ADA University
School of Information Technology and Engineering

Senior Design Project

FINAL REPORT

Project Title: Named Entity Recognition for the Azerbaijani Language

Authors:

1. CS Mahammad Nahmadov
2. CS Elmar Karimov
3. IT Fagan Rasulov

Project Advisor: Dr. Samir Rustamov

Baku, April 2022

Table of Contents

Abstract – Named entity recognition (NER) is the process of locating and classifying named entities in a given text. This project focuses on developing a named entity recognition system for Azerbaijani language. Dataset containing 1500 sentences in Azerbaijani was manually collected from Wikipedia AZ due to the lack of structured corpus for the Azerbaijani language, and tokens were labeled using IOB2 tagging. In order to categorize entities and prepare data for training, Person (PER), Location (LOC), Date (DATE) and Organization (ORG) tags were used. Bidirectional LSTM with a CRF layer is applied to use sentence level tag information as well as future and past input features. Thanks to the Bi-LSTM-CRF architecture, the model shows promising results with a high accuracy of 98% for detecting entities.

keywords – named entity recognition; ner; ner for Azerbaijani; Bi-LSTM-CRF tagger

I. Introduction

A. Definition

Named-entity recognition (NER) is one of the most important topics in Natural Language Processing (NLP) that is used to locate and categorize named entities mentioned in textual data into pre-defined groups such as organizations, locations, person names, time expressions, quantities, monetary values, percentages, and so on.

NER's initial target languages are a number of resource-rich languages (e.g., English, Chinese, Spanish). Low-resource languages including Azerbaijani language are increasingly attracting the attention of scholars. The methods for building NER systems have also evolved significantly.

While rule-based methods are still used, many academics are focusing on statistical machine learning. Machine learning (e.g., Conditional Random Fields, Hidden Markov Models, and so on) originally were used first, followed by deep learning (e.g., Recurrent Neural Networks) and other approaches (e.g., transfer learning and knowledge base).

Building systems to recognize named entities require dataset in target language, that is properly labeled in order to achieve high accuracy. Moreover, a good selection of algorithm(s) is also another essential factor.

B. Purpose

The primary purpose of this project is to develop and improve a high accuracy Named Entity Recognition system for the Azerbaijani language. Since Azerbaijani is a low-resource language, there is a scarcity of research done in this field. Moreover, on the basis of this project, businesses operating in Azerbaijani language can automate answering customer queries by automatically detecting important information. In order to implement the most appropriate machine learning algorithm, a thorough investigation into the implementation of machine learning algorithms on NER systems was conducted, and several of them were tested. A large, labeled dataset, as well as the right choice of machine learning approach, are both important in achieving high accuracy for a Named Entity Recognition system.

Our system will also be useful for attracting attention of global organizations in the incorporation of Azerbaijani language to their already existing NLP technologies, in addition to its obvious contribution to local NLP technologies.

C. Project Objectives, Significance, Novelty

The goal of our research is to select the most effective and accurate machine learning algorithm for constructing named entity recognizer for Azerbaijani language with the usage of a large dataset in order to reduce the ambiguity of natural language in Azerbaijani and obtain high-accuracy NER results. In a time when Azerbaijan is developing NLP technologies such as sentiment analysis, voice recognition, and speech-to-text/text-to-speech systems, it is extremely useful to develop a NER solution to assist these technologies work as accurately as possible. The impacts of our named entity tagging method for Azerbaijani as a backbone for these NLP systems are quite important. Previously, there had been little to no research and application of named entity tagging for the Azerbaijani language. This leads to one of the project's major problems: the lack of a structured data corpus for the Azerbaijani language. Therefore, data had to be gathered and labeled manually by our team to prepare dataset to train the model. Consequently, this project will present a high-accuracy state-of-the-art model for tagging of entities in Azerbaijani as a novelty.

D. Problem Statement

One of the most essential concepts in information retrieval is named entity recognition. The practice of extracting relevant and usable information from unstructured raw text sources is known as information retrieval. NER locates and categorizes identified entities included in unstructured text into standard categories such as person names, locations, organizations, and so on, which can simplifiy tasks involving a language in a lot of spheres such as machine translation, semantic analysis and chatbots. Although there are numerous libraries and firmwares for NER available in English, Russian, Spanish, and other widely-used languages, there are quite limited number of resources available in Azerbaijani, necessitating the development of a model to detect named entities inside texts in this language. The primary ways of developing a fully-functional named entity recognition system consist of three approaches: lexicon, rule, and machine learning based approach [12]. To begin, the lexicon-based technique requires domain-specific expertise in order to develop a lexicon or dictionary for named items manually. The lengthy manual work of constructing a lexicon and the importability of a lexicon across domains are also downsides of this strategy. Moreover, maintaining such system requires continual manual involvement and is a time-consuming effort. Named entities in the field of medicine, for example, may not be enough in finance. Furthermore, because there is no available lexicon for NER in Azerbaijani, this method is not taken into account in our project. The second NER approach is rule-based, which recognizes named entities using pre-defined criteria . It keeps track of how many times a word appears as a named entity as well as its total appearance. As a result, the stated threshold aids in classifying whether a word is a named entity or just a plain word. Finally, a machine learning-based method can be used to solve the problem of classifying named entities using multiple methodologies. To train a model for detecting named entities, modern research has used Hidden Markov Models (HMM), Conditional Random Fields (CRF), and Deep Neural Networks (DNN). A dataset that has been clearly annotated, labeled, and is ready to be used is required by this approach. Although creating a fully annotated training corpus takes a long time, the trained model performs effectively in a variety of conditions due to the generalization of scenarios where named entities are present in the training data.

II. Literature Review

Although NER exists for other widely-spoken languages, there is little research in this domain when it comes to the

Azerbaijani language. Since Azerbaijani has Turkic roots, initally, we can explore similar research made on the Turkish language. E. K. Akkaya, and B. Can investigate various text representations of the same trained data in order to look for possible ways to increase accuracy for Turkish in [1]. Additionally, they have used same methods for text in English to illustrate the differences. In [2], authors mainly focus on news data in order to research rule-based NER systems yet experience poor results because of domain improbability. As a low-resource language like Azerbaijani, in [3] Das, D. Ganguly, and U. Garain built a model for solving Named-Entity-Recognition tasks for the Bengali language. Although this implies the used data was scarce, which did not produce effective results, implementing embeddings in addition to utilizing standard classification is suggested by the authors instead. There has been some notable progress for Azerbaijani language only in recent months: a paper published by Mrs. Natavan Akhundova [4] is about a Named Entity Recognition system for the Azerbaijani language that detects named entity tags, as well as any proper nouns without classification. This was produced by using 620631 sentences for training from Wikipedia AZ, and the first approach used by the author was a statistical method that stores capitalized words and their occurrences, then calculates the probability of being a named entity based on this data. The second approach was using Convolutional Neural Networks CNN machine learning algorithm for sequence tagging as proposed by Collobert et al. [5] using spaCy library, which includes functions such as Part-of-Speech tagging, sentence segmentation, lemmatization, NER feature, tokenization, and other components of NLP. While the paper mentions 70% accuracy, which was mainly due to the scarcity of data and the automated labeling of the training data, our plans for this project include starting from complete scratch by building our own corpus for Azerbaijani language and building a system that has higher recognition accuracy.

III. Design Concept

A. Alternative Solutions/Approaches/Technologies

Rule based

One of the methods to build a NER system is a statistically-based rule-based technique in which each capitalized word is saved and the number of times it appears is determined. The procedure is divided into two steps: first, the amount of occurrence for each word is measured; second, if a word is capitalized, indicating that it is a proper noun, the number of occurrences for this word is also computed individually. As a result, we may create a list of all proper nouns, as well as the total number of times they appeared in the dataset and the amount they were used as a named item in the phrase. The rules used in this classification are typically created by humans, which demands a significant amount of manual labor. Because describing a sequence of laws by hand is inflexible, this process should be automated. To automate labeling, the Brill tagger algorithm [6] is used, which determines the set of labeling criteria that best mark the data and reduce NER tagging errors.

Probabilistic Apprach

One of the most well-known methods for named entity recognition is the Probabilistic Approach. Conditional Random Fields, a sort of probabilistic technique, have been utilized in various languages to address not only Entity recognition challenges, but also Parts of Speech tagging [7]. The conditional probability distribution P(y|x) is used by

CRF, a Discriminative Probabilistic Classifier. This algorithm not only recognizes feature relationships, but also adapts to patterns while taking into account future observations. CRF calculates probability of label sequences, given an observation sequence, which are based on arbitrary, non-independent aspects of the observation sequence. It calculates transition probabilities not only for the current transition as well as for the next and earlier tags. Furthermore, the Hidden Markov Model is a generative model that is also utilized for entity recognition [8]. But unlike discriminative probabilistic technique, generative models use the P probability distribution to combine the probabilities (x,y). HMM takes into account two types of probabilities: first, it calculates the likelihoods of a word given its potential tags. Then, to find relative probabilities, each word is taken as a sequence. The Viterbi Algorithm is a dynamic algorithm that is used to find these probabilities.

B. Detailed description of Solutions/ Approaches/Technologies of choice

Train Dataset

Finding a train dataset is tough, especially for languages with limited resources. Considering Azerbaijani is a turkic language, it makes the task a little more difficult because the language is morphologically complex. Words in sentences take on different forms based on where they appear in the sentence. This complicates NER systems' responsibilities since the same term may appear in the training dataset but in a different form in the test dataset. Many research in this field have noted the difficulty of the nature of turkic languages. Azerbaijani, like Turkish, has equal problems finding datasets. The creation of a Named Entity Recognition model for a low-resource language like Azerbaijani is a difficult task because there is a lack of data to work with in order to eventually construct an effective model. The dataset used to train a NER model in this project was collected from Wikipedia AZ.

Labeling

The obtained sentences were split into tokens to include only one word per line to apply IOB2 tagging. The IOB format is typically used for tagging tokens in computational linguistics, especially named entity recognition. For the purposes of this project, person, organization, location and date entities were used and therefore appropriate tags (PER, ORG, LOC, DATE) were used on the chunks, manually. Depending on the position of the chunk in the sentence, B- and I- prefixes were used. In the beginning of every chunk the first prefix is used, whereas relevant chunks that immediately follow the initial chunk is tagged using the latter prefix. All other chunks were labeled as O (Other). The following table illustrates a sample sentence split to chunks and tagged using IOB2 tagging:

Azərbaycan	B-LOC
Respublikası	I-LOC
Prezidentinin	O
13	B-DATE
yanvar	I-DATE
2014	I-DATE
-	I-DATE
cü	I-DATE
il	I-DATE
tarixli	O
Sərəncamı	O
ilə	O
Azərbaycan	B-ORG
Respublikası	I-ORG
Xarici	I-ORG
İşlər	I-ORG

Nazirliyinin	I-ORG
Diplomatik	I-ORG
Akademiyasının	I-ORG
və	O
Azərbaycan	B-ORG
Respublikasında	I-ORG
İnformasiya	I-ORG
Texnologiyaları	I-ORG
Universitetinin	I-ORG
əsasında	O
ADA	B-ORG
Universiteti	I-ORG
yaradılmışdır	O
.	O

Table 1. IOB2 Tagging

Conditional Random Fields (CRFs)

Probabilistic approaches are frequently used in named entity recognition. This apprach uses the frequency of occurrence of specified tag sequence patterns to calculate probability. Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) are the algorithms on which probabilistic approaches are based.

Lafferty et al. (2001) first introduced CRFs as a probabilistic model for segmenting and labeling sequential data in [7]. It is possible to accomplish classification while being aware of the structure's context using CRFs. Because there are generally links between different portions of speech/sentence, this strategy allows the algorithm to generate more accurate predictions (e.g., in Azerbaijani language adverbs usually come before the verb). CRFs have a number of features that distinguish them from HMMs (Hidden Markov Models) or MEEMs, proposed McCallum et al. in [9] which are also used for question-answering systems as well as named entity tagging. CRFs reduce the degree of independency and constraints present in these algorithms by employing sequential patterns to extract the probability of named entity occurrences. CRFs are used in a variety of NLP systems, including Parts-of-Speech (PoS) tagging, named entity recognition (NER), shallow parsing, and a wide range of computer vision applications including but not limited to object recognition.

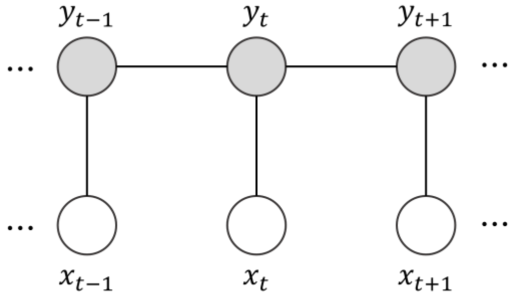


Figure 1. Conditional Random Fields (CRFs)

Recurrent Neural Networks (RNNs)

Deep Neural Network Algorithms are increasingly being utilized to address NLP-related challenges these days. Recurrent Neural Networks, in particular, are a special type of neural networks that are utilized in a wide range of applications, including Parts-of-Speech tagging and Named Entity Recognition. Because of its internal memory capability, this algorithm is one of the most used, and outperforms CRF by 14% according to [10]. RNNs can save crucial data that they process in their internal memory, allowing them to predict subsequent parts. Because in named entity tagging sequences of chunks are valued more individually than individual components, RNN is one of the key approaches that are used.

RNNs have greater advantages than neural networks, also known as feed-forwarding neural networks. In feed-forwarding neural networks, data is not looped and is only sent across layers once. The data never flowes through the same node twice. Feed Forwarding solely considers current input, whereas recurrent neural networks take a different approach.

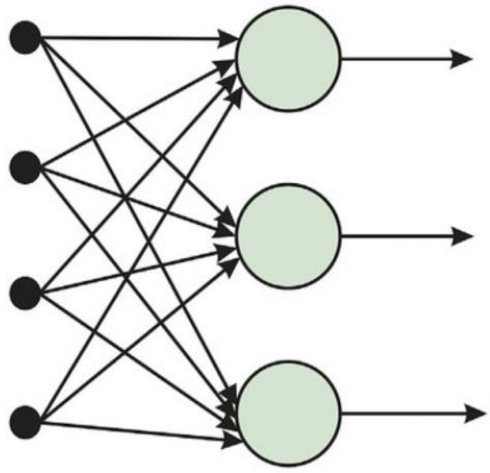


Figure 2. Feed-forwarding Neural Network

RNN cycles all of the network's inputs and learns the pattern faster than traditional neural networks. It takes into account and weights both prior and current inputs. RNN input outputs are mapped to one to many, many to one and many to many. Backpropagation Through Time is one of the crucial concepts in Recurrent Neural Networks. This enables the algorithm to update the weights for each timestamp.

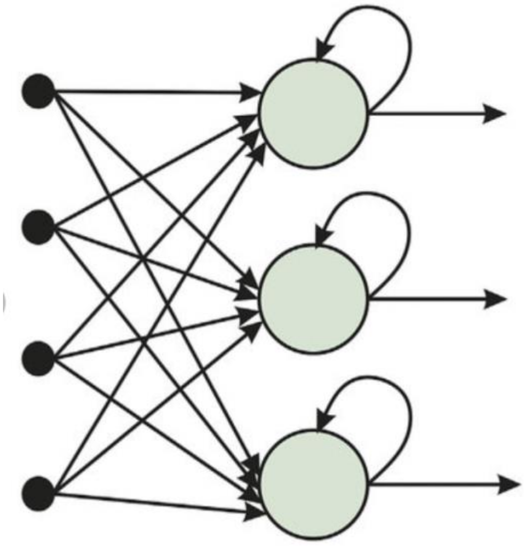


Figure 3. Recurrent Neural Network (RNN)

However, because RNN's short-term memory causes issues, numerous specialized neural networks, such as LSTM, have been developed to address this issue. Long Short-Term Memory (LSTM) is one of them, and it's a more advanced version of Recurrent Neural Networks that deals with extended memory. It is sometimes referred to as a better form of RNN because it learns critical experiences over a lengthy period of time. The LSTM is a gated cell that uses input, forget, and output gates to decide whether to keep or discard information. The sigmoid function in LSTM gates squashes values between 0 and 1.

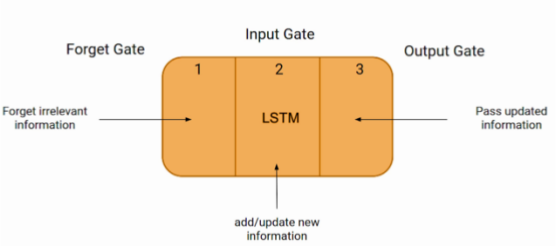


Figure 4. Long Short-Term Memory (LSTM)

The "Forget gate" takes the sigmoid function's value and decides if to preserve or discard it based on that value. Values close to 0 are also discarded:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

The input gate receives the prior hidden state as well as the current input. Both states are sent via the sigmoid and tanh functions before being multiplied and a cell state is created.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$C_t = tanh(W_C[h_{t-1}, x_t] + b_C)$$

The value is transmitted through the output gate, which provides the hidden state for next network.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh\,(C_t)$$

Bi-directional LSTM, proposed by Zhiheng Huang, Wei Xu and Kai Yu in [11] is a superior variant of LSTM that provides greater example prediction rates. Instead of one, two LSTMs are used. One of them provides results based on the word sequence as it is. The second, on the other hand, accepts input in the opposite order.

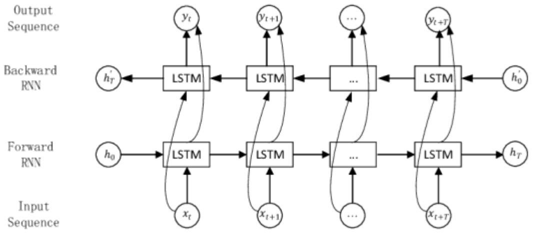


Figure 5. Bidirectional LSTM (Bi-LSTM)

C. Research Methodology and Techniques

A thorough paper search was undertaken on researchgate.net, scholar.google.com, ieeexplore.ieee.org in order to conduct the conceptual literature analysis for the project. The key terms used for these findings include "Named Entity Recognition for agglutinative languages", "Named Entity Recognition for Turkish language", "Named Entity Recognition using LSTM", and "Named Entity Recognition using CRF". The reason is that Azerbaijani is an agglutinative language, and having Turkic roots, shares similarities with other Turkic languages in terms of syntactic structure. We conducted our research by examining peer-reviewed literature. To include most recent and relevant information, publications written before 2000 were filtered off.

D. Architecture, Model, Diagram description

Bi-LSTM-CRF

The proposed approach is intended to develop NER system for the Azerbaijani language. The algorithms BI-LSTM, and CRF are utilized for classification and automatic labeling to achieve this goal. The system's core logic is written in Python, as well as the user interface, which is delivered via a web application built using the Python Django framework.

A LSTM network and a CRF network are combined to generate the LSTM-CRF model. A LSTM layer and a CRF layer allow this network to efficiently utilize past input features and sentence level tag information, respectively. Lines connect consecutive output layers to represent a CRF layer. A state transition matrix is one of the parameters of a CRF layer. We may utilize past and future tags to forecast the present tag with such a layer, which is similar to using past and future input features via a bidirectional LSTM network.

A bidirectional LSTM network and a CRF network are merged together to generate a BI-LSTM-CRF network, which is quite similar to an LSTM-CRF network. A BILSTM-CRF model is able to utilize future input features alongside the past input features and sentence level tag information used in an LSTM-CRF model. Various papers suggest that the added features can improve tagging accuracy.

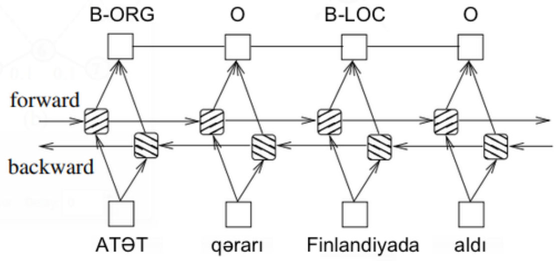


Figure 6. Bi-LSTM-CRF

Word Embedding

fastText

It has been proven (Collobert et al., 2011) [13] that word embedding is critical for improving performance of sequence tagging in NLP related tasks. We downloaded the pre-trained word embeddings, which comprises 50k words and equates to a 300-dimensional embedding vector for each word.

E. Social and environmental impact

Today's rapidly evolving and improving technology has become an inextricable part of people's daily lives in the twenty-first century. It has become an unavoidable truth in society, particularly during the last two years of lockdown and epidemic, that there can be no vision of the future without modern technologies supporting us. Development of technologies that incorporate modern AI and ML applications is anticipated to have a significant positive impact on societal progress, particularly in Azerbaijan, which is a few steps behind western countries in terms of new technology advancement. Language is the foundation of communication, and communication is one of a person's most basic needs. It's unsurprising that technology is being used to enhance human communication through natural language processing. We can certainly say that the social and environmental impact of our project is excessive based on this hypothesis and the gap in named entity recognition in NLP applications for Azerbaijani language. It can also be said that no comprehensive research on building a named entity recognition system for the Azerbaijani language has been undertaken in the worldwide domain, making our work useful for future development and research in this subject by global scientists and enterprises.

IV. Implementation

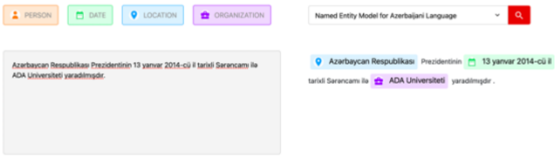
A. Software Design

The software of the project, which is a website written in Python programming language and developed in the Django framework, enables users to input Azerbaijani sentences and see tagged entities. This tagging is done using CRF and Bi-directional LSTM algoritms.

User Interface

The front-end of the web application was built with simplicity in mind to allow users focus on key task of the system – detecting named entities.

A large text box is included to accept input data from the user to be processed by the model. Above that, 4 buttons are available each showing the tags the system can detect. Clicking on each button toggles displaying of the detected entities to corresponding tag.
Upon clicking the search button, the data inside the textbox is sent to the server and result is shown on the right side with tagged entities properly highlighted.



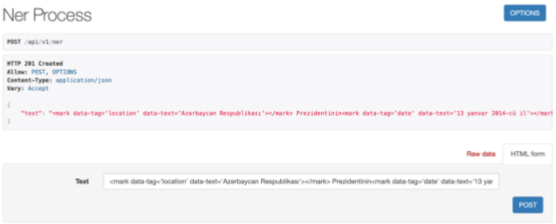
Azərbaycan Respublikası Prezidentinin 13 yanvar 2014-cü il tarixli Sərəncamı ilə ADA Universiteti yaradılmışdır.

Azərbaycan Respublikası Prezidentinin 13 yanvar 2014-cü il tarixli Sərəncamı ilə ADA Universiteti yaradılmışdır .

Figure 7. Software Interface

API

The web app provides a demo API at */api/v1/ner* location to conveniently use the entity recognition system. The data is sent to the API and text with detected and tagged entities are returned in HTML format by the API.



```
POST /api/v1/ner
```

```
HTTP 201 Created
Allow: POST, OPTIONS
Content-Type: application/json
Vary: Accept

{
    "text": "<mark data-tag='location' data-text='Azərbaycan Respublikası'></mark> Prezidentinin<mark data-tag='date' data-text='13 yanvar 2014-cü il'></mar
}
```

Text <mark data-tag='location' data-text='Azərbaycan Respublikası'></mark> Prezidentinin<mark data-tag='date' data-text='13 yar

Figure 8. API Interface

B. Timeline and Gantt chart

The schedule for each stage involved in improving the system was determined from the early stages of the project. Extensive study was conducted over the first two and a half months of the project, from September to December, in order to thoroughly understand the named entity recognition systems and algorithms to implement them, as well as the fundamental grammatical notions in the Azerbaijani language. The design documentation for the project's system architecture was written near the end of the topic research period, when a complete understanding of the functioning principles of similar projects had been achieved. The data collecting and labeling method was used extensively from the completion of the design documents to the end of the project timetable. A week following the start of the data gathering and labeling period, the linguistic study was completed, and the machine learning method was built. Until April, the data labeling and algorithm development processes were carried out in parallel. Testing documentation was written a week before the algorithm development process was completed, and the testing procedure was completed during the final week of the process. To test and refine algorithms as well as their accuracy, and optimize system results, the testing and outcome procedure was run concurrently with the algorithm creation process. During this time, a final report was written and interface changes for the system's web structure were implemented, followed by ultimate deployment to a web server.

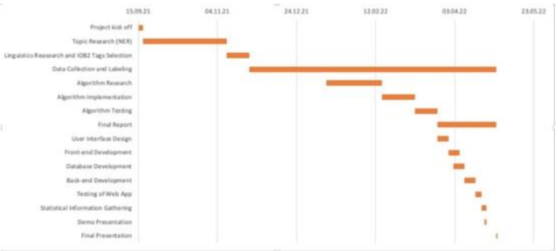


Figure 9. Gantt Chart

C. Testing/Verification/Validation of Results

To achieve and ensure the best outcomes and ensure high accuracy of the model, testing and verification of the algorithms were done periodically. Overall, promising results with up to 98% accuracy for detecting and labeling entities was achieved over 1500 sentence dataset. As for individual tags, Person tags had the highest accuracy, up to 100%, whereas Location had lowest with 94%.

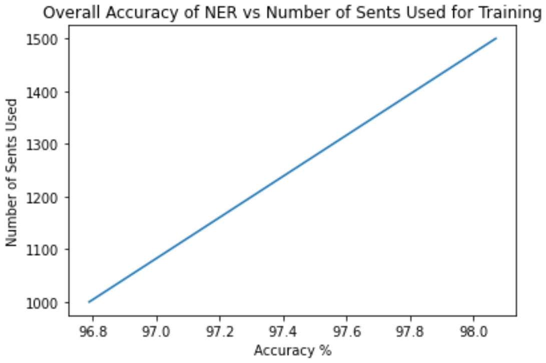


Figure 10. Overall Accuracy

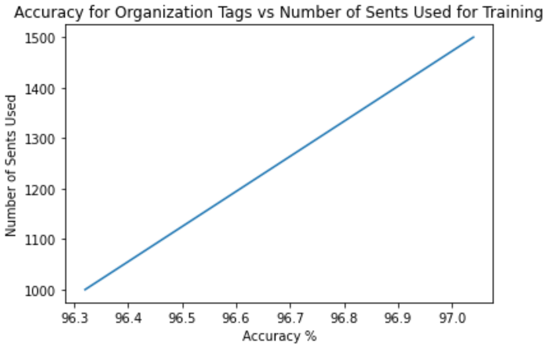


Figure 11. Organization Accuracy

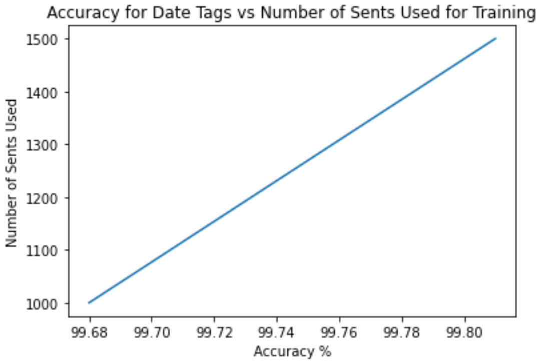


Figure 12. Date Accuracy

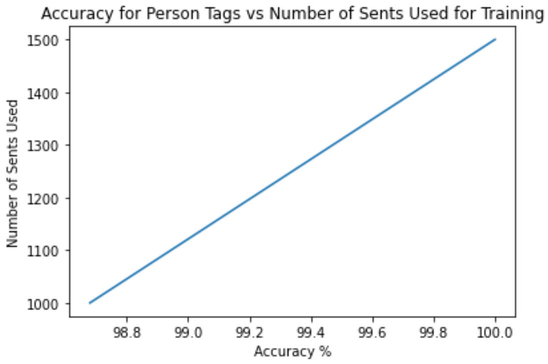


Figure 13. Person Accuracy

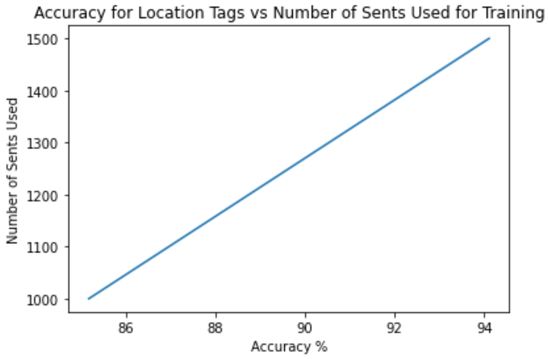


Figure 14. Location Accuracy

Accuracy	98.0724
F1-Score	0.9996
Loss	0.1399

Table 2. Statistics for Bi-LSTM-CRF

V. Conclusion

A. Discussion of Results

In conclusion, it is expected that the developed system would make a significant contribution to the development of NLP systems for the Azerbaijani language. In this project, data in the Azerbaijani language is manually collected and labeled for the corpus building. Following the completion of a tagged corpus with over 1500 sentences, different techniques were used to determine the optimum accuracy for named entity recognition of Azerbaijani language. As a consequence, two of the algorithms, Bidirectional LSTM with CRF, out of all the others examined, performed well and yielded high accuracy scores. On a 1500 sentenced data corpus, the accuracy ratings of the given algorithms are 98 percent – a promising result for a low-resource language such as Azerbaijani.

B. Future Work

A corpus of 1500 sentences was created over the course of the project's development. The accuracy score for the corpus ranges from 95 percent to 98 percent, based on multiple techniques. The project's long-term goal is to expand the corpus set to at least 10k sentences, with an accuracy range of more than 99 percent for all implemented methods. Only web applications are supported by the existing platform and interface for testing and using the established system. It is planned to make the API public, followed by proper documentation of the API and offer the service for outside access for a component of any system in the future development process.

ACKNOWLEDGMENT

This project was carried on at Center for Data Analytics Research (CeDAR) at ADA University, and we would like to thank Dr. Samir Rustamov for supervising the project for the whole duration of SDP I & II courses.

References

[1] E. K. Akkaya, and B. Can, “Transfer Learning for Turkish Named Entity Recognition on Noisy Text,” Natural Language Engineering 27, no. 1(2021): 35–64. doi:10.1017/S1351324919000627

[2] D. Kucuk, and A. Yazıcı, “Named Entity Recognition Experiments on Turkish Texts,” Flexible Query Answering Systems, 2009, 524–35.https://doi.org/10.1007/978-3-642-04957-6 45.

[3] A. Das, D. Ganguly, and U. Garain, “Named Entity Recognition with WordEmbeddings and Wikipedia Categories for allow-Resource Language,” ACM Transactions on Asian and Low-Resource Language Information Processing 16, no. 3 (2017): 1–19.https://doi.org/10.1145/3015467

[4] Akhundova, Natavan. (2021). Named Entity Recognition for the Azerbaijani Language. 1-7. 10.1109/AICT52784.2021.9620336.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research (JMLR).

[6] Brill, E. (1992). A Simple Rule-Based Part-of-Speech Tagger. In Proceedings of the workshop on Speech and Natural Language, pages 112–116. Association for Computational Linguistics.

[7] Lafferty, John & Mccallum, Andrew & Pereira, Fernando. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning.

[8] Morwal, Sudha & Jahan, Nusrat & Chopra, Deepti. (2012). Named Entity Recognition using Hidden Markov Model (HMM). International Journal on Natural Language Computing. 1. 15-23. 10.5121/ijnlc.2012.1402.

[9] A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. Proceedings of ICML.

[10] Mesnil, Grégoire, et al. “Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding.” Interspeech 2013, ISCA, 2013, pp. 3771–75. DOI.org (Crossref), https://doi.org/10.21437/Interspeech.2013-596.

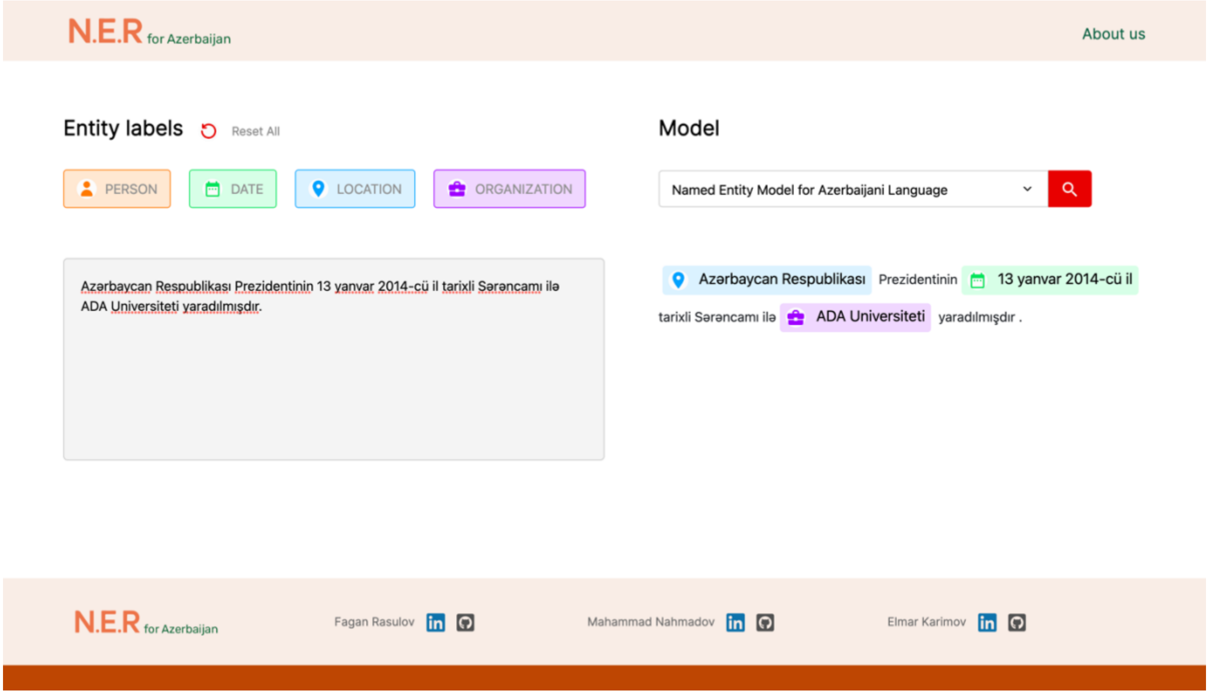
[11] Huang, Zhiheng, et al. “Bidirectional LSTM-CRF Models for Sequence Tagging.” ArXiv:1508.01991 [Cs], Aug. 2015. arXiv.org, http://arxiv.org/abs/1508.01991.

[12] Gudivada, Venkat N., and Kamyar Arbabifard. “Open-Source Libraries, Application Frameworks, and Workflow Systems for NLP.” Handbook of Statistics, vol. 38, Elsevier, 2018, pp. 31–50. DOI.org (Crossref), https://doi.org/10.1016/bs.host.2018.07.007.

[13] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research (JMLR).

Appendices

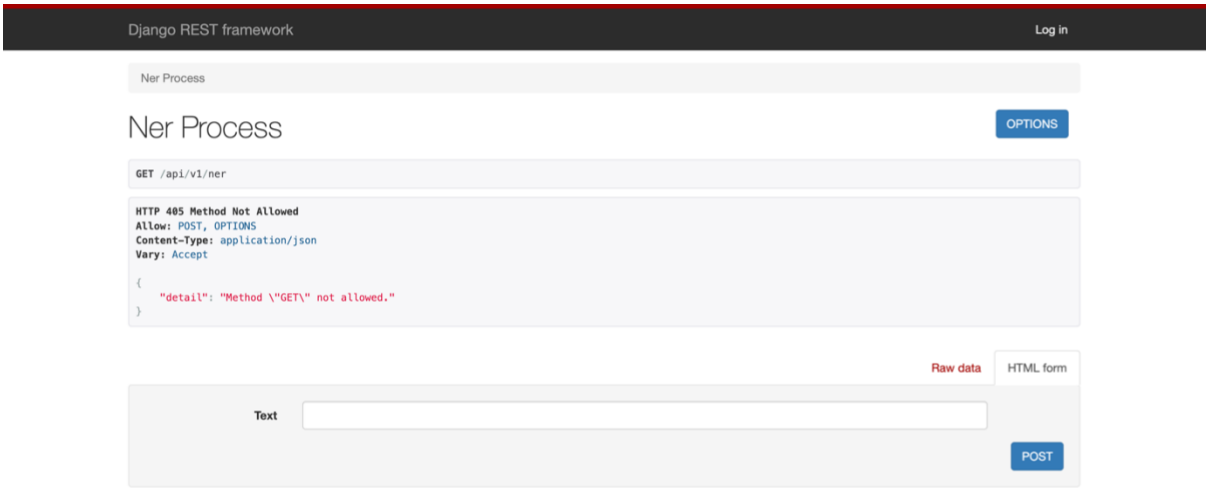
Screenshots of the Software Interface



Azərbaycan Respublikası Prezidentinin 13 yanvar 2014-cü il tarixli Sərəncamı ilə ADA Universiteti yaradılmışdır.

Azərbaycan Respublikası Prezidentinin 13 yanvar 2014-cü il tarixli Sərəncamı ilə ADA Universiteti yaradılmışdır .

API screenshot



```
GET /api/v1/ner
```

```
HTTP 405 Method Not Allowed
Allow: POST, OPTIONS
Content-Type: application/json
Vary: Accept

{
    "detail": "Method \"GET\" not allowed."
}
```

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Explanation
API	Application Programming Interface
CRF	Conditional Random Field
CNN	Convolutional Neural Network
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
MEMM	Maximum Entropy Markov Model
NLP	Natural Language Processing
NER	Named Entity Recognition
PoS	Parts-of-Speech
RNN	Recurrent Neural Network