13** | **89.99** | 91.21 | **78.80** | **84.55** |
| | LOC | 94.59 | **88.94** | **91.68** | 92.65 | **86.90** | **89.68** |
| | ORG | 93.29 | **72.33** | **81.48** | 88.08 | 72.88 | **79.76** |
| | Overall | **94.44** | **82.13** | **87.86** | 90.65 | **79.52** | **84.72** |

TABLE II. BEST RESULTS ON ANERCORP DATASET

| Type | P | R | F |
|---|---|---|---|
| PERS | 95.44 | 85.13 | 89.99 |
| LOC | 94.59 | 88.94 | 91.68 |
| ORG | 93.29 | 72.33 | 81.48 |
| Overall | 94.44 | 82.13 | 87.86 |

TABLE III. BEST RESULTS ON ALTEC DATASET

| Type | P | R | F |
|---|---|---|---|
| PERS | 91.21 | 78.80 | 84.55 |
| LOC | 92.65 | 86.90 | 89.68 |
| ORG | 88.08 | 72.88 | 79.76 |
| Overall | 90.65 | 79.52 | 84.72 |

TABLE IV. COMPARISON WITH OTHER ARABIC NER SYSTEMS ON ANERCORP DATASET

| System | Person | Location | organization | Overall |
|---|---|---|---|---|
| | F-Measure | F-Measure | F-Measure | F-Measure |
| CRF-based system [4] | 73.35 | 89.74 | 65.76 | 79.21 |
| Abdul-Hamid and Darwish [17] | 82.00 | 88.00 | 73.00 | 81.00 |
| Kareem Darwish [18] | 82.10 | 90.00 | 72.90 | 84.30 |
| **Our System** | **89.99** | **91.68** | **81.48** | **87.86** |

## REFERENCES

[1] K. Shaalan, "A survey of Arabic named entity recognition and classification." *Computational Linguistics* vol. 40, no. 4, pp. 469-510, 2014.

[2] S. AbdelRahman, M. Elarnaoty, M. Magdy, & A. Fahmy, "Integrated machine learning techniques for Arabic named entity recognition." IJCSI, vol. 7, pp. 27-36, 2010.

[3] N. F. Mohammed, and N. Omar, "Arabic named entity recognition using artificial neural network." *Journal of Computer Science* vol. 8, no. 8, pp. 1285, (2012).

[4] Y. Benajiba, and P. Rosso, "Arabic named entity recognition using conditional random fields." *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*. vol. 8. 2008.

[5] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition using optimized feature sets." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

[6] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition: A feature-driven study." *Audio, Speech, and Language Processing, IEEE Transactions.*, vol. 17, no. 5, pp. 926-934, 2009.

[7] K. Shaalan, and H. Raza, "Person name entity recognition for Arabic."*Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Association for Computational Linguistics, 2007.

[8] K. Shaalan, and H. Raza, "Arabic named entity recognition from diverse text types." *Advances in Natural Language Processing*. Springer Berlin Heidelberg, pp. 440-451, 2008.

[9] K. Shaalan, and H. Raza, "NERA: Named entity recognition for Arabic." *Journal of the American Society for Information Science and Technology*. vol. 60, no. 8, pp. 1652-1663, 2009.

[10] F. Luo, H. Xiao, and W. Chang, "Product named entity recognition using conditional random fields." *Business Intelligence and Financial Engineering (BIFE), 2011 Fourth International Conference on*. IEEE, 2011.

[11] U. Yasavur, R. Amini, C. L. Lisetti, & R. Napthali, "Ontology-Based Named Entity Recognizer for Behavioral Health." FLAIRS Conference. 2013.

[12] D. Sánchez, D. Cisneros, and F. A. Gali, "UEM-UC3M: an ontology-based named entity recognition system for biomedical texts." Association for Computational Linguistics, 2013.

[13] M. Aboaoga, and M. I. Ab Aziz, "Arabic person names recognition by using a rule based approach." *Journal of Computer Science. vol.* 9, no. 7, pp. 922, 2013.

[14] A. Elsebai, F. Meziane, and F. Z. Belkredim, "A rule based persons names Arabic extraction system." *Communications of the IBIMA*. vol. 11, no. 6, pp. 53-59, 2009.

[15] Y. Benajiba, M. Diab, and P. Rosso, "Arabic named entity recognition using optimized feature sets." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

[16] K. Darwish, and W. Gao, "Simple Effective Microblog Named Entity Recognition: Arabic as an Example." 2014.

[17] A. Abdul-Hamid, and K. Darwish, "Simplified feature set for Arabic named entity recognition." *Proceedings of the 2010 Named Entities Workshop*. Association for Computational Linguistics, 2010.

[18] K. Darwish, "Named Entity Recognition using Cross-lingual Resources: Arabic as an Example." *ACL (1)*. 2013.

[19] A. Zirikly, and M. Diab, "Named Entity Recognition for Dialectal Arabic."*ANLP 2014*, pp. 78 2014.

[20] Y. Benajiba, M. Diab, and P. Rosso, "Using Language Independent and Language Specific Features to Enhance Arabic Named Entity Recognition." *Int. Arab J. Inf. Technol.* vol. 6, no. 5, pp. 463-471, 2009.

[21] Y. Benajiba, P. Rosso, and J. M. Benedíruiz, "Anersys: An arabic named entity recognition system based on maximum entropy."*Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg, pp. 143-153, 2007.

[22] O. Zayed, S. El-Beltagy, and O. Haggag, "An Approach for Extracting and Disambiguating Arabic Persons' Names Using Clustered Dictionaries and Scored Patterns." *Natural Language Processing and Information Systems*. Springer Berlin Heidelberg, pp. 201-212, 2013.

[23] A. Hassan, H. Fahmy, and H. Hassan, "Improving named entity translation by exploiting comparable and parallel corpora." *AMML07* 2007.

[24] D. Samy, A. Moreno, and J. M. Guirao, "A proposal for an Arabic named entity tagger leveraging a parallel corpus." *International Conference RANLP, Borovets, Bulgaria*. 2005.

[25] E. Kamal, M. Rashwan, and S. Alansary, "High Quality Arabic Lexical Ontology Based on MUHIT, WordNet, SUMO and DBpedia."*Computational Linguistics and Intelligent Text Processing*. Springer International Publishing, pp. 98-111, 2015.

[26] S. Elkateb, W. Black, and P. Vossen, "Building a wordnet for arabic." *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*. 2006.

[27] M. Althobaiti, U. Kruschwitz, and M. Poesio, "A Semi-supervised Learning Approach to Arabic Named Entity Recognition." *RANLP*. 2013.

[28] O. Zayed, S. El-Beltagy, and O. Haggag, "A novel approach for detecting Arabic persons' names using limited resources." *Complementary Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. 2013.

[29] M. Althobaiti, U. Kruschwitz, and M. Poesio, "Combining Minimally-supervised Methods for Arabic Named Entity Recognition." *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 243-255, 2015.

[30] T. T. Nguyen, and T. H. Cao, "Linguistically Motivated and Ontological Features for Vietnamese Named Entity Recognition." *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2012 IEEE RIVF International Conference on*. IEEE, 2012.

[31] W. Zaghouani, "RENAR: A rule-based Arabic named entity recognition system." *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 11, no. 1, 2012.

[32] J. Lafferty, A. McCallum, and F. CN. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." 2001.

[33] I. Boujelben, S. Jamoussi, and A. Hamadou, "A hybrid method for extracting relations between Arabic named entities." *Journal of King Saud University-Computer and Information Sciences*, vol. 26, no. 4, pp. 425-440, 2014.

[34] A. De Sitter, T. Calders, and W. Daelemans, "A formal framework for evaluation of information extraction." *Online http://www. cnts. ua. ac. be/Publications/2004/DCD04*, 2004.