Recognition System for Libyan Entity Names

Abdelsalam A. Almarimi and Ezzedin M. Enbiah

Abstract — Named Entity Recognition (NER) is a computational linguistic concept that is used to find and classify appropriate nouns in a text such as person names, geographical locations, and organizations. Such a concept is fundamental in the field of natural language processing. In Libya, many private and public institutions suffer from using the proper translation of entity names from Arabic language into English. Therefore, in this paper, we are concerned with analyzing Arabic articles to extract and recognize entity names. A recognition system is developed for recognizing names of persons, academic institutions, and cities in Libya. At first, a training corpus and dictionaries are built for the intended entity names in this research. Then, the aspects of the entity names are studied, and their patterns and rules are designed. Then, the implementation is performed using Nooj linguistic language. The recognition of person names and Libyan cities and academic institutions was carried out. Statistics showed the frequencies of the appearance rate of person names, academic institutions, and cities in our training corpus. The obtained results are promised and met the research goals for tackling the problem of Arabic named entity recognition.

Index Terms — NLP, Arabic Named Entities, Named Entities Recognition, Translations, Transliterations, Nooj Linguistic Platform.

I. INTRODUCTION

Natural language processing (NLP) is an area that can be used to realize and process natural language text to accomplish valuable tasks by computers. The aim of NLP comprises building of real computational applications to explore the features of natural language used by human and to develop systems including the smart processing of human written language via computer. It deals with both concept and uses for a variety of applications. NLP still has a lot of interest in software development and research. The most common applications of NLP comprise information retrieval and extraction, machine translation, question answering, data classification and summarization, dialogue applications, and named entities recognition (NER) [1].

NER can be defined as the task that attempts to find and classify named entities into predefined classes or types in open-domain and unstructured texts, such as newspaper articles. It is a part of data classification applications in specific fields such as sport, agriculture, medicine, etc.

Basically, the NER is a subtask of information extraction which involves the identification and classification into a set of predefined sorts of importance. In NER, the expression “named entity” covers not only proper names but also includes temporal expressions and some numerical expressions to extract time, date, and address expressions from letters. Named Entities occur frequently in Arabic texts, and their recognition is essential. Recognizing and categorizing entity names and translating them properly from Arabic to English. We have used Nooj linguistic language for developing the proposed system [7].

The remainder of this paper is organized as follows. Section II introduces the related work. Section III presents the architecture of the proposed transliteration system. Section IV shows the experimental results. Section V presents the evaluation of proposed system. Finally, conclusions and future work are presented in Section VI.

II. RELATED WORK

Research in natural language processing has a lot of interest in software development and research. Named Entity Recognition (NER) has become a motivating research area. In the nineteen’s, in particular at the Message Understanding Conferences, Named Entity Recognition was first introduced as an information extraction task and deemed important by the research community [8].

A NER system is considered an important preprocessing tool for tasks such as document classification of clustering, machine translation, information extraction, information retrieval tasks, indexing and search and other text processing applications. It is very valuable for many applications such as web mining and information retrieval. Such applications are used for recognizing and retrieving relevant information from a set of data which matched to an input query [9].

Several research works have been achieved using local grammars as finite state local automata have been used to recognize person names in textual documents [10], [11]. Reference [12] introduced a rule-based approach for
recognizing Arabic person names. Traboulsi in [13] discusses the use of corpus linguistics methods and techniques, in conjunction with the so-called local grammar formalism to identify patterns of person names in Arabic news texts. Shaalan in [14] has presented a review for Arabic NER and classification and describing the recent activities and growth regarding Arabic NER study and the importance of Arabic NER features of languages are emphasized. Likewise, Abdallah in [15] has defined a local grammar as a way of describing syntactic restrictions of certain subsets of sentences.

III. PROPOSED MODEL

The architecture of the proposed system is shown in Fig. 1. It is implemented using Nooj linguistic language.

![Fig. 1. The Architecture of the Proposed System.](image)

The main components of the proposed system consist of:
1) A Nooj corpus is built and used as a storage of the related Libyan data and articles which we have collected. The data is represented in the corpus as a set of annotations. Such a corpus was extended by implementing a specific web crawler that we have developed.
2) Lexical resources are designed and implemented. They comprise dictionaries and morphological grammars. A dictionary of a language contains all of the lemmas (entries) associated with possible meaning. Fig. 2 shows sample of person names and cities’ dictionary.
3) Morphological grammars are a finite state automaton that reads (recognizes) sequences of letters of a word and associate them with linguistic information.
4) Syntactic grammars are set of patterns designed and represented using Nooj’s linguistic engine as Transducers for recognizing the intended entity names.
5) Concordance tables are used for the output to display the matched results of the applied grammar. A concordance table as shown below can be contained four columns, the first column shows the source file of the result (Text), the second, and fourth columns (After, Before) show 5 words before and after recognized entity, and the third column (Seq.) displays the matched result.
6) RDBMS stands for Relational Database Management System is MySql server that used for storing the obtained results of concordance tables.
7) The web interface is the interactive screen that enabled users to search the translation/transliteration of the recognized entity names by querying MySql database using PHP code.

![Fig. 2. Sample of person names and cities’ dictionary.](image)

IV. DESIGNING THE LINGUISTIC GRAMMARS

The intended entity names are classified as formalized patterns. These patterns are designed or expressed as: three basic patterns for person names, cities, and institutions. These patterns are represented with grammars using Nooj.

A. Designing the Grammar of Person Names

Fig. 3 shows the designed recognition grammar for person names. This grammar is a finite state graph consists of set of nodes that starts with initial state, followed by transition states, and ends with final state. Syntactic grammar reads text word by word (sequence of tokens). The \(<N+FirstName>\) is a transition state represents a rule that the token is a noun and a first-person name conforms an entry in the person name dictionary.

To improve the performance of the local grammars, we added new local grammars as trigger (TriggerParent) to involve the relative relations of persons such as ‘Son of:’. Also, we added another local grammar as trigger (TriggerPerson) to involve the adjectives of person names such Eng., Dr., Prof., Mr., Mrs, etc.

![Fig. 3. The Grammar of Person Names.](image)

B. Designing the Grammar of Cities

Recognizing city names is based on the designed linguistic grammar and dictionary. Fig. 4 shows cities recognition grammar. The grammar can recognize (City: Municipality: Municipal Council:)

DOI: http://dx.doi.org/10.24018/ejce.2020.4.6.263
C. Designing the Grammar of Academic Institutions
Recognizing academic institutions is based on the designed formal linguistic patterns for the academic institutions in Libya. The academic institutions are classified into three patterns: universities, faculties, and higher institutes. For example, the patterns of faculty name are classified into 3 patterns as follows:

\[
\text{Pattern1}::= \text{Faculty} \text{N-Speciality} / / \text{Name} \\
\text{Pattern2}::= \text{Faculty} \text{N-Speciality} \text{Conj} \text{N-Speciality} / / \text{Name} \\
\text{Pattern3}::= \text{Faculty} \text{Trigger} \text{ADJ-Speciality} / / \text{Name}
\]

Building the local grammars of our designed patterns as shown in Fig. 5 which represent the main graph (Transducer). We classified institutions into University, Faculty, Higher Institute, and Research Center. Fig. 6 shows a sample of academic organization recognition grammar.

A. Recognition of Person Names
Fig. 7 shows sample of the obtained results of the recognized person names.

B. Recognition of Libyan Cities
Recognition of Libyan cities are obtained by applying the designed grammars and cities’ Nooj dictionary on the created corpus. Fig. 9 shows sample of the obtained results of the recognized Libyan city person names.
C. Recognition of Libyan Academic Institutions

Recognition of Libyan academic institutions are obtained by applying the designed grammars and org’ Nooj dictionary on the created corpus. Fig. 10 shows a sample of the obtained recognized results.

VI. SYSTEM EVALUATION

For the evaluation and performance of our designed system, we have statistically showed the appearance rate of person name, academic institutions, and cities frequencies in our training corpus. Fig. 11, 12 and 13 show such rate respectively.

The evaluation metrics we have used are the well-known information retrieval measures: (1) recall, (2) precision and (3) f-measure. While the recall measures the quantity of relevant responses of the system compared to the ideal number of responses, the precision measures the number of correct recognitions, and f-measure is a combination of precision and recall for penalizing the very large inequalities between these two measures and computed by the formula 

\[
(2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})
\]

The precision was computed to measure the accurate of the obtained results by computing the number of true results which represent the correct recognition. Table I shows the true recognition of sample of person names and Libyan cities that are used for testing our system.

![Fig. 10. Sample of Academic Institution Recognition Grammar.](image1)

![Fig. 11. The appearance rate of person names.](image2)

![Fig. 12. The appearance rate of Academic Institutions.](image3)

![Fig. 13. The appearance rate of Libyan Cities.](image4)

TABLE I: THE PRECISION OF TRUE RESULTS

|                        | Sample Total | True Results | Precision |
|------------------------|--------------|--------------|-----------|
| Person Names           | 730          | 688          | 94.25%    |
| Cities                 | 1406         | 1405         | 99.93%    |
| Academic Institutions  | 87           | 78           | 89.66     |

The values obtained in the evaluation of recall and f-measure are shown in Table II. The lack of precision of academic institutions is related to some institutions are not Libyan.

TABLE II: THE PRECISION OF TRUE RESULTS

|                        | Precision | Recall | f-measure |
|------------------------|-----------|--------|-----------|
| Person Names           | 94.25%    | 93%    | 93.62%    |
| Cities                 | 99.93%    | 98%    | 98.95%    |
| Academic Institutions  | 89.66     | 83%    | 86.2%     |

VII. CONCLUSION

In this paper, we introduced a system for recognizing Arabic person names and Libyan cities and academic institutions from electronic news texts. The contribution of this paper is to provide an approach for recognizing Libyan entity names. We used Nooj linguistic Environment to design and implement our recognition model. In our experiments, we found that the recognition of a person names may lead into wrong result in the case of names are adjectives such as “Saeed: سعيد”, “Intisar: إنتصار”. The result of cities recognition is very accurate due to the use of our designed dictionary which contains all Libyan cities. Ambiguity can be obtained such as “Bishir city: مدينة بشر” which can be denoted for “Human: بشر” meaning or Bishir city, “Nisma city: نسمة” which is denoted also for people “Nasama: نسمة”.

DOI: http://dx.doi.org/10.24018/ejece.2020.4.6.263
One significant reason that has importantly influenced the above realized results is the non-standardization of writing Arabic manuscript. Most of them are free loaded with variations forms of Arabic script. Standard can support reach for more accurate results. The developed system was evaluated on a testing corpus and the obtained results were very promising. In the future, we are going to achieve the following:

1) Extending our corpus for large coverage of recognizing the intended names by our developed crawler system to extract texts of daily Libyan magazines to be stored as XML files and then added to the corpus.

2) Extending the designed grammars to be used for translating the intended entity names from Arabic into English.

3) Developing a friendly GUI for using the proposed system.

ACKNOWLEDGMENT

This work was supported by funding from Libyan Authority of Natural Science Research and Technology, Ministry of Education, Libya.

REFERENCES

[1] S. Bird, E. Klein, and E. Loper, “Natural Language Processing with Python”, Published by O’Reilly Medi, 2009.

[2] N. Y. Habash, “Introduction to Arabic Natural Language Processing”, Morgan & Claypool, vol. 10, 2010.

[3] S. Mesfar, “Named entity recognition for arabic using syntactic grammars”, In Natural Language Processing and Information Systems, Springer, pp. 305-316, 2007.

[4] R. Salah and L. Zakaria, “A Comparative Review of Machine Learning for Arabic Named Entity Recognition”, International Journal on Advanced Science, Engineering and Information Technology, vol.7, 2017.

[5] K. Shaalan and H. Raza, “NERA: Named entity recognition for Arabic”, Journal of the American Society for Information Science and Technology, vol. 60, pp. 1652-1663, 2009.

[6] W. Zaghouani, “RENAR: A rule-based Arabic named entity recognition system”, ACM Transactions on Asian Language Information Processing, vol. 11, p. 2, 2012.

[7] S. Max, “NooJ Manual”, 2003. Available for download at: www.nooj4nlp.net (Accessed in Nov, 2019).

[8] N. Chineho, “MUC-7 named entity task definition version 3.5”, In Proceedings of the Seventh Message Understanding Conference, Washington, DC, 1997.

[9] J. Baptista, “A local grammar of proper nouns”, In Proc. Seminarios de Linguistica 2, Faro: Universidade do Algarve, 1998, pp. 21-37.

[10] N. Fruberger and D. Maurel, “Finite-state transducer cascades to extract named entities in texts”, Theoretical Computer Science, vol. 313, no. 1, pp. 94-104, 2004.

[11] J.-S. Name and K.-S. Choi, “A Local Grammar-based Approach to Recognizing of Proper Names in Korean Texts”, In Proc. Workshop on Very Large Corpora, ACL, Tsing-hua Univ. and Hong-Kong University of Science and Technology, pp. 273-288, 1997.

[12] M. Abosaqa and M. Ab Aziz, “Arabic person names recognition by using a rule-based approach”, Journal of Computer Science, vol. 9, p. 922, 2013.

[13] H. Traboulsi, “Named Entity Recognition: A Local Grammar-based Approach”, Ph.D. dissertation, Dept. of Computing, Surrey Univ. Guildford, U.K, 2006.

[14] S. Abdallah, K., Shaalan, M. Shoaib, “Integrating Rule-Based System with Classification for Arabic NER’,”, pp. 311–322. Springer-Verlag, Heidelberg, 2012.

[15] K. Shaalan, “A survey of Arabic named entity recognition and classification”, Computational Linguistics, vol. 40, pp. 469-510, 2014.