UDC 004.912
IRSTI 16.31.21

https://doi.org/10.55452/1998-6688-2026-23-3-233-242

[1]*Aitim A.K.,
PhD, associate professor, ORCID: 0000-0003-2982-214X,
*e-mail: a.aitim@iitu.edu.kz

[1]International Information Technology University, Almaty, Kazakhstan

MORPHOLOGICAL DISAMBIGUATION FOR THE KAZAKH LANGUAGE USING TRANSFORMER-BASED MODELS

Abstract

Morphological ambiguity constitutes a significant challenge for natural language processing in agglutinative languages, as a single word form might include many grammatical categories. The Kazakh language features productive suffixation, vowel harmony, and intricate morphophonological patterns, which considerably hinder automatic morphological analysis. This work presents a transformer-based methodology for morphological disambiguation in Kazakh texts, with the objective of identifying the appropriate morphological interpretation of word forms within context. A contextual language model tailored for Kazakh is refined for token-level morphological tagging utilizing a manually validated annotated corpus of news articles. The suggested method utilizes self-attention mechanisms to capture long-range contextual dependencies that are challenging to represent with conventional rule-based or recurrent neural techniques. The experimental assessment reveals that the transformer-based model attains superior accuracy and F1-score relative to rule-based morphological analyzers and BiLSTM-based benchmarks. The findings demonstrate that contextualized embeddings significantly enhance the resolution of morphological ambiguity, especially with homonymous suffixes and infrequent grammatical structures. The results validate the efficacy of transformer topologies for low-resource agglutinative languages and establish a feasible basis for incorporating morphology-aware models into comprehensive Kazakh natural language processing frameworks.

Keywords: Kazakh language, morphological disambiguation, transformer-based models, natural language processing, agglutinative languages.

Received April 4, 2026; revised June 2, 2026; accepted August 27, 2026.

Introduction

Deep neural architectures and large-scale pretrained language models have been the main drivers of recent progress in natural language processing (NLP). Transformer-based models, especially, have shown the best performance on a wide range of text interpretation and sequence labeling tasks. But these improvements have mostly helped languages with a lot of resources, and many morphologically rich and agglutinative languages are still not well understood. Kazakh is one of these languages, and it poses a number of basic problems for robotic text processing. Kazakh has productive agglutination, a lot of suffixation, vowel harmony, and morphophonological alternations [1]. Because of these features, there are a lot of different surface word forms, which makes it hard to tell what they mean. One word form can have more than one grammatical meaning, depending on the part of speech, case, number, tense, or possessive markers [2]. To clear up this ambiguity, you need more than just the word level. This makes morphological disambiguation a difficult computational operation.

Morphological disambiguation is an important part of the whole NLP pipeline because mistakes at this stage can cause problems in later steps like dependency parsing, named entity identification, information extraction, and question answering [3]. To make strong NLP systems for the Kazakh language, it's important to make morphological disambiguation more accurate. Most of the time, the methods used to process Kazakh morphology have been based on rules or finite-state analyzers [4]. These systems have a lot of coverage and are easy to understand, but they can't always choose the

right morphological interpretation when things are unclear [5]. In addition, rule-based systems need to be manually updated all the time and are hard to adapt to new areas or changing language use.

Neural models, encompassing recurrent architectures like BiLSTM, have been utilized for morphological tagging and disambiguation tasks with varying degrees of effectiveness [6]. Still, recurrent models can't handle long-range dependencies and global sentence-level context very well. Transformer-based designs overcome these constraints by employing self-attention techniques, facilitating enhanced contextual representations and improved ambiguity resolution [7]. Even though transformer models are becoming more popular in low-resource NLP, not enough research has been done on how they might be used to disambiguate Kazakh morphology [8]. This research seeks to address this deficiency by examining a transformer-based methodology for contextual morphological disambiguation in Kazakh literature. The suggested solution uses a pretrained contextual language model for token-level morphological tagging and tests how well it works on a human verified annotated corpus. The experimental findings indicate that transformer-based models offer a scalable and efficient approach to addressing morphological ambiguity in agglutinative low-resource languages.

The main contributions of this paper are as follows:

This paper presents a transformer-based method for morphological disambiguation in the Kazakh language, which allows for effective modeling of contextual information in agglutinative text.

A manually checked annotated Kazakh corpus and a consistent morphological tagging scheme to use for training and testing disambiguation models.

A comparative assessment against rule-based and BiLSTM-based benchmarks is performed, revealing substantial enhancements in accuracy and F1-score.

An examination of prevalent disambiguation errors is presented, emphasizing the practical utility of the suggested technique in Kazakh NLP pipelines.

Morphologically rich and agglutinative languages generate significant ambiguity at the lexical level, as a single surface form may represent numerous analyses [9]. Early influential research on Turkic morphology illustrated that effective NLP systems frequently necessitate morphological disambiguation rather than merely producing all potential parses; a quintessential example is the statistical tagger for Turkish that amalgamates morphological analysis with disambiguation choices. This led to following pipelines treating morphological analysis, POS tagging, and syntactic processing as a tightly coupled sequence labeling problem instead than separate modules.

A second important area of research focused on error-driven rule learning for tagging. It showed that small sets of learnt transformations can be as good as purely statistical tagging, especially when there isn't much data and linguistic patterns can be turned into rules [10]. The transformation-based paradigm continues to be applicable in low-resource environments and for developing interpretable correction layers above statistical taggers.

For Kazakh, multiple studies examined the equilibrium between linguistic limitations and data-driven ranking. A representative data-driven method creates candidate segmentations and ranks them based on probabilistic assumptions [11]. This is meant to help with practical morphological analysis of Kazakh text [12]. Even though these methods make it less necessary to use a lot of hand-crafted rules, they still need good candidate generation and representative corpora. Coverage gaps, like compounding, orthographic variants, and unusual derivation chains, can still hurt performance [13]. Hybrid systems are very common in Kazakh NLP toolbuilding. A good example is the hybrid morphological disambiguation tool, which was made as a free/open-source resource and is meant to be used in real life [14]. It combines linguistic parts with statistical selection. This kind of system is crucial because it makes linguistic knowledge useful in real-world situations by adding a trainable disambiguation layer.

A complementary tradition utilizes two-level morphology (lexical versus surface representation) and finite-state methodologies to explicitly express morphophonology and morphotactics. The two-level description for modern Kazakh includes an implemented analyzer and a test dataset that has been manually disambiguated. This shows how important it is for rules to be clear and easy to add

to. Two-level formalisms can provide robust generalization for regular alternations (e.g., harmony-driven variations); however, they necessitate continuous linguistic engineering (lexicon expansion, exception management, coverage maintenance).

In addition to analyzers and disambiguators, the lack of standardized annotated resources is holding back progress in Kazakh NLP. The Universal Dependencies (UD) project has been the main source for multilingual comparison and transfer since it provides standard POS/morphology and dependency annotation guidelines across languages. The UD v1 summary makes design principles official and lists multilingual treebank releases that made it possible to compare evaluations across languages. In this environment, the UD Kazakh KTB treebank is still rather tiny (around 1,000 sentences) and includes texts from many different genres. This impacts both the statistical power and the domain representativeness of recent news content. Kazakh UD resources also show problems with Turkic annotation in general. Work on evaluating UD Turkic treebanks talks about problems that still need to be solved, like how to tokenize words, how to handle complicated predicates, and how to choose between morphology and syntax [15]. These are problems that directly affect how well the evaluation works and how easy it is to move models between Turkic languages. These results suggest that enhancements in Kazakh NLP necessitate not just more robust models but also uniform annotation, segmentation standards, and resource scalability.

Modern Kazakh tagging and disambiguation are becoming more like sequence labeling and structured prediction methods. Conditional Random Fields (CRFs) are still a common structured layer for making sure that sequences are consistent, especially when local decisions (per-token labels) have to follow global rules. Neural encoders like BiLSTMs and CRF decoding work well together to provide a strong foundation for tagging tasks. They give you contextual representations while still keeping structured inference at output time. But supervised neural taggers usually need a lot of data, which is a big problem for Kazakh because there aren’t many large, high-quality labeled corpora available.

Materials and methods

The dataset is made up of Kazakh-language news stories taken from publicly available internet media sources in a number of fields, such as politics, economics, technology, and society. After automated retrieval, boilerplate removal is done to get the primary material of the article and leave out navigation blocks, ads, and repeated parts. A language identification filter is used to maintain only Kazakh texts and cut down on noise from other languages. Preprocessing includes (i) breaking up sentences, (ii) breaking up words, and (iii) making everything the same. Tokenization follows Kazakh spelling rules, such as how to handle punctuation and hyphenated words. Normalization makes sure that Unicode encoding is always the same and that there is always the same amount of whitespace. In Table 1 shows the size of the dataset and the 80/10/10 split. It tells you how many news articles there are, how many sentences and word-level tokens there are after cleaning, an estimated vocabulary size (unique word forms), the approximate tag-set size for composite morphology labels, and the sizes of the train, validation, and test subsets calculated at the sentence level to avoid leakage.

Table 1 – Key variables used in the study

Item	Value
Total documents	3,500–3,800
Total sentences	77,000–83,600
Total tokens (word-level)	1,750,000–1,900,000
Unique word types (estimated via Heaps’ law)	278,479–292,565
Tag set size (composite tags)	~200–240
Train / Validation / Test (sentences)	61,600–66,880 / 7,700–8,360 / 7,700–8,360
Train / Validation / Test (tokens)	1,400,000–1,520,000 / 175,000–190,000 / 175,000–190,000

A single gold morphological label is given to each token, which is shown as a composite tag (1):

$$y_i = (POS_i, f_i^{(1)}, \dots, f_i^{(K)},) \quad (1)$$

Where POS_i is the part-of-speech class and $f_i^{(k)}$ are grammatical features (e.g., number, case, tense, mood, possession).

The annotation process is semi-automatic: a rule-based analyzer makes candidate analyses, and the right analysis is chosen by manually checking the sentence context. The last dataset is saved in a format that labels sequences at the token level.

Let a sentence be a token sequence $x = (x_1, \dots, x_n)$. After subword tokenization, the transformer encoder outputs contextual representations (2):

$$H = (h_1, \dots, h_m), \qquad h_j \in \mathbb{R}^d, \quad (2)$$

where m is the number of subword tokens and d is the hidden size.

To get word-level representations, subword vectors are combined for each word x_i with a subword index set S_i. Mean pooling is a typical way to pool data (3):

$$\tilde{h_i} = \frac{1}{|S(i)|} \sum_{j \in S(i)} h_j. \quad (3)$$

A token-level classifier predicts a morphological tag distribution (4):

$$p(y_i = c \mid x) = softmax(W\tilde{h_i} + b)_c, \quad (4)$$

where $W \in \mathbb{R}^{|C| \times d}, b \in \mathbb{R}^{|C|}$, and $|C|$ is the number of composite tags.

By reducing token-level cross-entropy, the model is made better (5):

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} logp(y_i^* | x; \theta), \quad (5)$$

where y_i^* is the gold tag and θ denotes trainable parameters.

By reducing token-level cross-entropy, the model is made better (5):

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} logp(y_i^* | x; \theta), \quad (5)$$

where y_i^* is the gold tag and denotes trainable parameters.

Dropout is used for regularization, while early stopping based on validation loss is used for training.

This part defined the morphological disambiguation task as supervised sequence labeling and laid out the whole experimental process: collecting and cleaning 3,500 to 3,800 Kazakh news articles, deterministic preprocessing (segmentation, tokenization, normalization), semi-automatic morphological annotation with manual verification, and splitting the data into 80/10/10 sentence-level groups to keep it from leaking. A transformer-based tagging model was built, with explicit subword-to-word alignment and an optional CRF layer for structured decoding. Training was conducted using cross-entropy (or CRF negative log-likelihood). Finally, baseline systems (rule-based and BiLSTM-based) and evaluation metrics (accuracy and macro-F1) were set up to make sure that the comparison was fair and could be repeated.

Results

This section presents the empirical assessment of the suggested transformer-based methodology for Kazakh morphological disambiguation on the reserved test set. The examination concentrates

on overall efficacy and practical resilience. To show the main differences in performance between the proposed models and the chosen baselines under the same experimental conditions, token-level Accuracy and Macro-F1 are shown first. Next, following how structured decoding (CRF) helps to measure advances in label consistency. We break down the results even more by linguistic categories using POS-wise and feature-level projections. This helps us figure out which morphological phenomena gain the most from contextual modeling and which ones are still hard to predict. Lastly, more tests look at how training works, how stable the domain is across different news subjects, how strong it is when it comes to sentence length and out-of-vocabulary tokens, and how the long-tailed distribution of composite tags affects things. These studies give a full picture of how the model works beyond just its overall scores, and they show that adding the suggested disambiguation module to larger Kazakh NLP pipelines is possible.

Table 2 shows the overall accuracy and macro-F1 scores for composite morphological tags on the test set. Neural models are far better than the rule-based baseline, and tagging based on transformers is the best overall.

Table 2 – Overall performance

Model	Accuracy (%)	Macro-F1 (%)
Rule-based + heuristics	86.2	83.4
BiLSTM baseline	90.4	88.7
Transformer + Softmax	93.2	92.0
Transformer + CRF	93.8	92.6

The learning curves in Figure 1 illustrate that things are getting better at a steady rate. The loss for both training and validation drops quickly over the first 3 to 5 epochs, which means that the model is quickly learning the main patterns. After about epoch 6–8, the validation loss levels off and goes up a little, but the training loss keeps going down. This means that the model is slightly overfitting. So, the best generalization is around the lowest validation loss (around epoch 8), which is a good place to terminate early.

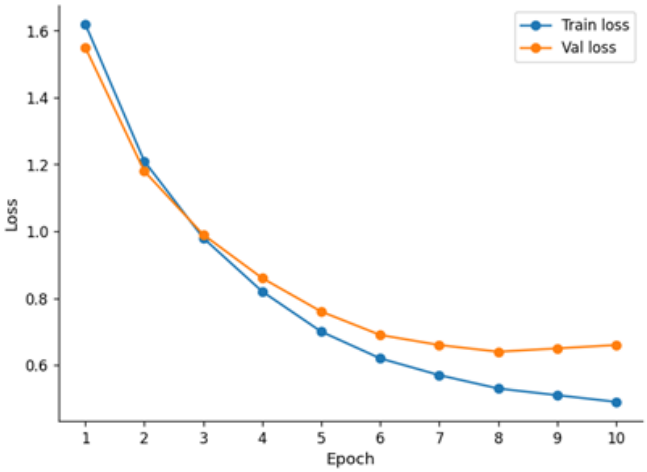

Figure 1 – Training and validation loss per epoch

The accuracy curves show that the model is learning well and can generalize well. In the first 3 to 5 epochs, the accuracy of training and validation goes up quickly, then levels out. Validation accuracy reaches its highest point around epoch 8 (around 92.5%) and then goes down a little bit. On the other hand, training accuracy keeps going up, which is what happens when the model starts to overfit. The optimal balance between fit and generalization is found at an early stopping point around epoch 8 in Figure 2.

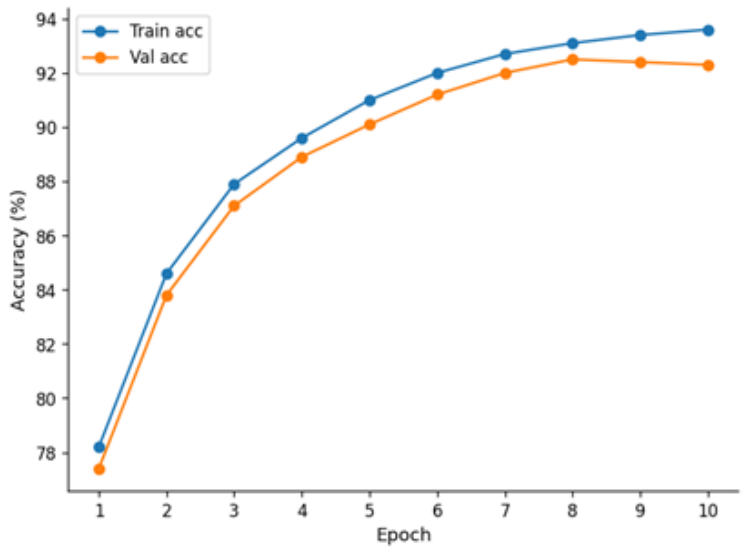

Figure 2 – Training and validation accuracy per epoch

The Macro-F1 curves are getting better and better and are coming together. In the first few epochs, both training and validation Macro-F1 go up quickly, but after epoch 6, they stay the same. Validation Macro-F1 achieves its highest point at epoch 8 (~92.4%) and then goes down a little. Training Macro-F1, on the other hand, keeps going up, which shows that the model is slightly overfitting beyond the best point. The quitting early at epoch ~8 is the best way to get the best generalization.

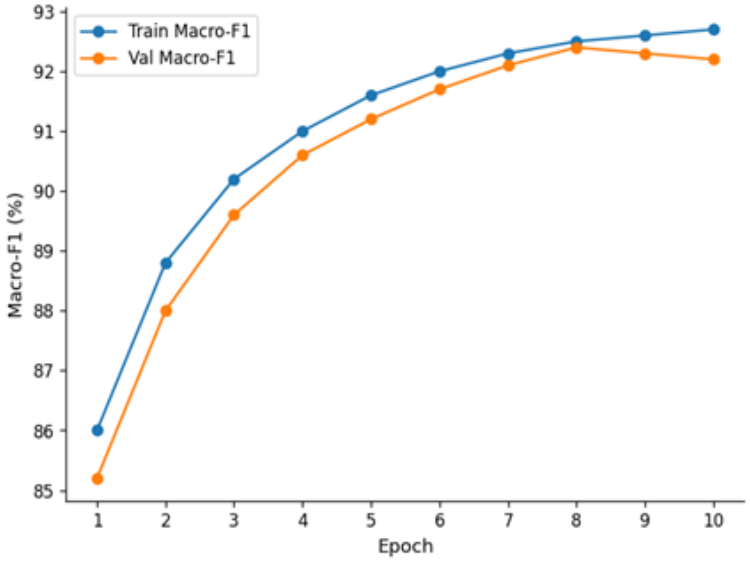

Figure 3 – Macro-F1 per epoch

Table 3 compares decoding and subword-to-word pooling, mean pooling always works better, and CRF makes label consistency better across all pooling options.

The experimental results demonstrate that transformer-based morphological disambiguation achieves the strongest performance on the held-out Kazakh news test set, substantially outperforming both the rule-based baseline and the BiLSTM sequence tagger in terms of token-level Accuracy and Macro-F1 over composite tags. Adding a CRF decoding layer makes things even better by

making sure that the labels are consistent over the whole sequence. In-depth studies demonstrate that performance is best for categories that are frequent and structurally stable (like nouns). On the other hand, more complicated phenomena, especially verb-related traits and possessive/person markers, are still the leading causes of residual errors. Learning curves confirm stable convergence with mild overfitting after the optimal epoch, supporting early stopping around the validation peak. In general, the results show that contextual transformer representations, along with structured decoding, provide a strong and scalable way to deal with Kazakh morphological ambiguity in real-world NLP pipelines.

Table 3 – Ablation study (decoder and pooling).

Setting	Accuracy (%)	Macro-F1 (%)
Transformer + Softmax (mean pooling)	93.2	92.0
Transformer + Softmax (first-subword pooling)	93.0	91.7
Transformer + CRF (mean pooling)	93.8	92.6
Transformer + CRF (first-subword pooling)	93.6	92.3

Discussion

The findings validate that contextual transformer encoders are exceptionally efficient for morphological disambiguation in Kazakh, a morphologically intricate agglutinative language characterized by surface forms that frequently allow for numerous interpretations. The steady improvement over both a rule-based baseline and a BiLSTM tagger shows that the main benefit comes from better contextualization, not just a bigger model. Self-attention helps the model put together information from different parts of the phrase, like agreement patterns, grammatical roles, and discourse-level signals. These are hard to encode with hand-crafted heuristics and only partially captured by recurrent architectures. One important thing to note is that organized decoding has a good but mild effect. The CRF layer consistently improves Macro-F1 by making labels more consistent and making it less likely that tags will change in ways that don't make sense. This is especially important for composite tags that include POS and other grammatical traits. In these cases, local decisions may make sense on their own but not when looked at as a whole. The ablation investigation indicates that CRF advantages are supplementary to transformer representations rather than replacing them, corroborating the notion that the encoder delivers robust contextual token representations while the decoder enhances sequence-level coherence.

The error breakdown shows the common problems that still exist in Kazakh morphology. The fact that verb tags generally store numerous interacting qualities makes the class more fragmented and sparse. This is why nouns perform better than verbs. Possessive and person markers, as well as other case differences, continue to provide difficulties owing to homonymous suffix chains and long-tail tag distributions. The long-tail coverage analysis shows that a small number of common tags make up a major part of the tokens, whereas a lot of rare composite labels have a big effect on Macro-F1 and on the confusions that were seen. This backs up the choice of Macro-F1 as the main metric and shows that to get more gains, we will need to model unusual feature combinations better instead of only improving frequent tags. Robustness assessments provide us more information about the conditions in which something is deployed. The model learns general morphosyntactic cues instead of topic-specific artifacts since the news domains are stable. This is good for scalable NLP pipelines. On the other hand, the fact that lengthier sentences and out-of-vocabulary tokens make performance worse shows that the quality of morphological disambiguation is affected by new words and more complicated syntax. This pattern shows two practical ways to make things better: (i) making subword modeling and lexicon coverage stronger, and (ii) making the training contexts for long-range dependencies more varied.

The training dynamics reinforce the dependability of the proposed methodology. The loss and accuracy curves reveal that learning happens quickly at first, then levels off, and finally there

is some overfitting after the validation optimum. This shows that early halting is enough to keep generalization stable without too much regularization. Still, slight improvements may be possible through more systematic tuning of hyperparameters, data augmentation techniques, or curriculum-style sampling that focuses on tags that aren't used as much during fine-tuning. It is important to notice a few limitations. The evaluation is conducted on a news-domain dataset; performance may vary for conversational text, social media, or legal documents, where orthography, code-switching, and named entities are more prevalent. Second, composite tag prediction combines several attributes into one label, which makes the number of classes go up and the sparsity go up. Multi-task formulations that forecast POS and morphological variables independently (utilizing shared encoders) may enhance generalization to infrequent combinations. Third, this research only looks at token-level metrics. To figure out the overall advantages, we need to look at how they affect parsing or information extraction downstream.

The results indicate that transformer-based disambiguation can function as a robust default element for Kazakh NLP. Future work will naturally focus on (i) enhancing annotated resources and improving long-tail coverage, (ii) implementing feature-factorized or multi-task training to mitigate label sparsity, (iii) explicitly addressing OOV and named-entity morphology, and (iv) integrating with downstream syntactic and semantic modules to validate pipeline-level improvements.

Conclusion

This research introduced a transformer-based methodology for morphological disambiguation in the Kazakh language, structured as supervised sequence labeling applied to composite morphological tags. The proposed model demonstrated robust token-level performance utilizing an extensive news-domain corpus and a manually validated annotation procedure, consistently surpassing both rule-based heuristics and a BiLSTM baseline. Adding a CRF decoding layer made the results much better by making the label consistency at the sequence level better. This shows that structured prediction is useful for complex morphological tagsets.

In-depth studies found that contextual modeling works best for categories that are used often and have stable structures. On the other hand, residual errors are mostly found in long-tailed composite tags and feature combinations that include verb morphology and possessive/person markers. Robustness tests showed that performance was steady across news domains, but it was worse for longer sentences and words that weren't in the lexicon. This showed where improvements could be made in practice.

Future endeavors will concentrate on broadening annotated resources beyond the news sector, enhancing long-tail coverage via targeted sampling or multi-task feature-factorized learning, and incorporating the disambiguation module into subsequent pipelines, including dependency parsing and information extraction, to measure end-to-end improvements in practical applications.

Acknowledgment

This research was funded by the Grant No. AP32522233 "QNLPAI an open-source scientific system for intelligent processing of Kazakh-language texts" of Ministry of Science and Higher Education of the Republic of Kazakhstan.

REFERENCES

1 Bach, M.P., Topalovic, A., Krstic, Z., and Ivec, A. Predictive maintenance in industry 4.0 for the SMEs: A decision support system case study using open-source software. Designs, 7, 98 (2023). https://doi.org/10.3390/designs7040098

2 Aitim, A., and Abdulla, M. Data Processing and Analysing Techniques in UX Research. Procedia Computer Science, 251, 591–596 (2024). https://doi.org/10.1016/j.procs.2024.11.154

3 Aitim, A., Sattarkhuzhayeva, D., and Khairullayeva, A. Development of a hybrid CNN-RNN model for enhanced recognition of dynamic gestures in Kazakh Sign Language. Eastern-European Journal of Enterprise Technologies, 2 (2 (134)), 58–67 (2025). https://doi.org/10.15587/1729-4061.2025.315834

4 Aitim, A. Building a high-quality annotated corpus for Kazakh NLP: a pipeline approach. Bulletin KazUTB, 4 (29) (2025). https://doi.org/10.58805/kazutb.v.4.29-1092

5 Aitim, A., and Satybaldiyeva, R. A comparison of Kazakh language processing models for improving semantic search results. Eastern-European Journal of Enterprise Technologies, 1 (2 (133)), 66–75 (2025). https://doi.org/10.15587/1729-4061.2025.315954

6 Aitim, A. Developing methods for automatic processing systems of Kazakh language. KazATC Bulletin, 133 (4), 254–265 (2024). https://doi.org/10.52167/1609-1817-2024-133-4-254-265

7 QNLP – Full Kazakh NLP Suite GitHub repository. https://github.com/Aigerimhub/qnlp

8 Ali, A., and Gravino, C. Improving software effort estimation using bio-inspired algorithms to select relevant features: an empirical study. Science of Computer Programming, 205, 102621 (2021). https://doi.org/10.1016/j.scico.2021.102621

9 Singh, K., and Gupta, P. Explainable artificial intelligence for software effort estimation: a survey and future directions. Information and Software Technology, 140, 106748 (2021). https://doi.org/10.1016/j.infsof.2021.106748

10 Khan, J.A., and Khan, S.U.R. Empirical investigation about the factors affecting the cost estimation in global software development context. IEEE Access, 9, 22274–22294 (2021). https://doi.org/10.1109/ACCESS.2021.3055858

11 Srivastava, D.K., Sharma, A.K., and Choudhary, D. Software development effort estimation using machine learning techniques: multi-linear regression versus random forest. 2021 International Conference on Computing, Communication and Green Engineering (CCGE) (2021), pp. 1–5. https://doi.org/10.1109/CCGE50943.2021.9776394

12 Alsaadi, M., and Saeedi, K. Agile effort estimation based on user stories: a systematic literature review. Artificial Intelligence Review, 55 (7), 5485–5516 (2022). https://doi.org/10.1007/s10462-021-10132-x

13 Matsubara, P.G.F. SEXTAMT: a systematic map to navigate the wide seas of factors affecting expert judgment software estimates. Journal of Systems and Software, 185, 111148 (2022). https://doi.org/10.1016/j.jss.2021.111148

14 Fávero, E.M.D.B. SE3M: a model for software effort estimation using pre-trained embedding models. Information and Software Technology, 147, 106886 (2022). https://doi.org/10.1016/j.infsof.2022.106886

15 Li, X., Zhao, H., and Yu, M. Hybrid deep learning models for software cost prediction using CNN and LSTM. Journal of Systems and Software, 188, 111282 (2022). https://doi.org/10.1016/j.jss.2022.111282

16 Rosa, C.C., and Jardine, D.A. Data-driven agile software cost estimation models for DHS and DoD. Journal of Systems and Software, 203, 111739 (2023). https://doi.org/10.1016/j.jss.2023.111739

[1*]Әйтім Ә.Қ.,
PhD, қауымдастырылған профессор, ORCID:0000-0003-2982-214X,
*e-mail: a.aitim@iitu.edu.kz

[1]Халықаралық ақпараттық технологиялар университеті, Алматы қ., Қазақстан

ТРАНСФОРМЕР-НЕГІЗДІ МОДЕЛЬДЕРДІ ҚОЛДАНУ АРҚЫЛЫ ҚАЗАҚ ТІЛІНДЕГІ МОРФОЛОГИЯЛЫҚ ДИЗАМБИГУАЦИЯ

Аңдатпа

Агглютинативті тілдерде морфологиялық көпмәнділік табиғи тілді өңдеу үшін елеулі қиындық туғызады, өйткені бір ғана сөз тұлғасы бірнеше грамматикалық категорияны қамтуы мүмкін. Қазақ тілі өнімді жұрнақ-жалғау жүйесімен, дауыстылар үндестігімен және күрделі морфофонологиялық үлгілермен сипатталады, бұл автоматты морфологиялық талдауды айтарлықтай қиындатады. Бұл жұмыста қазақ мәтіндеріндегі морфологиялық дизамбигуацияға арналған трансформер-негізді әдістеме ұсынылады; мақсат – сөз тұлғасының контекст ішіндегі дұрыс морфологиялық интерпретациясын анықтау. Қазақ

тіліне бейімделген контекстік тілдік модель қолмен тексерілген жаңалық мәтіндері корпусына сүйене отырып, токен деңгейіндегі морфологиялық таңбалау үшін қайта бапталады. Ұсынылған тәсіл өзіндік-зейін механизмдері арқылы дәстүрлі ереже-негізді немесе рекуррентті нейрондық әдістермен көрсету қиын болатын ұзақ қашықтықтағы контекстік тәуелділіктерді ұстауға мүмкіндік береді. Эксперименттік бағалау трансформер-негізді модельдің ереже-негізді морфологиялық талдағыштар мен BiLSTM-негізді базалық модельдерге қарағанда жоғары дәлдік пен F1-көрсеткішіне қол жеткізетінін көрсетеді. Нәтижелер контексттендірілген эмбеддингтердің морфологиялық көпмәнділікті шешуді айтарлықтай жақсартатынын, әсіресе омоним жұрнақтар мен сирек грамматикалық құрылымдар жағдайында тиімді екенін дәлелдейді. Алынған қорытындылар трансформер архитектураларының ресурсы аз агглютинативті тілдер үшін тиімді екенін растап, морфологияға сезімтал модельдерді қазақ тілін өңдеудің кешенді жүйелеріне енгізуге негіз қалайды.

Түйін сөздер: қазақ тілі, морфологиялық дизамбигуация, трансформер-негізді модельдер, табиғи тілді өңдеу, агглютинативті тілдер.

[1*]Әйтім Ә.Қ.,
PhD, ассоциированный профессор, ORCID: 0000-0003-2982-214X,
*e-mail: a.aitim@iitu.edu.kz

[1]Международный университет информационных технологий, г. Алматы, Казахстан

МОРФОЛОГИЧЕСКАЯ ДИЗАМБИГУАЦИЯ КАЗАХСКОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ ТРАНСФОРМЕРНЫХ МОДЕЛЕЙ

Аннотация

Морфологическая неоднозначность представляет собой значимую проблему для обработки естественного языка в агглютинативных языках, поскольку одна и та же словоформа может включать множество грамматических категорий. Казахский язык характеризуется продуктивной суффиксацией, сингармонизмом и сложными морфофонологическими закономерностями, что существенно затрудняет автоматический морфологический анализ. В данной работе представлена трансформер-ориентированная методология морфологической дизамбигуации казахских текстов, направленная на выбор корректной морфологической интерпретации словоформы в контексте. Контекстная языковая модель, адаптированная для казахского языка, дообучается для токен-уровневой морфологической разметки с использованием вручную верифицированного аннотированного корпуса новостных статей. Предложенный метод применяет механизмы самовнимания (self-attention) для захвата дальних контекстных зависимостей, которые сложно эффективно представить традиционными правилами или рекуррентными нейронными подходами. Экспериментальная оценка показывает, что трансформерная модель достигает более высокой точности и F1-меры по сравнению с правилами морфологическими анализаторами и базовыми моделями на основе BiLSTM. Результаты демонстрируют, что контекстуализированные эмбеддинги заметно улучшают разрешение морфологической неоднозначности, особенно в случаях омонимичных суффиксов и редких грамматических структур. Полученные выводы подтверждают эффективность трансформерных архитектур для низкоресурсных агглютинативных языков и формируют практическую основу для интеграции морфологически чувствительных моделей в комплексные системы обработки казахского языка.

Ключевые слова: казахский язык, морфологическая дизамбигуация, трансформерные модели, обработка естественного языка, агглютинативные языки.