University of Zawia

Faculty of science

Department of computer science

Research Proposal

Project title

Named Entity Recognition for Libyan Dialect using a Machine Learning Algorithm

Prepared by: Lila Saad Arebi.

Supervised by: Dr.Ramadan Alsayed Alfaered

**Named Entity Recognition for Libyan Dialect using a**

**Machine Learning Algorithm**

**Abstract**—The difficulty of locating and categorizing Named Entities (NE) in Arabic texts written in the Libyan dialect is examined in this proposed. A brand-new lexicon will be created with named entity seeds taken from the Twitter social media network corpus in the Libyan dialect. A statistical model (Machine Learning) applies to obtain this lexicon that will be useful for many NLP applications.

**Introduction**

In the recent years, Named Entity Recognition (NER) becomes very popular and has been a topic of active research area in natural language processing [1][2]. Many researches on Arabic NER in the several domains are published [3], [4], [5]. However, little work has been done on Arabic dialects [6]. NER is a subtask of information extraction that classifies proper names from unstructured texts into categories of names. These names can be personal, geographic named expressions, company or organization as well as temporal and numeric ones. The main idea of NER is to improve the performance of many natural language processing applications such as, Machine Translation (MT) [7], Information Retrieval (IR) [8] and Question Answering QA[9].

In general, NER systems have used three techniques [10], [11]: linguistic grammar-based techniques, statistical models (i.e. machine learning) and hybrid method (a rule-based and a statistical model). The rule-based model uses a set of morphological rules to identify a NE. On the other hand, the statistical model is based on association rules and it is built by employing unsupervised learning of context patterns that indicate the presence of a person name, company and location. The data collection will be obtained from Twitter. Therefore, twitter is a social media network website, where, its users can post their tweets.

**Related Work**

There are several different approaches to constructing Named Entity Recognition tools, First is based on simple machine learning strategies such as Support Vector Machines (SVM), Conditional Random Fields (CRF), or Random Forest (RF), where the algorithm is trained on a labeled dataset and tasked with classifying new tokens based on statistical trends ([12] ; [13]). Deep learning methods employ a similar principle, but introduce vastly more complex architecture with far more sensitive capacity for capturing latent trends. Models based on bidirectional transformer stack architecture. In the paper presented by [14]; [15] have proven to be very promising, but a number of other designs including Pooled-GRU and CNN deserve to be examined. While deep learning systems are more powerful and accurate, they tend to have higher computational requirements and very long training times, which limit the extent of their practical usefulness [16]. Another are used grammatical rule to recognize named entities, for example using genitive rules to discover multiword entities. It's also possible to construct knowledge bases that include gazetteers of named entities and semantic interpretation frameworks (ontologies) and use those resources to improve the recognition process [17]. The paper[6] presents the problem of finding and classing named entity in Arabic Libyan Dialect Text. A rule-based model is used for extraction these entities. The results of this study obtained are 8583 recognizing named entities which are (71.59%, 26.13% and 02.27%) in (person, location and company) respectively.

**Motivations**

Machine learning has been used in numerous earlier researches to analyze Arabic dialects. However, Named Entity Recognition in the Libyan dialect did not employ this method. This study closes the gap. I have especially looked into this study proposal to fill in this gap by providing a training dataset for machine learning algorithms using texts from social media.

**Objectives**

1. Creating a dictionary for the Libyan dialect in order to build an integrated system for named entity recognition by using one of the machine learning methods.

**Problem Statement**

We can assess, enhance, forecast, and reduce the incidence of various problems and then make judgments to solve them by using one of the machine learning methods.

**Methodology**

* Data gathering (tweets, dialect vocabulary(
* Pre-processing data
* Use a machine learning algorithm to analyze the data and evaluation our results

**Timeline**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Months  phase | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Background and reading |  |  |  |  |  |  |  |  |  |  |  |  |
| System design |  |  |  |  |  |  |  |  |  |  |  |  |
| Evaluation |  |  |  |  |  |  |  |  |  |  |  |  |
| Writing up |  |  |  |  |  |  |  |  |  |  |  |  |

**References**

[1] X. Liu, N. Yang, Y. Jiang, L. Gu and X. Shi. (2020). “A parallel computing-based deep attention model for named entity recognition,” *Journal of Supercomputing*, vol. *76*, no. *2*, pp. 814–830.

[2] M. Xiaofeng, W. Wei and X. Aiping. (2020). “Incorporating token-level dictionary feature into neural model for named entity recognition,” *Neurocomputing*, vol. *375*, no. *6*, pp. 43–50.

[3] M. Abdul-Mageed, A. Elmadany, and El M. Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. arXiv preprint arXiv:2101.01785.

[4] M. Alkhatib and K. Shaalan. 2020. Boosting arabic named entity recognition transliteration with deep learning. In The thirty-third international flairs conference.

[5] N. Alsaaran and M. Alrabiah. 2021. Classical arabic named entity recognition using variant deep neural network architectures and bert. IEEE Access, 9:91537–91547.

[6] Ramadan Alsayed Alfared, Hussain Alhammi , Libyan Dialect Named Entity Recognition On Rule-Based Approach, The Libyan International Conference on Electronic Engineering and Technology, LICEET 2018, Tripoli, Libya.

[7] Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, and Conghu Yuan. 2018. An Empirical Study of Machine Translation for the Shared Task of WMT18. In Proceedings of the Third Conference on Machine Translation, Brussels, Belgium. Associ ation for Computational Linguistics

[8] K. Roberts. TREC-Covid: rationale and structure of an information retrieval shared task for covid-19. J. Am. Med. Inform. Assoc. 27, 1431–1436 (2020).

[9] Tang, R. et al. Rapidly bootstrapping a question answering dataset for COVID-19. Preprint at arXiv:2004.11339 (2020).

[10] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, pages 9–15 (LREC 2020), Marseille, 11–16 May.

[11] Wissam Antoun, Fady Baly, and Hazem Hajj., Araelectra: Pre-training text discriminators for arabic language understanding., In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 191–195 Kyiv, Ukraine (Virtual), April 19, 2021

[12] Marwa Muhammad, Muhammad Rohaim, Alaa Hamouda, and Salah Abdel- Mageid. 2020. A comparison between conditional random field and structured support vector machine for arabic named entity recognition.

[13] Mohammad Hudhud, Hamed Abdelhaq, and Fadi Mohsen. 2021. Arabianer: A system to extract named entities from arabic content. In ICAART (1), pages 489–497.

[14] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 9–15.

[15] Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. arXiv preprint arXiv:2101.01785

[16] Chadi Helwe and Shady Elbassuoni. 2019. Arabic named entity recognition via deep co- learning. Artificial Intelligence Review, 52(1):197–215.

[17] Hussein Khalil, Taha Osman, and Mohammed Miltan. 2020. Extracting arabic composite names using genitive principles of arabic grammar. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(4):1–16.