Hybrid Named Entity Recognition - Application to Arabic Language

Mohamed A. Meselhi[1], Hitham M. Abo Bakr[1], Ibrahim Ziedan[1], and Khaled Shaalan[2]
[1]Derpartment of Computer and System Engineering; Faculty of Engineering, Zagazig University, Egypt,
mohatef@zu.edu.eg, hithamab@yahoo.com, iziedan@yahoo.com
[2]The British University; Dubai, UAE, Khaled.Shaalan@buid.ac.ae

Abstract— **Most Named Entity Recognition (NER) systems follow either a rule-based approach or machine learning approach. In this paper, we introduce out attempt at developing a hybrid NER system, which combines the rule-based approach with a machine learning approach in order to obtain the advantages of both approaches and overcomes their problems [1]. The system is able to recognize eight types of named entities including Location, Person, Organization, Date, Time, Price, Measurement and Percent. Experimental results on ANERcorp dataset indicated that our hybrid approach outperforms the rule-based approach and the machine learning approach when they are processed separately. Moreover, our hybrid approach outperforms the state-of-the-art of Arabic NER.**

I. INTRODUCTION

In 1995, the Named Entity Recognition (NER) task was first introduced by the Message Understanding Conference (MUC-6)[1]. Three subtasks were mainly defined: ENAMEX (for the Person, Location, and Organization), TIMEX (for Date and Time expressions), and NUMEX (for monetary amounts and percentages). Named Entity Recognition (NER) task is an essential preprocessing task for many Natural Language Processing (NLP) applications such as text summarization, document categorization, Information Retrieval, among others [2]. NER seeks to identify the sequence of words in a document that can be classified under a predefined category of named entity such as Person, Organization, and Location names in addition to temporal and monetary expression. In this paper we concentrate on NER for Arabic. Two main types of approaches are utilized to develop NER system including rule-based approach and machine learning approach. The rule-based approach has better results for specific domains and its ability to recognize complex entities. However, the maintenance of rule-based system requires high cost and consumes a lot of time. In addition, this approach does not adapt to other domains or languages. The machine learning approach is updatable and domain independent but requires a large amount of training data in order to obtain a good performance [3]. We integrate a rule-based NER approach, a reproduction of NERA [4], with a machine learning NER approach, in particular SVM, in order to–improve the performance of Arabic NER. The rule-based component depends on a set of grammar rules. Whereas, the machine learning component depends on a set of features extracted from the annotated text. We could identify some recognition errors by using the tag selection and correction component which compares between the results of the rule-based component and machine learning component.

The remainder of this paper is organized as follows. Section 2 introduces a background about Arabic Language Characteristics. Section 3 describes the structure of our proposed hybrid system and its components. Experimental results are discussed in Section 4. In Section 5 we give some concluding remarks.

II. ARABIC LANGUAGE CHARACTERISTICS

Arabic is a highly inflected language, with a rich morphology and complex syntax [5]. Generally, the significance of Arabic worldwide is too obvious to enumerate. The language is spoken by Arab world, and Islamic countries. The main challenges and characteristics of Arabic NER task are as follows:

Lack of capital letters in the Arabic orthography: A named entity (NE) in Latin languages is usually distinguished by a capital letter at the word beginning. However, this is not the case in Arabic which makes the detection of NE in text based on the case of letters more difficult. Some efforts to overcome this problem have used lexical triggers that are derived from analyzing the NE's surrounding context [6] while some others have used the gloss feature of the English translation of the NE produced by the Morphological Analysis and Disambiguation for Arabic (MADA) tool [2][7].

Complex Morphology: Arabic morphology is very complex because of the language agglutinative nature [8]. There are three types of agglutinative morphemes: stems, affixes and clitics. The primitive form of the word is the stem. Affix letters are attached to the stem which has three types: prefixes attached before the stem, suffixes attached after the stem, and circumfixes that surround the stem. Clitics are also attached to the stem after affixes. They play a syntactic role at the phase level. Clitics are either proclitics that precede the word or enclitics that follow the word. The conjunction and "و" (*wa*[3], and) [9] and object pronoun "هن" (*hn*, they-3rdP-fem) are examples of proclitics and enclitics, respectively.

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

[2] http://www1.ccls.columbia.edu/MADA/

[3] Habash-Soudi-Buckwalter transliteration scheme

- **Ambiguity:** diacritics are rarely used in modern Arabic texts which most often lead to ambiguous situations i.e. different meaning [6]. For example, consider the word مصر which if it is diacritized as مِصر miSr it means the noun 'Egypt' but if it is diacritized as مُصِرّ muSir~ it means the adjective 'insistent'. In addition to the diacritization problem, Arabic words can differ in meanings depending on the context in which it appears. Consider the following sentence: هناك تواجد أمني مكثف بجوار مصطفى محمود (There is a heavy security presence beside Mustafa Mahmoud). In this sentence, مصطفى محمود (Mustafa Mahmoud) might represent either a Person or a Location name which can be resolved from the context.
- **Lack of Resources:** The lack of Arabic NER freely available resources or the expense of creating or licensing these important Arabic NER resources make NER task far more challenging. So, researchers had to exert considerable effort building their own resources. We used both ANERcorp[4] [10] and Automatic Content Extraction (ACE 2005)[5] corpora in order to train and evaluate our proposed system.

III. THE APPROACH

Both machine learning-based and rule-based approaches have their own strengths and weaknesses and by combining them in one system they could achieve a better performance than applying each of them separately. Our proposed system tries to avoid sequential implications of the most closely related research presented by Oudah and Shaalan [11] using pipeline hybrid approach. As a result of using two dependent components including rule-based component and machine learning component where the output of rule-based component is passed as input features to machine learning component. Firstly, any error in the classification by the rule-based component will be used by the classifier in the training phase which leads to inaccurate results in the test phase. Secondly, any modification in the grammar rules or updating the system gazetteers requires retraining the classifier. Our proposed hybrid system consists of three main components: machine learning component, rule-based component and tag selection and correction component.

A. The Machine Learning Component

Arabic NER is challenging due to the highly ambiguous nature of Arabic named entities (NEs). A NE can very well appear as a non-NE in Arabic text. Moreover, the modern style of Arabic writings allows for optional diacritization and different methods for performing transliterations. So, these peculiarities of the language raise the need for Machine Learning (ML) algorithms which lend itself to the Arabic NER task. These algorithms usually involve a selected set of features, extracted from datasets annotated with NEs, which is used to generate a statistical model for NE prediction.

In the literature, the ML approaches used in NER systems are Maximum Entropy (ME) [12], Hidden Markov Model (HMM) [13], Conditional Random Fields (CRFs) [14], [15] and Support Vector Machines (SVM) [16].

The machine learning component uses SVM because of its robustness to noise and ability to deal with a large number of features effectively [17]. SVM is a training algorithm that builds a model from large marked training examples that belongs to one of two classes in order to classify test examples into one class [16]. The SVM task isn't limited to find just any hyperplane between the two classes but it searches for the best hyperplane which maximizes the distance from example points. We used YamCha (Yet Another Multipurpose Chunk Annotator) toolkit[6] in order to train and test SVM models.

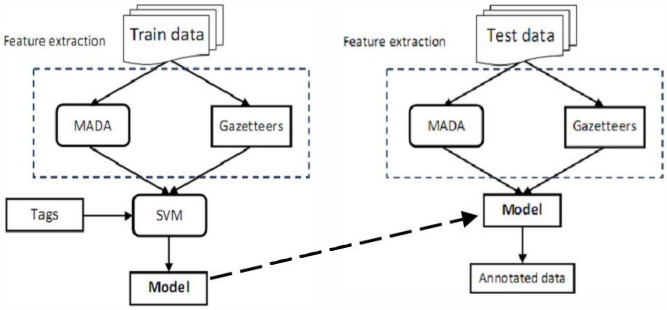

Figure 1. Training and test phases

In Machine Learning-based NER approach, there are two phases: training phase and test phase. The training phase generates the classifier (model) by using a set of classification features of the input text, where each word in the input text is represented by a set of features and its actual NE's type. In the test phase, the classifier generated by the training phase is utilized to predict a class for each token (word) in the test dataset. The architecture of training and test phase in the employed system is illustrated in Figure 1.

B. The Feature Set

Feature selection refers to the task of identifying a useful subset of features chosen to represent elements of a larger set (i.e., the feature set). In the rest of this section, we discuss the feature space.

Contextual word feature (CXT): The features of a sliding window comprising a word *n-gram* that includes the targeted word, along with preceding and succeeding words For example, in the training corpus the verb "حضر" (HDr, attend) appears frequently before a NE of type Person. So this feature will allow the classifier to mostly assign a Person NE tag to the word after this verb.

Gazetteer features (GAZ): This binary feature represents the presence of the word from a respective gazetteer which indicates NE type. We rely on the following gazetteers:

- Location Gazetteer (countries, cities, rivers and mountains)
- Person Gazetteer (names of people)

[4] http://www1.ccls.columbia.edu/~ybenajiba/
[5] https://catalog.ldc.upenn.edu/LDC2006T06
[6] http://chasen.org/~taku/software/yamcha/

- Organizations Gazetteer (companies and other organizations)

Morphological features (MORPH): A set of morphological information performed through MADA. The morphological features produced include the following:

- Aspect: One of the three aspects of the Arabic verb: perfective (ماضي, mADy), imperfective (مضارع, mDArE), or imperative (أمر, >mr).
- Case: One of three cases of Arabic nominals: nominative (مرفوع, mrfwE), accusative (منصوب, mnSwb), or genitive (مجرور, mjrwr).
- Gender: A binary value indicating the gender of the word, i.e. masculine or feminine.
- Number: Any of the three values indicating the number of the word, i.e. singular, dual, or plural.
- Mood: Any of the three Arabic moods that only vary for the imperfective verb: indicative (مرفوع, mrfwE), subjunctive (منصوب, mnSwb), or jussive (مجزوم, mjzwm).
- State: Any of the three values: definite, indefinite, or construct.
- Voice: A binary value indicating either passive or active voice.
- Proclitics and Enclitics: exact clitics that are attached to the stem.

Part-of-Speech (POS): A binary value indicating whether or not the POS tag (extracted by MADA) is a noun or proper noun.

Gloss: A binary value indicating whether the English translation (gloss) provided by MADA starts with a capital letter.

Lexical features (LEX): This feature considers the orthography of each token in the text. The usefulness of lexical features mostly appears when the same NE occurs in different places of the text but with some difference in the orthography. For example, in the sentence: أوباما يدافع عن برامج مراقبة الهواتف والإنترنت (*>wbAmA ydAfE En brAmj mrAqbp AlhwAtf wAl<ntrnt*, Obama defends program of surveillance phones and the Internet). Transliteration variant of foreign names is a common problem in Arabic writing. For example, Obama may be transliterated to Arabic as (أوباما, >wbAmA) or (أباما, >bAmA). So, in the training corpus, if Obama has only appeared with the first transliteration, the classifier cannot classify the second transliteration. However, when we used the last characters of the word as a feature, it would help the classifier to classify second transliteration also.

C. The Rule-based Component

The initial version of the rule-based component is developed using the design documents of NERA, a NER system for Arabic that is implemented using GATE[7]. Figure 2 shows the construction of the Rule-based component. The system applies three procedures: firstly, tokenizing the target document using Arabic Tokenizer tool within GATE which divides input text into tokens where each is either a word or a number or a punctuation mark. Secondly, recognizing and classifying NEs in text by the exact matching with gazetteers entries. Samples of Location gazetteers with transliteration and English translation are shown in Table I.

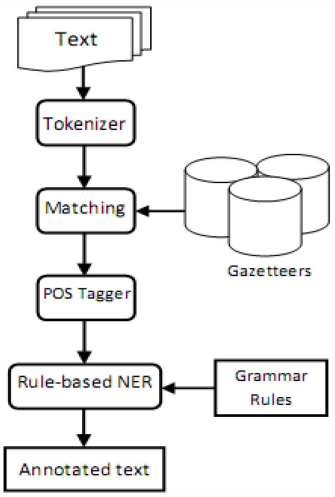

Figure 2. The Rule-based component

TABLE I. SAMPLES FOR LOCATION GAZETTEERS

Gazetteer name	No. of entries	Entry Translation Transliteration
City	3221	القاهرة Cairo AlqAhrp
Country	243	كندا Canada kndA
Location_Indicator	68	مدينة City mdynp
Continents	45	أسيا Asia |syA
Rivers	240	نهر الفرات Euphrates River nhr AlfrAt

Finally, execute a finite-state transducer, based on a set of local grammar rules that are implemented using JAPE. JAPE (Java Annotation Patterns Engine) [18] is a pattern matching language that provides finite state transduction over annotations based on regular expressions. A JAPE grammar consists of a set of phases, each phase consists of a set of rules. Each JAPE rule consists of Left and Right sides. The Left Hand Side describes the pattern to be matched, while the Right Hand Side allocates annotations to be created. Annotations matched on The Left Hand Side of a rule are referred to the

[7] http://gate.ac.uk/

Right Hand Side by means of labels. The following example illustrates an implemented rule for recognizing an organization name. This rule identifies an organization name (token) that starts with Organization indicator (e.g.”نادي”, club) and then any token, which is optionally followed by a token (“في”), and then followed by a nationality or any Location.

```
Priority:30
(({Lookup.majorType==" Organization_Indicator "}
({Token})+):Org2
({Token.string == "في"})?
({Lookup.majorType==" Nationality"}|
{Lookup.minorType=="Location"}))
-->
:Org2.Organization= {rule="OrganizationRule2"}
```

Figure 3. A rule implemented in JAPE for recognizing organization name using surrounding indicators

This rule may detect a false positive (i.e., annotated wrongly) Organization named entity as in the following example:

قرر الإلتحاق بأى جامعة موجودة في مصر

He decided to join any university existing in Egypt

The expression (“جامعة موجودة”, “university existing”) was recognized as an Organization named entity. This problem can be solved using Stanford POS Tagger[8] that assigns part of speech category to each word in the text. It can limit words which come after the indicator in specific types including noun and proper noun. We modified the preceding rule to verify whether or not the targeted token is a noun or a proper noun in following example.

```
Priority:30
(({Lookup.majorType==" Organization_Indicator "}
({Token,Token.category == noun_prop }|
{Token,Token.category == noun})+):Org2
({Token.string == "في"})?
({Lookup.majorType==" Nationality"}|
{Lookup.minorType=="Location"}))
-->
:Org2.Organization= {rule="OrganizationRule2"}
```

Figure 4. A rule implemented in JAPE for recognizing organization name using its POS category

D. *Tag Selection and Correction Component*

Figure 5 shows the architecture of our proposed Arabic hybrid NER system. The input data is applied in parallel to the rule-based and machine learning-based Arabic NER components. The tagging results for each component are compared with each other in order to agree on the final tags. From the analysis of each component results we conclude that the performance of the machine learning classifier is more accurate than the performance of the Rule-based component. So, the role of this component is to fine-tune the machine learning system’s output by checking the most false negatives (i.e., missing annotations) and applying the correction using the tagging decisions determined by the rule-based component as shown in Figure 6.

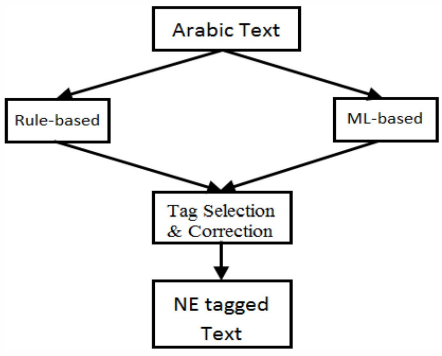

Figure 5. System Architecture

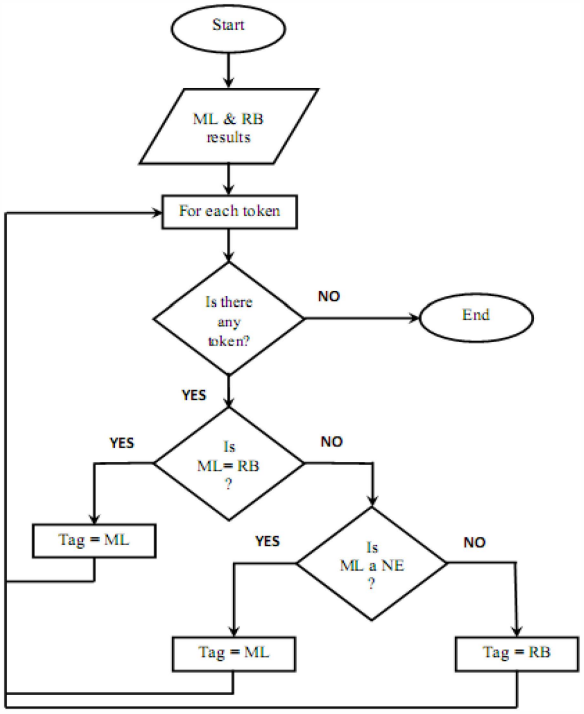

Figure 6. Tag selection and correction

IV. EXPERIMENTS AND RESULTS

In order to study the impact of the NER approaches we have examined each component separately before combining the results. In other words, we are dealing with three annotated outputs from the machine learning component, rule-based component, and the overall hybrid output, respectively. The experimental setting for the machine learning component uses splits of the ANERcorp dataset into three datasets: 90% as a training dataset, 5% as a development dataset, and 5% as a test dataset.

[8] http://nlp.stanford.edu/software/tagger.shtml

As far as the machine learning component is concerned, the experiment proceeds in three stages. In the first stage, the training is applied on the training dataset using selected feature set and the results are analyzed to determine the best feature set. In the second stage, the training is applied on the combined training and development datasets using the best selected feature set. In the third stage, the classifier is applied on the test dataset and the results are reported and discussed.

The baseline feature set consists of tokens at window size that ranges from -1/+1 to -4/+4.We found that a context size of one previous token and one subsequent token (i.e. window size is 3) achieves the best performance in this task. The baseline model has achieved a precision of 93.09% and recall of 68.98%. This indicates that adding extra features would improve the performance by increasing its coverage.

The selection of the subset to be utilized by a classifier is a very critical to the NER system's performance such that when optimized it can enhance the quality of the machine learning component in particular and the entire proposed system in general. So, the first step in the proposed hybrid approach is to select the significant features from the training dataset. Then, study the impact of each feature one at a time. Next, we sort them in a decreasing order according to the performance achieved. Table II shows the order of the features according to the obtained results.

TABLE II. FEATURES ORDER ACCORDING TO ITS IMPACT

Order	Feature(s)
1	GAZ
2	POS
3	MORPH
4	Gloss
5	LEX

Finally, the system is trained on the first feature and the performance is measured (i.e. GAZ). Then the system is trained on the first two features and the performance is measured (i.e. GAZ+POS), and so on until the best performance is obtained. Table III shows the obtained results carried out on ANERcorp dataset obtained from machine learning component for different feature sets in terms of Precision, Recall and F-measures [19], for Location, Person and Organization. These results show the impact of using successive addition of different features. The Gloss feature has a noticeable impact on the NER results. When the Gloss feature is included, it achieved an overall improvement in terms of F-measure by 0.76% which emphasizes the important role of the corresponding English capitalization in this task. Best results are obtained when all the features are selected.

As far as the rule-based component is concerned, the untagged version of the reference dataset (i.e. ANERcorp) is entered to GATE for processing, where the annotated result can be automatically evaluated using *Annotation Diff* tool of GATE which has an elegant GUI for presenting the results. The obtained results carried out on ANERcorp dataset obtained from rule-based component for Location, Person and Organization are shown in Table IV.

As far as the tag selection and correction component is concerned, the results of machine learning and rule-based components are compared in order to agree the final tags. Table V shows the results obtained from the tag selection and correction component.

TABLE III. RESULTS FOR SUCCESSIVE ADDITION OF FEATURES

Feature set	Type	P	R	F
CTX	PER	0.9769	0.6259	0.763
	ORG	0.8807	0.6575	0.7529
	LOC	0.919	0.7782	0.8428
	Overall	0.9309	0.6898	0.7924
CTX+GAZ	PER	0.9609	0.9111	0.9354
	ORG	0.8394	0.7877	0.8127
	LOC	0.9298	0.8548	0.8908
	Overall	0.9227	0.863	0.8918
CTX+POS+GAZ	PER	0.9764	0.9185	0.9466
	ORG	0.8633	0.8219	0.8421
	LOC	0.9221	0.8589	0.8894
	Overall	0.9311	0.875	0.9022
CTX+POS+GAZ+MORPH	PER	0.9767	0.9296	0.9526
	ORG	0.8707	0.8767	0.8737
	LOC	0.9339	0.8548	0.8926
	Overall	0.9366	0.8901	0.9127
CTX+POS+GAZ+MORPH+Gloss	PER	0.9844	0.9333	0.9582
	ORG	0.8889	0.8767	0.8828
	LOC	0.9389	0.8669	0.9015
	Overall	0.9459	0.8961	0.9203
All	PER	0.9767	0.9333	0.9545
	ORG	0.8919	0.88	0.8859
	LOC	0.9612	0.8884	0.9234
	Overall	0.9514	0.9046	0.9274

TABLE IV. OBTAINED RESULTS FROM RULE-BASED COMPONENT

Type	P	R	F
PER	0.9798	0.9	0.9382
ORG	0.8129	0.9267	0.866
LOC	0.9526	0.8805	0.9151
Overall	0.9263	0.8987	0.9123

TABLE V. OBTAINED RESULTS FROM THE TAG SELECTION AND CORRECTION COMPONENT

Type	P	R	F
PER	0.9701	0.963	0.9665
ORG	0.9	0.96	0.929
LOC	0.9518	0.9442	0.948
Overall	0.9468	0.9553	0.951

The previous tables show that the results for extracting Location, Person and Organization named entities from machine learning approach (SVM) are better than those from rule-based approach whereas the hybrid system performs the best. On the other hand, the extraction of named entities of type Date, Time, Price, Measurement and Percent can rely only on rule-based component, because it can be identified directly from the text. Table VI shows the results carried out on ACE 2005 corpus. This table shows the high results achieved by using only the rule-based component. Figure 7 shows that our hybrid system outperforms the state-of-the-art Arabic hybrid NER system by [11] when carried out on ANERcorp dataset for extracting Location, Person and Organization named entities and ACE 2005 BN dataset for the rest of the NE types.

TABLE VI. RESULTS OF TIMEX AND NUMEX NAMED ENTITIES WHEN APPLIED ON ACE 2005 NW AND ACE 2005 BN

Dataset	Type	P	R	F
ACE 2005 NW	Date	1.00	0.966	0.983
	Time	0.995	0.953	0.974
	Price	1.00	1.00	1.00
	Measurement	1.00	0.974	0.987
	Percent	1.00	1.00	1.00
ACE 2005 BN	Date	1.00	0.959	0.979
	Time	1.00	1.00	1.00
	Price	1.00	1.00	1.00
	Measurement	1.00	0.986	0.994
	Percent	1.00	1.00	1.00

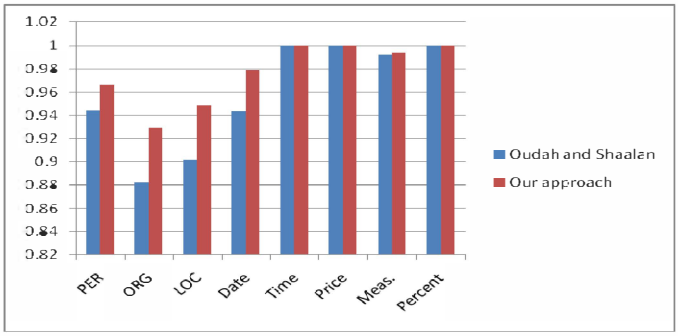

Figure 7. Comparing best results of (Oudah and Shaalan, 2012) with our best results

V. CONCLUSIONS

Our proposed hybrid NER approach integrates the rule-based approach with the ML-based approach in order to optimize overall performance. The two components responsible for the integrated approach are processed in parallel. A tag selection and correction component is used in order to fine-tune the machine learning system's output by checking the most false negatives (i.e., missing annotations) and applying the correction using the tagging decisions determined by the rule-based component.

Experimental results show that the proposed hybrid system achieves a better performance than when applying each of the two components separately and outperforms the state-of-the-art of the Arabic hybrid NER system. Our study on the impact of the features indicates that when window size is 3 it achieves the best performance. For future work, the authors would like to increase the capability of the system in identifying other types of named entities. With regard to machine learning component, we expect to use large annotated corpora which cover a large number of named entities and add new features to the feature set. We are also considering the possibility of investigating different machine learning techniques other than SVM and study their impact on the overall performance of the hybrid NER system. On the other hand, we aim to enhance the results of rule-based component for recognizing Location, Person and Organization named entities by improving both grammatical rules and gazetteers. Grammatical rules will be improved through revising the current rules and adding new ones as well as gazetteers will be updated and expanded automatically In order to reduce the cost and effort of manual expansion.

REFERENCES

[1] Mohamed A. Meselhi, Hitham M. Abo Bakr, Ibrahim Ziedan, and Khaled Shaalan, "A Novel Hybrid Approach to Arabic Named Entity Recognition," in *Proceedings of the 10th China workshop on Machine Translation (CWMT),* in press.

[2] Yassine Benajiba, Mona Diab, and Paolo Rosso, "Arabic named entity recognition using optimized feature sets," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 284–293, Honolulu, October 2008.

[3] Alireza Mansouri, Lilly Suriani Affendey, and Ali Mamat, "Named Entity Recognition Using a New Fuzzy Support Vector Machine," *IJCSNS International Journal of Computer Science and Network Security*, February 2008.

[4] Khaled Shaalan and Hafsa Raza, "NERA: Named entity recognition for Arabic", Journal of the American Society for Information Science and Technology, pp. 1652–1663, 2009.

[5] Imad A. Al-Sughaiyer and Ibrahim A. Al-Kharashi, "Arabic morphological analysis techniques: a comprehensive survey," Journal of the American Society for Information Science and Technology, 55(2004), pp. 189–213, 2004.

[6] Khaled Shaalan, "A Survey of Arabic Named Entity Recognition and Classification," Computational Linguistics, 40:2, MIT Press, 2014.

[7] Benjamin Farber, Dayne Freitag, Nizar Habash and Owen Rambow, "Improving NER in Arabic Using a Morphological Tagger," in *Proceedings of LREC'08*, Marrakech, 2008, pp. 2,509–2,514.

[8] Nizar Habash, "Introduction to Arabic Natural Language Processing," Mogran & Claypool Publisher, 2010.

[9] Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter, "On Arabic transliteration," In Antal van den Bosch and Abdelhadi Soudi, editors, Arabic Computational Morphology: Knowledge-based and Empirical Methods. Kluwer/Springer, 2007.

[10] Yassine Benajiba, Paolo Rosso, and José-Miguel Benedí, "ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy," in *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2007)*, Berlin, 2007, pp. 143–153.

[11] Mai Oudah, and Khaled Shaalan, "A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach," in *Proceedings of the International Conference on Computational Linguistics* , Mumbai, 2012, pp. 2,159–2,176.

[12] Andrew Borthwick, "A Maximum Entropy Approach to Named Entity Recognition," Ph.D. thesis, Computer Science Department, New York University, 1999.

[13] Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel, "An Algorithm that Learns What's in a Name," Machine Learning, 34(1-3):211-231, 1999.

[14] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," In ICML, pages 282-289, 2001.

[15] Andrew McCallum, and Wei Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons," in *Proceedings of Seventh Conference on Natural Language Learning, CoNLL 2003*.

[16] Vladimir Vapnik, "The Nature of Statistical Learning Theory," Springer Verlag, 1995.

[17] Yassine Benajiba, Mona Diab, and Paolo Rosso, "Arabic Named Entity Recognition: An SVM-Based Approach," in *Proceedings of Arab International Conference on Information Technology, ACIT-2008*, 2008, pp. 16-18.

[18] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 168-175.

[19] An De Sitter, Toon Calders, and Walter Daelemans, "A formal framework for evaluation of information extraction," University of Antwerp, Department of Mathematics and Computer Science, Technical Report TR 2004–0, 2004.