NER-Based Token Classification for Turkish Sentence Structure Analysis

Abdurrahim Gün
Artifical Intelligence
TOBB University of Economics and Technology
a.gun@etu.edu.tr

Ali Şahin
Artifical Intelligence
TOBB University of Economics and Technology
alisahin@etu.edu.tr

***Abstract*—Understanding sentence structure is crucial for NLP applications. This project aims to identify Turkish sentence elements using Named Entity Recognition (NER) in a token classification framework. The agglutinative and rich morphological structure of Turkish poses challenges for NLP tasks. This research aims to develop a robust model to accurately classify and identify sentence components such as subjects, objects, predicates and determiners.**

A deep learning approach is used to fine-tune a pre-trained BERTurk model on a dataset of labeled Turkish sentences. The model is trained to recognize and label tokens corresponding to various sentence elements, which allows for detailed sentence structure analysis.

This study aims to provide a dataset that can meet the needs of future work in this field and to contribute to deep learning-based shallow parser approaches.

***Index Terms*—Named Entity Recognition, BERT, Token Classification, Natural Language Processing, Turkish Language**

I. Introduction

Analyzing sentence structure in the field of Natural Language Processing (NLP) is crucial for applications. This project focuses on identifying sentence elements in Turkish by leveraging Named Entity Recognition (NER) techniques within a token classification framework. The primary objective of the project is to develop a robust model capable of accurately classifying and identifying sentence components such as subjects, objects, predicates, and modifiers.

Identified Sentence Elements:

- Subjects: The fundamental elements of the sentence, indicating who is performing the action or experiencing the state.
- Objects: The elements affected by or directed towards the verb.
- Predicates: The main verb of the sentence, defining its meaning.
- Adverbial Complements: Elements that modify the verb, indicating aspects such as place, time, amount, and reason.
- Indirect Complements: Elements indicating to whom or what the action of the verb is directed.

Complexity of Analyzed Sentences: The project aims to analyze sentences with varying levels of linguistic complexity. The following linguistic phenomena will be considered:

- Nested Structures: Sentences containing other embedded sentences, creating complex structures.
- Dependencies: Identifying dependency relationships between words in a sentence is crucial, such as the relationship between a verb and its subject or object.
- Morphological Variations: Given that Turkish is an agglutinative language, the affixes attached to words and the changes in meaning these affixes bring must be considered.

Labeled Sentence Components: Accurate labeling of sentence components is critical for the success of the model. This labeling process involves identifying the role of each word or word group in the sentence. For example:

- Subject
- Direct Object
- Indirect Object
- Predicate
- Adverbial Complement
- Indirect Complement
- Punctuation
- Extraneous Elements

This detailed approach will allow for more accurate and comprehensive analyses of Turkish sentences. During the project, deep learning and NER techniques will be utilized to label these sentence components and address linguistic phenomena.

One of the most significant challenges in this field is the lack of a suitable dataset that meets the needs of such analyses. Therefore, one of the fundamental aims of this project is to create a dataset that can also benefit future research. Details on how the dataset will be constructed are thoroughly explained in section 3.

Another critical gap identified in the literature is the lack of sufficient and comprehensive studies on the use of deep learning and NER tools for this specific purpose. NER provides a useful and important infrastructure for the intended task. Breaking down Turkish sentences into their components determines the roles of words in a sentence (such as subject, object, predicate, adverbial complement, and indirect complement). In this process, the type and position of the word in the sentence are determined. NER attempts to identify named entities (such as persons, places, organizations, dates, etc.) in a sentence. These named entities are often used as specific components in a sentence (such as subjects or objects). When the labels for words or word groups are correctly structured, the problem becomes solvable using an NER approach. NER algorithms can utilize information derived from syntactic analyses to better understand the context of words. This allows for a more successful breakdown of the sentence into its components.

In this study, we will fine-tune the pre-trained BERTurk model using the dataset we create. Research has shown that the base model performs well for many NLP tasks. Therefore, the base model will be used initially. If its performance is found to be insufficient, we will transition to the larger model.

The detailed steps of the project are explained in section 3.

II. Related Works

[1] worked about providing an extensive overview of deep learning-based solutions for Named Entity Recognition (NER), aimed at assisting new researchers in gaining a thorough understanding of the field. The comprehensive survey covers the historical background of NER, traditional approaches, current state-of-the-art techniques, and future research directions. It consolidates essential NER resources, including tagged corpora and readily available systems, with a particular emphasis on general domain and English language applications. The authors introduce basic NER concepts, evaluation metrics, and deep learning fundamentals. They review various deep learning models and categorize them using a novel taxonomy, which aids in understanding their applicability to different NER challenges. Additionally, the survey discusses the latest deep learning methods applied to NER, highlighting new problem settings and applications. This detailed exploration helps frame the potential and limitations of NER technologies, guiding our approach to developing a DL-based NER model for Turkish sentence structure analysis.

This work [2] aims to address the nested name-identity recognition (NER) problem using fine-tuned, pre-trained BERT-based language models. By leveraging transfer learning with BERT models, the research provides a simpler and more effective solution compared to existing complex models. The approach transforms the nested NER problem into a flat NER problem by using a common labeling technique to handle nested entities, thus enabling the use of traditional NER models. The proposed models are evaluated using two nested NER datasets (GENIA and GermEval 2014) and one flat NER dataset (JNLPBA). The results reveal that the fine-tuned BERT models significantly outperform traditional models such as CRF and Bi-LSTM-CRF. Based on this study, it is concluded that the elements of Turkish sentences can be successfully classified by fine-tuning the BERTurk model with a well-structured dataset.

In general, this paper [3] describes the shift in NER applications in NLP from LSTM-based models such as ULMFIT and ELMO to transformer-based models such as BERT and GPT. They are supported by extensive text corpora, advanced hardware (e.g. GPUs) and advanced optimization techniques such as Adam and slanted triangle learning rates. The transfer learning paradigm involves fine-tuning pre-trained language models for specific subtasks. By challenging the traditional assumption regarding the pre-training of neural language models, this work demonstrates that domain-specific pre-training from scratch can significantly exceed mixed-domain pre-training. This approach has achieved state-of-the-art results in various biomedical NLP applications. To facilitate this research, BLURB, a comprehensive benchmark for biomedical NLP covering tasks such as Named Entity Recognition (NER), relation extraction, document classification and question answering, was developed. Based on these approaches, it is concluded that pre-trained BERT models will outperform LSTM-based methods on NER tasks. Within the scope of the project, it was decided to use the Bi-LSTM model, one of the LSTM-based models, for performance evaluation.

Shallow discourse parsing is recognized as an important step towards discourse understanding and an important contribution to NLU research in general. However, despite its importance, most existing studies remain limited to English, leaving the field largely understudied in non-English contexts. To address this issue, the researchers performed several subtasks of the shallow discourse parsing pipeline on Turkish [4]. Although their work is not a complete end-to-end parser, it is the most comprehensive work on Turkish to date. All tasks are modeled as multi-class text classification problems using the Turkish BERT model. The results showed satisfactory accuracy comparable to English results despite limited training data. In our project, we aim to identify sentence components in Turkish using shallow parsing. Implementing a Turkish BERT model as shown in the referenced work will improve our accuracy in identifying subjects, objects and predicates. Despite limited training data, the results show that effective shallow parsing is achievable. Furthermore, exploring cross-lingual techniques can help overcome data limitations. In summary, the insights gained from this study will guide us in developing an accurate and efficient shallow parsing system for Turkish sentences.

In another shallow parsing study conducted by Isik University [5], the authors performed an analysis of sentence constituents to identify elements such as subjects, objects and other groups of words with specific grammatical functions. Their research involved applying shallow parsing to 1400 Turkish sentences annotated by seven different annotators. They evaluated the performance of six different models in the context of shallow parsing, including classifiers such as Decision Tree (C45), Naive Bayes, K-Nearest Neighbors (KNN), Linear Perceptron and Multilayer Perceptron. These word embeddings were then used to improve the classifiers, providing continuous features for words without additional feature engineering. Based on this paper, we decided on the state-of-the-art machine learning based models that will be used to compare the performance of the fine-tuned BERTurk model.

III. Project Scope and Methodology

As the first step of the project, due to the lack of an appropriate dataset, we will create our own dataset. The dataset will be constructed using the BOUN Treebank data created by TABILAB. The raw dataset has the following characteristics: The BOUN Treebank contains a total of 9,761 manually annotated sentences from various topics, including biographical texts, national newspapers, educational texts, popular culture articles, and essays. The texts are sourced from the Turkish National Corpus (TNC). The dependency relations in the BOUN Treebank are manually annotated within the Universal Dependencies (UD) framework. Morphological features and UPOS information are initially obtained from the morphological parser by Sak et al. (2011) and automatically converted to UD morphology using our script.

From each category of biographical texts, essays, educational texts, national newspapers, and popular culture articles, the first 250 sentences for training, and the first 30 sentences for testing and validation, were sampled to create the dataset used for labeling. This results in a total of 1,250 sentences for training and 150 sentences each for testing and validation.

The data is in .xlsx (Excel) format. As described above, each word has various attributes, such as root, word type, singularity-plurality, etc. These attributes will be used in state-of-the-art methods for NER and Shallow Parsing concepts. However, for the fine-tuning of the BERTurk model, which is the ultimate goal of the project, these attributes will not be used. Instead, labels created with IOB tagging will be used. IOB tagging (Inside-Outside-Beginning) is a labeling method used for identifying named entities in texts.

- B (Beginning): Indicates the beginning word of the entity. For example, the first word ”John” in the name ”John”.
- I (Inside): Indicates subsequent words within the entity. For example, the second word ”York” in the name ”New York”.
- O (Outside): Indicates words that are not part of any entity.

In the project, elements outside the sentence will be labeled as ”outside.” Punctuation marks will be labeled as ”PUNC.” Other words and phrases forming the core purpose of the project will be labeled as ”B-Subject,” ”I-Subject,” ”B-Predicate,” ”I-Predicate,” ”B-Indirect Object,” ”I-Indirect Object,” ”B-Adverbial Complement,” ”I-Adverbial Complement,” ”B-Definite Object,” ”I-Definite Object,” ”B-Indefinite Object,” ”I-Indefinite Object.” At this stage, the dataset has been created as described, and the words started to be labeled.

Current methods used for NER and Shallow Parsing were researched. The details of the research findings are presented in section 2. Methods which are K-NN, Multi-Layer Perceptron, Naive-Bayes, CRF, and Bi-LSTM-CRF were selected for comparison with the fine-tuned BERTurk model's performance. Since the primary task is classification, commonly used metrics in machine learning which are precision, recall, accuracy, F1-Score, and macro average F1-Score, will be used for performance comparison, and confusion matrices will be obtained for classifiers. These metrics will be calculated with the help of Scikit-Learn, a Python machine learning library. Additionally, Cohen's Kappa metric will be examined to measure the overlap between two classifiers.

For the implementation of state-of-the-art models, Python libraries Scikit-Learn and PyTorch will be used, while the Transformer library and the pre-trained model from the Hugging Face platform will be utilized for the fine-tuning of the BERTurk model.

References

[1] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70, 2020.

[2] Ankit Agrawal, Sarsij Tripathi, Manu Vardhan, Vikas Sihag, Gaurav Choudhary, and Nicola Dragoni. Bert-based transfer-learning approach for nested named-entity recognition using joint labeling. *Applied Sciences*, 12(3):976, 2022.

[3] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[4] Ferhat Kutlu, Deniz Zeyrek, and Murathan Kurfalı. Toward a shallow discourse parser for turkish. *Natural Language Engineering*, pages 1–26, 2023.

[5] Ozan Topsakal, Onur Açıkgöz, Ali Tunca Gürkan, Ali Buğra Kanburoğlu, Burak Ertopçu, Berke Özenç, İlker Çam, Begüm Avar, Gökhan Ercan, and Olcay Taner Yıldız. Shallow parsing in turkish. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 480–485. IEEE, 2017.