HMM-Based Dari Named Entity Recognition For Information Extraction

Ghezal Ahmad Jan Zia and Ahmad Zia Sharifi

[1] Department of Models and Theory of Distributed Systems,
Technical University of Berlin, Berlin Germany,
zia@campus.tu-berlin.de
[2] Department of Networking and Communications,
Nangarhar University, Nangarhar, Afghanistan
sharifi@nu.edu.af

Abstract. Named Entity Recognition (NER) is the fundamental subtask of information extraction systems that labels elements into categories such as persons, organizations or locations. The task of NER is to detect and classify words that are parts of sentences. This paper describes a statistical approach to modeling NER in Dari language. Dari and Pashto are low resources languages, spoken as official languages in Afghanistan. Unlike other languages, named entity detection approaches differ in Dari. Since in Dari language there is no capitalization for identifying named entities. We seek to bridge the gap between Dari linguistic structure and supervised learning model that predict the sequences of words paired with a sequence of tags as outputs. Dari corpus was developed from the collection of news, reports and articles based on the original orthographic structure of the Dari language. The experimental result of named entity recognition performance presents 94% accuracy.

Keywords: Natural Language Processing (NLP), Hidden Markov Model (HMM), Named Entity Recognition (NER), Part-of-Speech (POS) Tagging.

1 Introduction

Named Entity Recognition (NER) is the fundamental subtask of information extraction systems. The task of NER is to classify words that are part of sentences into predefined sets of categories such as names of the persons, organizations, locations, etc[1].

Named Entities (NE) refers to every entity that are recognized with a proper name. Named Entity Recognition is useful for Information Extraction (IE), Question Answering (QA), Information Retrieval (IR), and etc [2][1]. NER is executed using a rule-based approach, this approach uses hand-written rules by human (linguist) or Machine Learning (ML) various approaches as Hidden Markov Model (HMM), Maximum Entropy (ME), Conditional Random Field (CRF), and etc. These approaches learn rules from large trained datasets (corpora) [3][5]. NER is primarily discussed in 1995 by the Message Understanding Conferences (MUC-6)[6].

Dhinaharan Nagamalai et al. (Eds) : SAI, NCO, SOFT, ICAITA, CDKP, CMC, SIGNAL - 2019
pp. 73-81, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90706

Consequently, they defined the tasks as: Entity Name Expression (ENAMEX) for the proper names, Time Expression (TIMEX) for temporal expressions of times and dates, Numeric Expression (NUMEX) for numeric expression of heights, monetary expression, percentage, and etc [7].

In this paper, we considered the ENAMEX that operates on the classification and extraction of proper names in the Dari language. Initially, the following tags that are used as a class labels in this paper are [7]:

- ORG (Organization): government or another civil organizational unit
- LOC (Location): place names
- PER (Person): person names
- O (Other): for everything else

In this paper, We considered on the role of NER under the domain of IE that extracts the relevant information from large unstructured texts. We have adopted the gap between Dari linguistic structure and supervised learning model that predict the sequences of words paired with the sequence of tags as outputs. The development of HMM-based Dari NER system is based on HMM which is one of the significant model of NLP [1]. It is a probabilistic model used to detect and classify NE from unstructured texts into relevant categories [8]. The unstructured texts are the source of information that contains proper nouns in texts. The unstructured texts are processed as an input. The system is locating NEs and their classes from the input. Considering the complexity of Dari language, the mechanism is to tokenize the texts into tokens. Initially, these tokens are nouns, adjectives, prepositions, verbs, article and etc. Now, the task of the system is to detect and classify proper nouns from these tokens. These proper nouns present names of a person, a place, or an organization [2].

Dari and Pashto are the official languages, spoken in Afghanistan. There has been a considerable amount of research on NER problem for other languages. This paper presents the first NLP application that we have developed for the Dari language.

2 Related work

Machine learning based NER can be classified into supervised learning that is based on labeled data or unsupervised learning that combines labeled data and unlabeled data. In supervised learning, most of the NER task is represented by HMM. On the stemming process, the first order of HMM based is used by Fadl which solves Arabic's inflection problem and ambiguity, and through that they achieved a combined precision and recall score of 77% and 73% respectively [7]. Chopra's HMM-based NER on Indian language present more accuracy than Morwal's, and obtained 97.14% accuracy [9].

Using optimized feature sets Yassine shows F-measure 83.5% in Arabic [10]. Similarly, a discriminative machine learning framework, named as Support Vector Machines is used, and implemented on the Arabic to define sets of features which are both language independent and language specific with F1=82.71 [11].

We searched to find systems being developed for the processing of the Dari language text. A rule-based system combining machine learning approach has been developed by [12] that integrated dictionaries of Persian named entities, Persian grammar rules and a Support Vector Machine (SVM). The system shows 90% accuracy. The approach that refers to entire dictionaries to locate proper nouns would not be an appropriate way to tackle NER problems; as ambiguous tokens/words that fall in this category are more likely to be used as non-proper nouns in the text. Consequently, however, there are some similarities between Persian and Dari but according to the orthographic, morphology and phonology structures, there is a lot of differences. For example, there are words exist in Dari but not in Persian, or the phonology in Persian is median but in Dari such a phonology is not existed such as the word برزیدن /barzidan/ which does not exist in Dari [4].

3 Linguistic Issues and Challenges

Dari language is a resource scarce language that has rich morphology and complex orthographic structure that makes it a highly inflected language, with the existence of affixes, suffixes as well as its ambiguity [4]. The figure 1 shows the Dari language morphology consisting of many interconnecting parts. Dari is a complex language to deal with, considering its orthographic features such as different word spellings and different word usages.

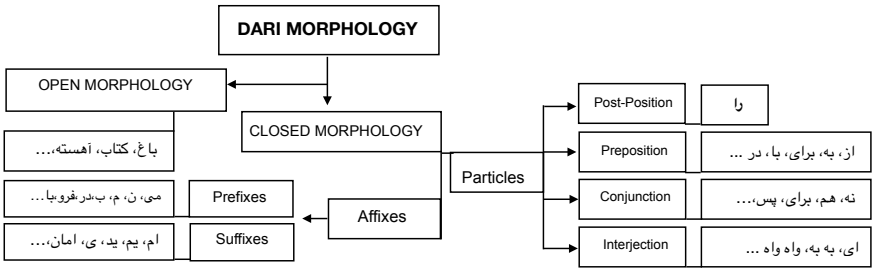

Fig. 1: Dari Morphology

Different levels of ambiguities are in Dari which complex morphology is one of the major challenges in detecting NEs. For instance, considering the following text that shows the name of an organization: تیم ملی کرکت /Team/Mili/Cricket/ (Cricket National Team). If we split these words into individual words, in results

there will be a different meaning. In this example, the word تیم /Team/ (team) is a noun and ملی /Mili/ (national) is an adjective, and it does not refer to NEs. But when it is used with the word کرکت /Cricket/, it refers to an organization.

To consider another example that one word implies different meanings and match with verb and nouns. The Dari word نیاز /niaz/ (need) might lead trigger words to denote three different forms. In this example نیاز /Niaz/ refers to a person name, such as نیاز محمد /Niaz/ /Mohammad/, or a verb من به یک کتاب نیاز دارم /man/ /ba/ /yak/ /kotab/ /niaz/ /daram/ (I need a book) or a trigger word for a noun. Moreover, the lack of standardization leads to variants of the same word with different spelling such as for the word "Spain" that refers to a location with different spelling هسپانیا /haspania/ or اسپانیا /espania/. Further, to consider another example, the word "Ismael" (Person name), can have different orthographic nature: اسماعیل or اسمعیل. Therefore, many other ambiguous words exist in the Dari language with different orthographic nature and structure that give rise to a conflict situation. Thus, in order to correctly target the relevant texts to answer the input queries for the NE, it needs a correct classification for the NE.

Dari language does not support capitalization to indicate a word or sequence of words as a NE. Thus, NER tasks would require a pre-processing of the data which is more challenging.

4 Building The Required Resource For NER

We earlier mentioned that Dari language is a low resource language. Therefore, the supervised learning approach needs annotated training data. There is no free corpus available to test and train the model for the NLP tasks in Dari language. Thus, we have developed a corpus oriented towards the newswire domain that contains 85K Dari words. We collected the data from three main sources that follow Dari pure orthographic structure. These sources daily publish the political, sports and information about the society. These main sources are the Voice of America (VOA Dari), Azadi Radio and for the person name, we used the data of Kankor (university entry national exam) from the Ministry of Higher Education of Afghanistan [14][15][16]. In addition, before starting manually labeling the data, we performed the pre-processing operation on the collected data to clean and tokenize the data. Table 1 shows our developed corpus by the ENAMEX based categories of class labels for the task of NER.

Table 1: Distribution of The Data

Dari NER Corpus			
Locations	Organizations	Persons	Other
2565	631	66012	17348

5 System Architecture

Our goal is to build a model which in input, it has a sequence of Dari words; and in output, a sequence of words annotated with NEs. To simplify the complexity of Dari language for the task of NER and to easily adapt it to the HMM model, we propose a system architecture as shown in Figure 2, that includes the following processes. We have an input file as a development data which is a series of sentences separated by one word per line. The training data file has a series of sentences separated with an empty line, one word and tag per line, separated by a space.

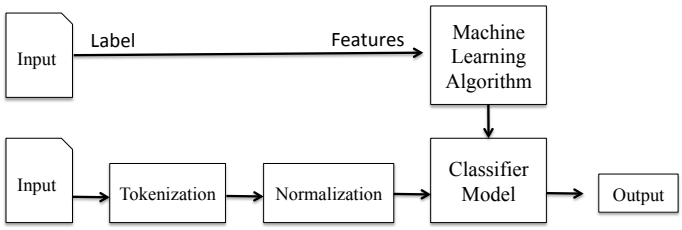

Fig. 2: System Architecture *(HMM Model)*.

5.1 Tokenization

In order to define the word boundaries and to clean the input data, we included the tokenization operation to the system process. To locate a NE in the text, we have to divide the sequence of texts into individual words tokens and separated by one word per line. In the tokenization process, we faced with words that have spaces between characters such as the word زنده گی /zinda gi/ (life). To tackle this problem, we used the library of NLTK (Natural Language Toolkit) with regex to tokenize the sequence of words. Consequently, for words with space between their characters, we used the regular expression to count it as single words. Table. 2, example of an ENAMEX based Dari corpus represented with labels. Giving such annotated data to a model like HMM or neural network can be trained to label new sentences.

5.2 Normalization

Normalization is a necessary phase to all machine learning systems to predict NERs properly. The objective behind this process is to give a particular property to all sets of data; In this phase, we tried to normalize the data and clean noises as there are words which are not neatly delimited by the tokenization process.

Table 2: An example of an ENAMEX annotated corpus

ENAMEX			
Dari	BuckWalter	English Trans.	Tags
غنی	Ghani	Ghani	PER
پروژه	prozha	Project	O
تاپی	TAPI	TAPI	ORG
بین	baine	Between	O
ترکمنستان	Turkmanistan	Turkmenistan	LOC
افغانستان	Afghanistan	Afghanistan	LOC
پاکستان	Pakistan	Pakistan	LOC
و	wa	and	O
هند	hind	India	LOC

6 HMM-Based Dari Named Entity Recognition

HMM is a generative probability model of words and hidden states [1]. The task of mapping the sequence of tags $T = t_1, ...t_n$ (length n states) to the sequence of observations (words) $W = w_1, ...w_n$ is also called sequence labeling problem. The HMM model will solve the classification problem in texts by moving through different states based on the transition probabilities in a time interval. The states are hidden (unobserved states) and the observation is shown (observed words). The emission probabilities, also called a sequence of observation likelihoods, that on a states (n) it emits from all sequences of tags for a sentence that maximizes the probabilities of an observation w_n [1].

To present the above explanation based on the bigram model for tagging, $P(T|W)$ that maximizes over sequences of tags and words, can be represented by the equation (1).

$$t_1^n = \text{argmax}_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \tag{1}$$

The two model components, t_i is the sequence of tags and w_i is the sequence of words (observation), $P(t_i|t_{i-1})$ is the prior probability distribution over the tag t and $P(w_i|t_i)$ is the probability of generating the input w given label t [1]. In general, probability of emitting a word (w_i) in a given state (t_i) corresponding to the emission probability and probability of moving from one state (t_i-1) of the model to the next (t_i), which is also called transition probability.

Our NER model is based on an HMM to estimate the $P(T|W)$. Therefore the HMM-based model for the task of NER consists of the following elements [13][1].

- a set of state $S = s_1, s_2, ..., s_N$, where N is the number of state. In the NER task a state is a NE type where in POS tagging it is corresponds to a POS. N is the number of NE type plus 1 and q_t means a state s in the time t.
- A is the transition probability matrix $A = a_{11}a_{12}...a_{n1}...a_{nn}$, here, a_{ij} means the probability of moving from state i to state j as follows: $a_{ij} = P(q_{t+1} = S_j|q_t = S_t)$, 1 ≤ i, ≤ N.
- a sequence of M observations, $V = w_1, ..., w_M$. In the NER, the number of observable symbol is the same as the size of the vocabulary.
- B is a sequence of observation likelihoods, also called emission probabilities, $B = b_j(k)$, defined as: $b_j(k) = P(q_t = w_k|q_t = S_i)$, 1≤ j ≤ N, 1 ≤ k ≤ M
- $\pi = \pi_1, ...\pi_N$ is the initial probability distribution over states. Where, π_i is the probability that the model will start in state i. Thus, some state j may have $\pi_j = 0$, meaning that they cannot be initial states.

In this paper, we used the bigram for the learning phase and detection phase. Therefore, calculating the probability of long sequences are still not effective. In the equation (1), the probability of a word depends only on its tag. The second, the bigram model seeks at the pairs of tags and uses the conditional probability that a tag is only dependent on the immediate previous tag [1]. Using this approach (1), in order to estimate the maximum likelihood probabilities, it is not tractable due to the exponentially large number of state sequences [1]. Finding the most likely tag sequence is exponential in the length of the input sentence. According to the equation (1), the $argmax$ is taken over sequence of tags. Each word will consider all possible tag sequences. There are w^n possible words and there are up to t^n possible tag sequences. Therefore, for longer sentences, this method will be inefficient, because we cannot enumerate all t^n possible tag sequences. To solve this problem, we used the Viterbi algorithm to exploit the independence assumptions efficiently in the HMM.

The Viterbi algorithm is a kind of dynamic programming that considers the most probable (best hidden) tags over all distributed tag sequences for the observation sequences [1]. The Viterbi makes HMM more efficient on the decoding of the NEs. It breaks up the big search problem (equation 1) into smaller subproblems. The data structure used to store the solution of these subproblems is called a trellis [13].

7 Discussion

To demonstrate the model's ability, we tested it with the CoNLL-2003 datasets which includes approximately the same labels schema (ENAMEX) as to Dari dataset. Simply we adjust the model to accept the CoNLL-2003 dataset format and did not include the POS tags in our models. Table 3 presents the score for the CoNLL-2003

dataset with our model that obtained F_1 of 92.4. Furthermore, we created a test dataset of 17K words from Dari dataset and tested with the model that presents F_1 of 93.7. For unseen and unlabeled data, we again tested the model with Dari texts on 5K words. The system reports the scores of 91.8.

Table 4 presents our comparisons with other models for named entity recognition in Dari. There were limitations to support Dari language text by most of the available models. Therefore, to compare our and other models; some models with or without the use of external labeled data such as gazetteers are required. Our model is not designed for any external labeled resources or gazetteers. Similarly, using a discriminative machine learning framework, namely, Support Vector Machines on Arabic language to define sets of features that are both language independent and language specific with $F_1 = 82.71$[11].

Table 3: Scores of HMM-Based Dari NER on CoNLL-2003 and Dari Dataset

Test#	#Datasets	F_1
1	CoNLL-2003	92.4%
2	Dari NER dataset	93.7%
3	5k (unseen and unlabeled Dari text)	91.8%

Table 4: Arabic SVM model indicates trained with Dari text

Test#	#Models	F_1
1	Arabic (SVM)	82.2%

Looking to the tables, the HMM-Based Dari NER is designed and configured that is more dependent on the Dari text other than English. It should be mentioned that the corpus contains equal words with different NEs that causes ambiguities. It is a problem to decide whether a noun refers a person, an organization or a location. For example, حامد کرزی /Hamid/ /Karzai/ was tagged as a person name in one sentence and on the another sentence, it is considered as an organization name (میدان هوایی بین المللی حامد کرزی /Maidan/ /Hawayee/ /Hamid/ /Karzai/ (Hamid Karzai International Airport). Therefore, in such a condition the classifier does not learn well and lead not a perfect performance.

8 Conclusion and Future Work

In this paper, we have developed a NER system using HMM to detect and extract the type of NEs from unstructured text. We also created the first Dari language

dataset that includes 85K words/tags and will be publicly available for further research. The dataset was trained to the model to predict accurately. Moreover, We defined a system architecture that the input data is preprocessed into a predefined set of features. These features are to detect words boundary, defining a word with a space between their characters, removing noises and etc.

The experimental results have been successful in producing a desired or intended result. Moreover, we tested the NER problem using a generative model. However, there is still some ambiguities exist in Dari language that need to be solved in the future. Analyzing the performance using other methods such as Conditional Random Fields (CRF) will be an interesting experiment.

References

1. Jurafsky, Dan, and James H. Martin. Speech and language processing. Vol. 3. London:: Pearson, (2014).
2. Prabhakar, Dinesh Kumar and Dubey, Shantanu and Goel, Bharti and Pal, Sukomal: ISM@ FIRE-2014: Named Entity Recognition for Indian Languages, 98–102, (2014)
3. Sarkar, Kamal:A hidden markov model based system for entity extraction from social media english text at fire (2015)
4. Prof. Dr. Mohammad Hussain Yamini. Dari Grammer, Morphology, Phonology and Syntax, (2015)
5. Morwal, Sudha and Jahan, Nusrat and Chopra, Deepti: Named entity recognition using hidden Markov model (HMM), International Journal on Natural Language Computing (IJNLC), 1,15–23, (2012)
6. Zaghouani, Wajdi. "RENAR: A rule-based Arabic named entity recognition system." ACM Transactions on Asian Language Information Processing (TALIP) 11.1 (2012)
7. Dahan, Fadl, Ameur Touir, and Hassan Mathkour. "First Order Hidden Markov Model for Automatic Arabic Name Entity Recognition." International Journal of Computer Applications 123.7 (2015).
8. Chung, Euisok, Yi-Gyu Hwang, and Myung-Gil Jang. "Korean named entity recognition using HMM and CoTraining model." Proceedings of the sixth international workshop on Information retrieval with Asian languages-Volume 11. Association for Computational Linguistics, (2003).
9. Chopra, Deepti, Nisheeth Joshi, and Iti Mathur. "Named Entity Recognition in Hindi Using Hidden Markov Model." Computational Intelligence & Communication Technology (CICT), 2016 Second International Conference on. IEEE, (2016).
10. Benajiba, Yassine, Mona Diab, and Paolo 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).
11. Benajiba, Yassine, Mona Diab, and Paolo Rosso. "Arabic named entity recognition: An svm-based approach." Proceedings of 2008 Arab International Conference on Information Technology (ACIT). (2008).
12. Dashtipour, Kia, et al. "Persian Named Entity Recognition." Cognitive Informatics Cognitive Computing (ICCI* CC), 2017 IEEE 16th International Conference on. IEEE, (2017).
13. Yun, Bo-Hyun. "HMM-based korean named entity recognition for information extraction." International Conference on Knowledge Science, Engineering and Management. Springer, Berlin, Heidelberg, (2007).
14. Voice of America, Dari, `https://www.darivoa.com`
15. Azadi Radio, Dari, `https://da.azadiradio.com`
16. Kankor Entry Exam, Dari, `https://http://kankor.edu.af`