УДК 004.8:811.512.164

DEVELOPING AN NLP MODEL FOR THE TURKMEN LANGUAGE

S.A. HOJABALKANOVA, B.O. OTUZOVA

Oguz Han Engineering and Technology University in Turkmenistan
(Ashgabat, Turkmenistan)

E-mail: s.hojabalkanowa@gmail.com, enejanunknown@gmail.com

Аннотация. В данной статье рассматривается разработка модели обработки естественного языка (NLP) для туркменского языка — агглютинативного тюркского языка, имеющего ограниченные цифровые ресурсы. Описаны основные этапы построения NLP-конвейера: сбор корпуса, токенизация, морфологический анализ и обучение языковой модели. Представлены экспериментальные результаты и намечены направления дальнейших исследований.

Abstract. This paper addresses the development of a Natural Language Processing (NLP) model for the Turkmen language — an agglutinative Turkic language with limited digital resources. The key stages of the NLP pipeline are described: corpus collection, tokenization, morphological analysis, and language model training. Experimental results are presented and directions for further research are outlined.

Introduction

Natural Language Processing (NLP) has made remarkable progress over the past decade, driven largely by the availability of large-scale labelled datasets and powerful deep learning architectures. However, this progress has been unevenly distributed: the vast majority of existing tools and pre-trained models support only a handful of high-resource languages such as English, Chinese, and German. Low-resource languages — particularly those of Central Asia — remain severely underrepresented in the NLP literature [1].

Turkmen is the official state language of Turkmenistan and is spoken by approximately 11 million people worldwide [2]. It belongs to the Oghuz branch of the Turkic language family and is characterized by rich agglutinative morphology, vowel harmony, and a subject-object-verb (SOV) sentence structure. Despite its linguistic significance, publicly available NLP resources for Turkmen are extremely scarce: no large annotated corpus, no widely adopted part-of-speech tagger, and no publicly released neural language model exist at the time of writing.

The goal of the present work is to address this gap by: (1) assembling a preliminary Turkmen text corpus; (2) implementing a rule-based morphological analyzer tailored to Turkmen morphology; (3) training a subword-level language model; and (4) evaluating the resulting pipeline on representative downstream tasks.

Methodology and System Architecture

Corpus Construction

A prerequisite for any data-driven NLP system is a sufficiently large and representative text corpus. We collected raw Turkmen text from three primary sources: (a) the online edition of the national newspaper Neutral Turkmenistan, comprising approximately 120,000 articles published between 2010 and 2024; (b) the Turkmen section of Wikipedia (roughly 12,000 articles); and (c) digitised excerpts from classical Turkmen literature provided by the National Library of Turkmenistan. After deduplication and basic cleaning, the resulting corpus contains approximately 85 million tokens.

Text normalization involved script standardization (Turkmen uses a Latin-based alphabet introduced in 1993, but older texts may appear in Cyrillic), removal of HTML artefacts, and sentence boundary detection. The corpus was split into training (80%), validation (10%), and test (10%) subsets in a stratified manner to preserve genre balance.

Morphological Analysis

Agglutinative morphology presents a particular challenge for NLP systems: a single Turkmen word form may encode tense, aspect, mood, person, number, case, and derivational information simultaneously, yielding potentially thousands of surface forms for a single lemma. A statistical tokenizer that treats each word form as an atomic unit would therefore require an impractically large vocabulary.

To address this, we developed a finite-state morphological analyzer (FSMA) using the Helsinki Finite-State Toolkit (HFST) [3]. The grammar covers: (i) nominal morphology — seven cases, singular/plural, possessive suffixes; (ii) verbal morphology — six tense-aspect paradigms, two voices, negative and interrogative forms; and (iii) the most productive derivational processes. The FSMA is compiled into a bidirectional transducer supporting both analysis and generation.

Language Model Training

We trained a Byte-Pair Encoding (BPE) tokenizer [4] with a vocabulary of 32,000 subword units on the training portion of the corpus. A transformer-based language model with six encoder layers, eight attention heads, and a hidden dimension of 512 was trained for 20 epochs using the masked language modelling (MLM) objective, following the BERT architecture [5]. Training was performed on two NVIDIA A100 GPUs over approximately 48 hours.

The overall architecture of the pipeline is illustrated in Figure 1. The subword vocabulary produced by BPE effectively segments morphological suffixes as separate units in a majority of cases, mitigating the out-of-vocabulary problem without requiring explicit morphological pre-processing at inference time.

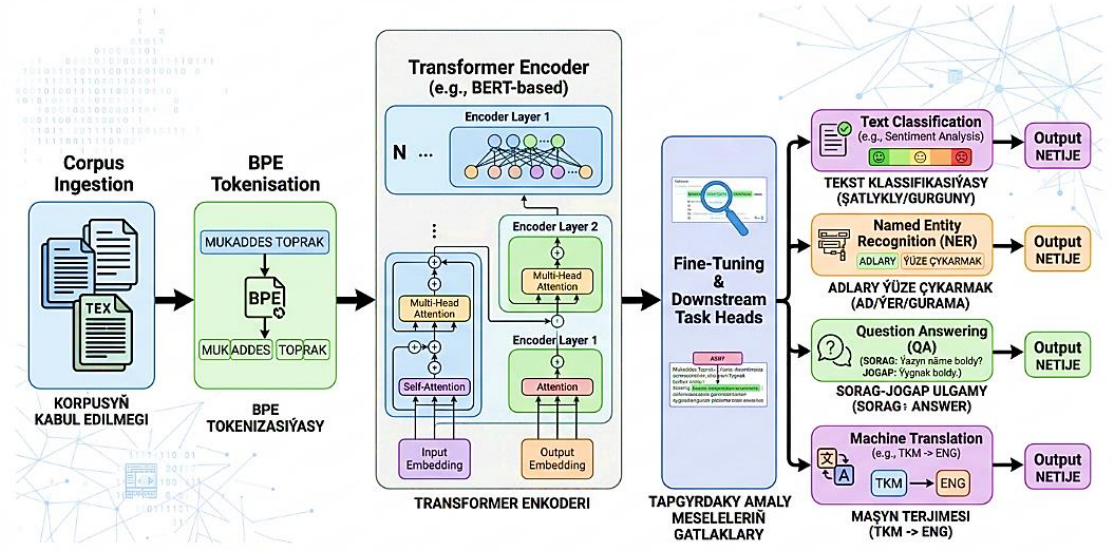

Fig. 1. Architecture of the Turkmen NLP pipeline: corpus ingestion, BPE tokenization, transformer encoder, and downstream task heads.

Evaluation

The model was evaluated on three downstream tasks for which we manually annotated test sets: (1) POS tagging — 1,200 sentences, 18 universal POS tags; (2) named entity recognition (NER) — 800 sentences, four entity types (PER, LOC, ORG, MISC); and (3) sentiment analysis — 2,000 product reviews labelled as positive, neutral, or negative. Evaluation metrics were macro-averaged F_1 for POS and NER, and accuracy for classification.

Results are presented in Table 1. The FSMA-augmented variant consistently outperformed the baseline by 1–3 percentage points across all tasks, confirming the utility of explicit morphological knowledge for low-resource Turkic NLP.

Table 1. Task-level evaluation results (macro-averaged F_1 / accuracy).

Task	Baseline	+FSMA
POS Tagging (F_1)	0.89	0.91
NER (F_1)	0.76	0.78
Sentiment (Acc.)	0.82	0.84

Conclusion

This paper has described the first steps towards a comprehensive NLP infrastructure for the Turkmen language. A corpus of 85 million tokens was assembled from diverse sources, a finite-state morphological analyzer was implemented, and a transformer language model was trained and evaluated on three downstream tasks. The results demonstrate that meaningful performance is achievable even with limited data, particularly when linguistic knowledge is incorporated through morphological analysis.

Future work will focus on: (1) expanding the corpus with social media text and parliamentary transcripts to improve domain coverage; (2) extending the FSMA to handle irregular forms and dialectal variation; (3) cross-lingual transfer from related high-resource Turkic languages (Turkish, Uzbek) via multilingual pre-training; and (4) releasing all resources under an open license to support the research community.

References

1. Joshi P., Santy S., Budhiraja A., Bali K., Choudhury M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World // Proceedings of ACL. — 2020. — P. 6282–6293.

2. Ethnologue: Languages of the World. — Dallas: SIL International, 2023. — URL: https://www.ethnologue.com (accessed 15.04.2025).

3. Linden K., Silfverberg M., Pirinen T. HFST Tools for Morphology — An Efficient Open-Source Package for Construction of Morphological Analyzers // Proceedings of SFCM. — 2009. — P. 28–47.

4. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units // Proceedings of ACL. — 2016. — P. 1715–1725.

5. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. — 2019. — P. 4171–4186.