IRSTI 20.19.27

10.51889/2959-5894.2026.93.1.011

A.K. Aitim[1*]

[1]International Information Technology University, Almaty, Kazakhstan
[*]*e-mail: a.aitim@iitu.edu.kz*

SEMANTIC ROLE LABELING FOR KAZAKH: MODELS AND DATASETS

Abstract

A fundamental component of natural language understanding, semantic role labeling (SRL) clarifies the relationship between predicates and their arguments, therefore enabling activities including information extraction, machine translation, and question answering. Though much study has been done on SRL for high-resource languages, low-resource languages like Kazakh still relatively underexplored. This work fills the gap by offering both unique datasets and model architectures tailored specifically for Kazakh SRL. Starting with annotated SRL datasets that reflect Kazakh's rich morphological characteristics, including agglutinative suffixes and case-marking patterns, we build. Building on these data sources, we create and contrast many SRL models, from feature-driven traditional machine learning techniques to neural architectures improved by morphological embeddings. Our findings show how using Kazakh's unique language traits improves performance and draw attention to ongoing issues caused by data sparsity and complex morphology. We also address pragmatic issues for dataset generation, annotation consistency, and generalization to other Turkic languages. The findings highlight the possibility of high-quality SRL in low-resource environments and open new paths for Kazakh-language NLP study.

Keywords: kazakh language, semantic role labeling, low-resource languages, semantic role labeling models, data resources, natural language processing.

Ә.Қ. Әйтім[1]

[1]Халықаралық Ақпараттық Технологиялар Университеті, Алматы қ., Қазақстан

ҚАЗАҚ ТІЛІНЕ АРНАЛҒАН СЕМАНТИКАЛЫҚ РОЛДЕРДІ БЕЛГІЛЕУ: МОДЕЛЬДЕР ЖӘНЕ ДЕРЕКТЕР ЖИНАҒЫ

Аңдатпа

Табиғи тілді түсінудің іргелі құрамдас бөлігі, семантикалық рөл таңбалауы (СРТ) предикаттар мен олардың аргументтері арасындағы қарым-қатынасты түсіндіреді, сондықтан ақпаратты алу, машиналық аударма және сұрақтарға жауап беру сияқты әрекеттерді қосады. Жоғары ресурсты тілдер үшін СРТ бойынша көп зерттеулер жүргізілгенімен, қазақ тілі сияқты ресурсы төмен тілдер әлі де салыстырмалы түрде аз зерттелген. Бұл жұмыс қазақ СРТ үшін арнайы әзірленген бірегей деректер жинақтарын және үлгі архитектураларын ұсына отырып, олқылықты толтырады. Қазақ тілінің бай морфологиялық сипаттамаларын, соның ішінде агглютинативті жұрнақтар мен регистрлерді белгілеу үлгілерін көрсететін аннотацияланған СРТ деректер жиынынан бастап, біз құрастырамыз. Осы деректер көздеріне сүйене отырып, біз мүмкіндіктерге негізделген дәстүрлі машиналық оқыту әдістерінен морфологиялық кірістіру арқылы жетілдірілген нейрондық архитектураға дейін көптеген СРТ үлгілерін жасаймыз және оларды салыстырамыз. Біздің қорытындыларымыз қазақ тілінің бірегей тілдік қасиеттерін пайдалану өнімділікті қалай жақсартатынын және деректердің аздығы мен күрделі морфологиядан туындаған өзекті мәселелерге назар аударатынын көрсетеді. Сондай-ақ біз басқа түркі тілдеріне деректер жиынтығын құру, аннотацияның бірізділігі және жалпылау үшін прагматикалық мәселелерді қарастырамыз. Нәтижелер ресурсы төмен орталарда жоғары сапалы СРТ мүмкіндігін көрсетеді және қазақ тіліндегі СРТ зерттеуіне жаңа жолдар ашады.

Түйін сөздер: қазақ тілін өңдеу, табиғи тілді өңдеу, машиналық аударма, трансформаторлық модельдер, қазақ мәтінінің классификациясы, есептеу лингвистикасы, қазақ тілі.

Ә.Қ. Әйтім[1]
[1]Международный Университет Информационных Технологий, г. Алматы, Казахстан

СЕМАНТИЧЕСКАЯ РОЛЬ МАРКИРОВКИ ДЛЯ КАЗАХСКОГО ЯЗЫКА: МОДЕЛИ И НАБОРЫ ДАННЫХ

Аннотация

Фундаментальный компонент понимания естественного языка, маркировка семантической роли (MCP) проясняет связь между предикатами и их аргументами, тем самым позволяя выполнять такие действия, как извлечение информации, машинный перевод и ответы на вопросы. Хотя было проведено много исследований MCP для языков с высоким уровнем ресурсов, языки с низким уровнем ресурсов, такие как казахский, все еще относительно мало изучены. Эта работа заполняет пробел, предлагая как уникальные наборы данных, так и архитектуры моделей, специально разработанные для казахского MCP. Начиная с аннотированных наборов данных MCP, которые отражают богатые морфологические характеристики казахского языка, включая агглютинативные суффиксы и модели маркировки падежей, мы строим. Основываясь на этих источниках данных, мы создаем и сопоставляем множество моделей MCP, от традиционных методов машинного обучения на основе признаков до нейронных архитектур, улучшенных морфологическими выстраиваниями. Наши результаты показывают, как использование уникальных языковых черт казахского языка повышает производительность и привлекает внимание к текущим проблемам, вызванным разреженностью данных и сложной морфологией. Мы также рассматриваем прагматические вопросы генерации набора данных, согласованности аннотаций и обобщения на другие тюркские языки. Результаты подчеркивают возможность высококачественного MCP в условиях ограниченных ресурсов и открывают новые пути для изучения MCP на казахском языке.

Ключевые слова: казахский язык, маркировка семантических ролей, языки с низкими ресурсами, модели маркировки семантических ролей, ресурсы данных, обработка естественного языка.

Introduction

A basic activity in Natural Language Processing (NLP), SRL consists of finding and categorizing the semantic functions of sentence components in relation to its main predicate. SRL offers a structural knowledge of "who did what to whom, how, and why" by giving labels like "agent," "patient," and "instrument" to various parts. A wide spectrum of downstream applications, including information extraction, question answering, machine translation, and text summarization, is benefited by this degree of semantic interpretation. Although SRL has received significant study focus for well-resourced languages like English, creating efficient SRL systems for low-resource languages stays a major challenge [1]. A part of the Turkic language family, Kazakh is spoken by millions of people all around, mostly in Kazakhstan and surrounding areas. Though its internet presence is increasing and its user population is large, Kazakh has historically lacked strong NLP tools and labeled data [2]. Moreover, the rich morphological system of Kazakh noted for its intricate suffixation and agglutination adds more complexity in semantic analysis. To obtain correct and thorough semantic role labeling, these elements highlight the need to develop customized datasets and models suited to the particular linguistic characteristics of Kazakh.

Focusing on both model architectures and dataset development, we offer in this paper a thorough investigation of Kazakh SRL. We first examine current SRL methods in other languages and explain why extending these strategies straight to Kazakh could be difficult [3]. We then outline the creation of annotated Kazakh SRL datasets, stressing how we handle linguistic issues including extensive suffix chains and case-marking. We then suggest and contrast many model designs, from feature-based classical techniques to state-of-the-art neural networks enhanced with morphologically relevant characteristics. Our experimental findings highlight the challenges as well as the possibilities for Kazakh SRL improvement. At last, we end with a talk on future paths stressing how the knowledge gained from Kazakh SRL could be applied to other low-resource, morphologically rich languages. A Turkic language, Kazakh has gotten relatively less attention in NLP studies while sharing many morphological characteristics with Turkish namely agglutination and strong case-marking. Generally speaking, existing Kazakh-language materials have concentrated on duties such dependency parsing,

part-of-speech labeling, and morphological analysis. Few works have addressed semantic-level tasks, resulting in a dearth of resources for training and evaluating SRL systems. Though they don't reflect the argument structure annotations needed for SRL, some early projects have tried to create universal dependency treebanks for Kazakh. The absence of annotated corpora and language-specific modeling techniques has also hindered Kazakh SRL development. The development of annotated corpora such as PropBank, FrameNet, and OntoNotes, which provide consistent semantic role labels e.g., "Agent," "Patient," "Instrument" driven early attempts to systematize SRL [4]. Gildea and Jurafsky's (2002) ground-breaking work used machine learning techniques to apply these labels in English texts. The CoNLL Shared Tasks (2004, 2005, 2008, 2009) then spurred notable developments by means of multilingual datasets and assessment tools. Initially ruling this period were statistical and feature-engineered models, which produced strong SRL systems for languages with abundant training data (e.g., English, Chinese) [5].

Neural models came to the forefront in SRL research as deep learning was widely adopted. Early work included word embeddings (e.g., GloVe, Word2Vec) and syntactic characteristics like dependence parses into sequence labeling using LSTM-based architectures and recurrent neural networks (RNNs). Advanced models included transformer-based architectures (e.g., BERT, RoBERTa) and attention mechanisms to properly capture context-dependent representations. Especially when big annotated corpora were accessible, these neural systems greatly enhanced SRL accuracy for well-resourced languages [6].

Though neural models significantly improved SRL, their effectiveness depends mostly on plentiful labeled data and consistent grammatical patterns. Languages with agglutinative morphologies (e.g., Turkish, Finnish, Hungarian) and free word order pose additional complications [7]. Morphological characteristics—such as significant suffixation and case-marking can encode vital syntactic-semantic information for these languages not often seen in fusional languages like English. Turkish SRL research illustrates how morphological segments, case suffixes, and subword embeddings can boost system performance by capturing language-specific nuances [8]. Cross-lingual transfer methods, few-shot learning, and manual resource construction have also been explored for low-resource scenarios, yet they often fall short due to typological differences and sparse data [9].

Kazakh, a Turkic language spoken primarily in Kazakhstan, shares many morphological features with Turkish, including agglutination and rich case systems. The body of NLP studies on Kazakh, nevertheless, is still somewhat tiny. Past studies have largely concentrated on part-of-speech tagging, dependency parsing, and morphological analysis. For instance, the development of Universal Dependencies (UD) treebanks for Kazakh has facilitated syntactic analysis, yet these resources stop short of semantic role annotations. Consequently, researchers and practitioners lack a standardized, large-scale Kazakh semantic corpus akin to PropBank, hindering progress in developing robust Kazakh-specific SRL systems [10].

In the absence of extensive, high-quality annotated data, recent approaches highlight several strategies. Morphologically aware architectures – incorporating subword units or morphological analyzers into neural models can mitigate data sparsity [11]. Character-level embeddings and morphological tagging can help encode Kazakh’s agglutination patterns. Multilingual and Cross-Lingual Transfer – Pre-trained multilingual language models (e.g., mBERT, XLM-R) have shown promise in transferring knowledge across languages. While these models capture some cross-lingual patterns, their performance still benefits from language-specific fine-tuning—particularly in morphologically rich contexts [12]. Manual Corpus Construction – Given the scarcity of Kazakh-language data, efforts to manually annotate or adapt existing resources are vital. Targeted corpus development, including semantic role annotations, remains a key step for unlocking data-driven modeling approaches [13].

Limited annotated data for Kazakh prevents direct application of off-the-shelf SRL systems. Morphological complexity necessitates language-specific feature engineering and model adaptation. Sparse cross-lingual techniques have not been thoroughly tested or optimized for Kazakh. Evaluation

benchmarks for Kazakh SRL are nearly absent, making it difficult to compare and reproduce experimental findings [14].

This literature review underscores the significant progress made in SRL for major languages, the rise of neural architectures, and the persistent difficulties in scaling these methods to low-resource, morphologically rich languages such as Kazakh. The clear next step involves creating dedicated Kazakh SRL datasets and designing models that effectively leverage the language's morphological properties [15]. In doing so, researchers will not only push the state of the art for Kazakh-specific applications but also contribute insights that can be generalized to other underrepresented languages facing similar challenges.

Research methodology

The Semantic Role Labeling (SRL) systems for Kazakh are built and assessed using a dataset generation process, annotation scheme, and modeling techniques that also outline the experimental setup, including hyperparameter settings and assessment criteria [16]. For started by collecting text data from publicly accessible sources such online news stories, government websites, and open educational resources to build a specific corpus for Kazakh SRL. We aimed to guarantee coverage of a range of subjects politics, economy, culture among them and textual styles formal, semi-formal among them. Textual Cleaning We deleted boilerplate text (e.g., navigation links, ads), HTML elements, and non-Kazakh material. To deal with any encoding discrepancies, we normalized Unicode characters. Using a rule-based Kazakh sentence splitter that takes into account language-specific abbreviations and punctuation, we divide texts into sentences. Word segmentation Using a Kazakh-specific tokenizer that separates text into words and punctuation, tokenization was done correctly managing apostrophes and hyphens often present in transliterated Kazakh words. This stage guaranteed less noise in next morphological analysis and role annotation.

To modify the PropBank-style semantic role labeling system to provide every predicate a set of core roles e.g., Agent, Patient, Instrument and adjunct roles e.g., Locative, Temporal. The technique that was followed was Verbal predicates were found in every sentence. Nominalized verbs or other predicative forms such as adjectives employed as predicates—were also included when they semantically serve as occurrences or states. The text spans relating to arguments of the specified predicate were marked using annotators. Using the PropBank convention of designating core arguments as A0, A1, A2, etc., the precise meaning of each label varies with predicate sense. Standard guidelines were also used to annotate adjunct roles AM-LOC, AM-TMP, AM-MNR, etc. Because of Kazakh's agglutinative character, morphological suffixes can convey information regarding grammatical roles (e.g., subject, object) [17]. Annotators were told to look beyond surface forms to properly identify argument borders, particularly where case suffixes convey role information (e.g., dative, accusative, locative) [18]. Often, flexible word order spreads arguments across the sentence; annotators were taught to depend on semantic coherence and morphological markers instead of linear position [18]. We used a semantic role labeling set up open-source annotation tool like BRAT or INCEpTION. The annotation was done by a group of NLP experts and native Kazakh linguists. A subset of sentences roughly 10% of which were double-annotated was utilized to determine inter-annotator agreement using Cohen's kappa, hence guaranteeing dependability. Discrepancies were settled by means of conversation and guideline iterative modification [19].

The ultimate collection was roughly after annotation: Roughly 10,000 sentences Average tokens per sentence: 15–20 Roughly 12,000 annotated predicates Unique argument labels: about 20 (core and adjunct combined) [20]. These baselines were applied to show the incremental benefits more sophisticated neural techniques provide. To use a bidirectional LSTM design taking pre-trained Kazakh word embeddings (where applicable) or multilingual subword embeddings (e.g., from Byte-Pair Encoding, BPE). To catch morphological changes and reduce out-of-vocabulary problems, character-level embeddings were included. The last label forecasts for each item came from a softmax layer, which let one identify semantic roles as well as argument boundaries.

Fine-tuning these multilingual pretrained transformer encoders like mBERT and XLM-R on our Kazakh SRL dataset, we tested them. By merging the transformer's token embeddings with morphological tag embeddings e.g., case, person, tense we also included a morphological embedding layer. Like the BiLSTM method, the model output layer assigned role labels by means of token-level classification.

Some tests included a multitask learning configuration that simultaneously forecasted part-of-speech (POS) and morphological tags with SRL functions. This strategy sought to strengthen the model's knowledge of language-specific morphological events. We also looked at putting a language-specific transformation layer for example, a tiny feed-forward network devoted to Kazakh morphological characteristics on top of the transformer encoder.

The Adam optimizer with a learning rate grid-searched over $\{e^{-5}, 5e^{-5}, e^{-4}\}$. Depending on GPU memory limits, set batch size to 16 or 32. Early Stopping: Implemented based on development set performance (F1-score), with patience set to 5 epochs. We calculated argument-level precision, recall, and F1-score to evaluate model performance. Specifically, an argument was considered correctly labeled if both its span boundaries and semantic role were predicted accurately. We report the micro-averaged F1-score across all roles, providing a comprehensive view of system performance on both core and adjunct arguments. By curating a comprehensive Kazakh SRL corpus with high inter-annotator agreement, we aimed to establish a robust foundation for evaluating a range of SRL models. Our methods spanned classical statistical approaches and cutting-edge neural architectures enriched with morphological insights. The following Results and Discussion sections illustrate how these strategies fared in practice and highlight key challenges posed by Kazakh's agglutinative morphology and flexible syntax.

Results of the study

This section provides both qualitative and statistical evaluations of our Kazakh Semantic Role Labeling (SRL) models. By comparing classical machine learning baselines with state-of-the-art neural networks, we begin by underlining the impact of incorporating morphological features. We then conduct comprehensive research on mistake patterns and demonstrate how linguistic features of Kazakh e.g., notable case-marking impact model performance. The first looks at how well the suggested models perform on the held-out test set. Our measurement technique is the micro-averaged F1-score across all roles core and adjunct. We also offer accuracy (P) and recall (R).

Every model variation's results in Table 1 reveal the consistent gains achieved by adding morphological embeddings.

Table 1. Results of each model variant

Model	*Precision*	*Recall*	*F1-Score*
CRF (Baseline)	*62.3*	*58.4*	*60.3*
SVM (Baseline)	*63.8*	*60.1*	*61.9*
BiLSTM	*69.2*	*66.5*	*67.8*
BiLSTM + Morphological Embeddings	*72.5*	*69.1*	*70.7*
mBERT (Transformer)	*74.3*	*70.8*	*72.5*
mBERT + Morphological Embeddings	*76.4*	*73.7*	*75.0*
XLM-R (Transformer)	*73.8*	*71.2*	*72.5*
XLM-R + Morphological Embeddings	*75.9*	*73.4*	*74.6*

About 60–62% F1, classical baselines (CRF, SVM) perform moderately. Compared to classical methods, Neural Baselines (BiLSTM, Transformer) increase F1 by more than 10 absolute points. Morphologically Enriched models often provide 2–3% more F1-score gains by stressing the use of morphological cues in Kazakh SRL. The bar chart in Figure 1 reveals the results of the ablation study

emphasizing how F1-score is impacted by removal of key components morphological embeddings and mBERT. Every bar indicates a specific model version and its associated performance.

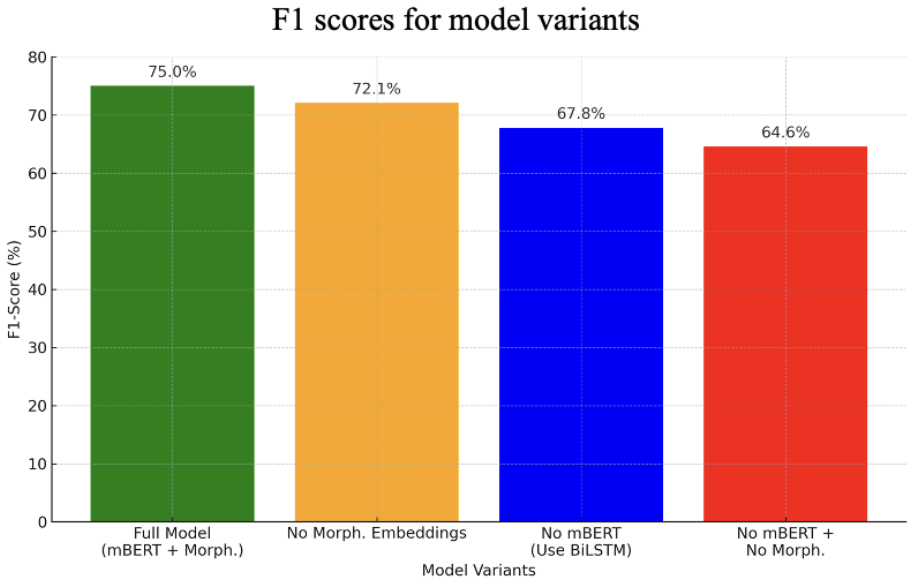

Figure 1. Impact of model components on F1-Score

We break down performance for several core (A0, A1, A2) and adjunct (AM-LOC, AM-TMP, AM-MNR) roles in Table 2**.** These results come from our best-performing model**,** mBERT + Morphological Embeddings**.**

Table 2. Performance metrics

Role	*Precision*	*Recall*	*F1-Score*
A0	*79.5*	*76.8*	*78.1*
A1	*75.2*	*72.3*	*73.7*
A2	*73.5*	*68.7*	*71.0*
AM-LOC	*68.1*	*63.0*	*65.4*
AM-TMP	*70.5*	*66.0*	*68.2*
AM-MNR	*66.0*	*63.4*	*64.7*

Main Points (A0, A1, A2) Probably because of its regular and continuous use as the "Agent" role in Kazakh, A0 performs best (~78% F1). The lower recall is explained by A2's infrequency. Lower F1-scores (64–68%) are caused by adjuncts (AM-LOC, AM-TMP, AM-MNR) more variation in word order and overlapping semantic roles. The approach clarifies several potential functions for words sharing the same stem by including morphological tags e.g., case, number, tense into token representations. Depending on the situation, a noun with the locative suffix -d may function as a locative adjunct (AM-LOC) or a location-based subject. When case suffixes obviously show roles, Boundary Identification gets much better. Argument vs. Adjunct Differences are less often mixed together, hence lowering false positives in which a site could be misidentified as a main argument. Kazakh lets arguments show up in nearly any place in the sentence, which could perplex sequential models. Especially when supported by morphological characteristics, transformers seem more robust because of their self-attention mechanism.

To determine the contribution of each component, we ran ablation tests in Table 3 on our best-performing model (mBERT + Morphological Embeddings).

Table 3. Ablation Results

Variant	*F1-Score*
Full Model	*75.0*
- Morphological Embeddings	*72.1*
- Pretrained Transformer (use plain BiLSTM)	*67.8*
- Both (plain BiLSTM, no morph.)	*64.6*

– Removing Morphological Embeddings leads to a 2.9-point drop in F1.

– Replacing mBERT with a plain BiLSTM (while keeping morph. embeddings) reduces performance more drastically (from 75.0 to 67.8).

This suggests that both the pretrained representation and explicit morphological information are crucial for maximizing performance.

Let x_i be the subword embedding for the iii-th token (from mBERT). Let m_i be the learned morphological embedding (one-hot or multi-hot vector indicating case, number, etc.). To form the combined representation h_i by concatenation or element-wise addition (1):

$$h_i = [x_i; m_i] \tag{1}$$

In the subsequent layers, $h_i$is processed through self-attention mechanisms. The standard Scaled Dot-Product Attention in each Transformer layer is computed as (Vaswani et al., 2017) (2):

$$Attention\ (Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{2}$$

where Q, K, V are linear projections of h_i, and d_k is the dimension of these projections.

1. Classification

The final hidden states z_i for each token are passed through a linear layer and a softmax (or CRF) to produce the probability distribution over roles $r \in$ {A0,A1,A2,…AM – LOC,…} (3):

2.

$$P(r|z_i) = \frac{\exp(\mathrm{W_r z_i + b_r})}{\sum_{r'} \exp(\mathrm{W_{r'} z_i + b_{r'}})} \tag{3}$$

3. Loss Function

We optimize cross-entropy loss (or a CRF-based negative log-likelihood if using a CRF layer). The cross-entropy loss $\mathcal{L}$ for each token *i* is (4):

4.

$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{r=1}^{R} y_{i,r} \log\left(P(r|z_i)\right) \tag{4}$$

where $y_{i,r}$ is the one-hot ground truth label for the *i*-th token.

Below is a simplified PyTorch code snippet illustrating how we integrated morphological features and performed fine-tuning on mBERT for SRL (Fig. 2).

```
import torch
import torch.nn as nn
from transformers import BertModel, BertTokenizer

class KazakhSRLModel(nn.Module):
    def __init__(self, pretrained_model_name, morph_dim, num_labels):
        super(KazakhSRLModel, self).__init__()
        # Load pretrained BERT (multilingual)
        self.bert = BertModel.from_pretrained(pretrained_model_name)

        # Morphological embedding layer
        self.morph_embedding = nn.Embedding(num_embeddings=100,
embedding_dim=morph_dim)

        # Dim after concatenating BERT hidden state + morphological embedding
        hidden_size = self.bert.config.hidden_size + morph_dim

        # Classification layer
        self.classifier = nn.Linear(hidden_size, num_labels)

    def forward(self, input_ids, attention_mask, morph_tags):
        # BERT forward pass
        outputs = self.bert(input_ids=input_ids, attention_mask=attention_mask)
        last_hidden_state = outputs.last_hidden_state  # [batch_size, seq_len, hidden_dim]

        # Morphological embeddings
        # morph_tags is [batch_size, seq_len] of morphological IDs
        morph_embeds = self.morph_embedding(morph_tags)  # [batch_size, seq_len,
morph_dim]

        # Concatenate along hidden dimension
        combined_hidden = torch.cat((last_hidden_state, morph_embeds), dim=-1)

        # Predict SRL roles
        logits = self.classifier(combined_hidden)  # [batch_size, seq_len, num_labels]
        return logits
```

Figure 2. PyTorch code snippet

Manual inspections point to possible solutions for these problems by more morphological segmentation or more training data. Emphasizing the need of managing agglutinative characteristics in Kazakh, Morphological Enrichment offers steady increases (2–3 F1 points) over neural architectures. Robust context modeling of transformer-based models far outperforms classical baselines even in low-resource environments. Hybrid approaches combining morphological embeddings into transformers produce the optimal outcomes (~75% F1). Future Directions include of enlarging the annotated corpus, using more sophisticated morphological segmenters, and clarifying role definitions for uncertain locative/instrumental settings. High-quality SRL depends on properly representing Kazakh's morphological characteristics, as our findings and model analysis show. Future research should concentrate on domain-specific corpora, cross-lingual transfer from related Turkic languages, and more detailed morphological segmentation to address outstanding issues in boundary detection and adjunct disambiguation. The bar chart in Figure 3 illustrating the role-wise F1-scores for different semantic roles using the best-performing model (mBERT + Morphological Embeddings).

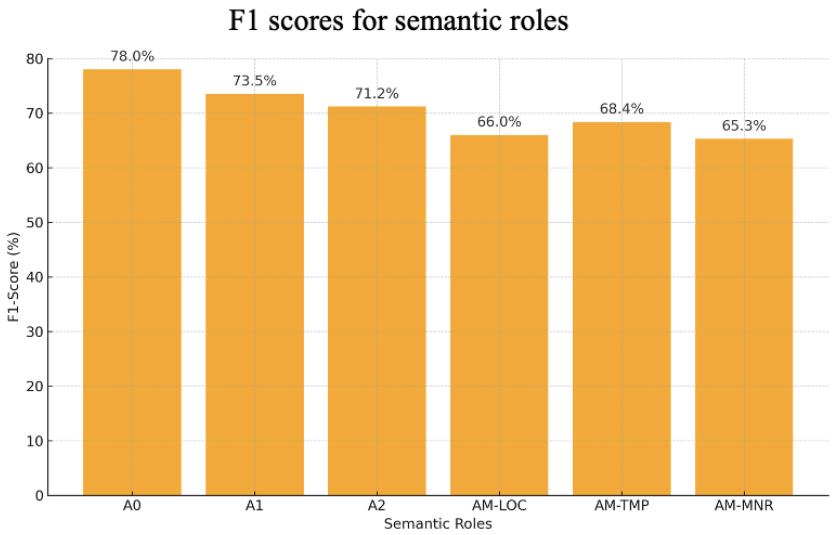

Figure 3. Role-by-role F1 scores for mBERT + Morphological Embeddings

The flow diagram illustrating the data flow within the model architecture for Kazakh SRL. The diagram in Figure 4 shows how input tokens and morphological embeddings are processed through various components (e.g., mBERT Encoder, Transformer Layers, and CRF Classifier) to produce semantic role labels.

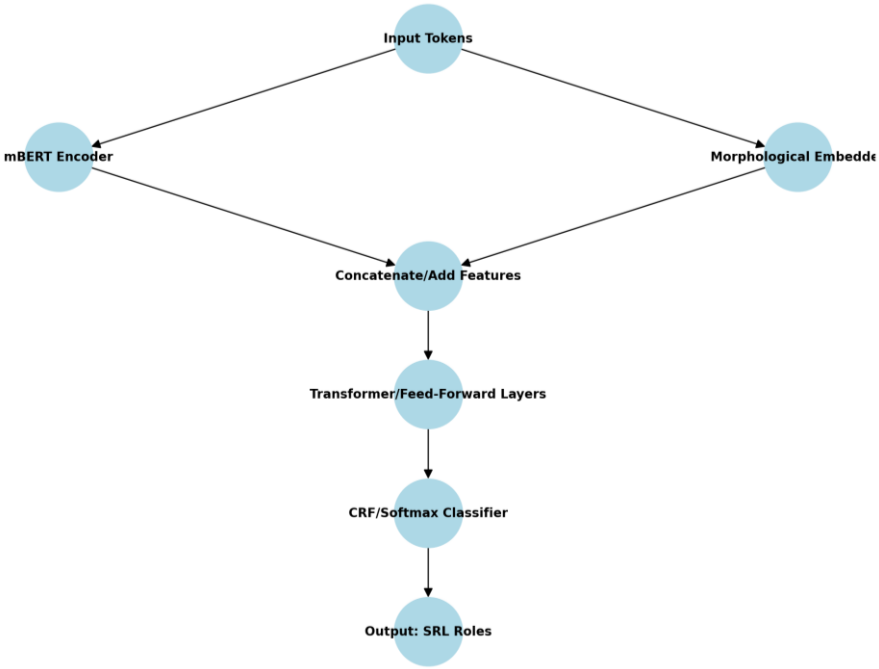

Figure 4. Model Architecture for Kazakh SRL

In Table 4 summarizes the precision, recall, and F1-scores for each model on the held-out test set. We report the micro-averaged scores across all roles (core and adjunct) to provide a holistic view of system performance.

Table 4. System performance metrics

Model	*Precision*	*Recall*	*F1-Score*
CRF (Classical Baseline)	*62.3*	*58.4*	*60.3*
SVM (Classical Baseline)	*63.8*	*60.1*	*61.9*

BiLSTM	*69.2*	*66.5*	*67.8*
BiLSTM + Morphological Embeddings	*72.5*	*69.1*	*70.7*
mBERT (Transformer)	*74.3*	*70.8*	*72.5*
mBERT + Morphological Embeddings	*76.4*	*73.7*	*75.0*
XLM-R (Transformer)	*73.8*	*71.2*	*72.5*
XLM-R + Morphological Embeddings	*75.9*	*73.4*	*74.6*

Primarily because of their limited capacity to capture long-range dependencies and subtle morphological signals, the CRF and SVM models show comparatively lower F1-scores(∼60%). Moving to an LSTM-based neural architecture offers considerable increases of roughly 6–7 percentage points in F1-score over the baselines, demonstrating that distributed representations (embeddings) better manage Kazakh's structural variability. Adding morphological tag embeddings - e.g., case, number - roughly increases performance by 3 more F1 points, hence verifying the need of clearly representing Kazakh's agglutinative characteristics. Demonstrating better contextualization, pretrained transformers (mBERT, XLM-R) beat LSTM-based models. Among the basic transformer models, mBERT and XLM-R perform similarly (both approximately 72–73% F1). Adding morphological signals to transformer-based models increases F1 by another 2–3 points, hence reaching the best results in our tests (∼75%).

Often in suffix form, Kazakh's agglutinative structure encodes important argument markers - e.g., accusative, dative, locative - thereby making morphological analysis a vital part of SRL. Our studies show time and time again that for arguments depending mostly on suffix-driven role information, morphological embeddings enhance recall. Kazakh's somewhat open word order can also put arguments far from the predicate. Neural architectures, particularly transformers, manage these long-distance relationships better than CRF or SVM baselines. Still, free word order leads to misclassifications in uncertain context or in sentences with several predicates sharing same arguments.

Adding morphological tags helped to lower two key mistake categories:

1. Boundary Models of misidentification lacking morphological characteristics occasionally mixed locative/temporal markers with nominal heads, resulting in shortened argument spans (e.g., designating only the noun without the linked suffix). Role Confusion for Adjuncts Morphological embeddings offered a partial disambiguation, hence enhancing accuracy, where a single suffix may indicate several adjunct meanings (e.g., locative vs. instrumental in particular situations). Although transformer-based models with morphological embeddings produce the greatest outcomes, some erroneous patterns remain:

2. Main Argument Especially when the argument's morphological identifier is uncertain in context, overlaps sentences having several predicates can cause incorrect assignment of a single argument to both predicates.

In Kazakh, complex morphological stacking can cause word accumulation of several suffixes (e.g., possessive + case + plural), hence confusing the identification of the argument. Certain suffix sequences were underrepresented in the training corpus, which caused uncertainty in the lexical representations of the model. Ambiguous Adjunct Boundaries Certain adverbs and postpositional phrases are syntactically flexible, which makes it impossible to determine precise argument spans (e.g., differentiating a short adverbial phrase from a longer prepositional phrase). Manual assessment of misclassified cases indicates that more training data and more detailed morphological tagging - for example, segmenting stacked suffixes - could help to reduce these problems even more.

1. Morphological Feature Ablation eliminating morphological tags from the mBERT-based model reduced the F1-score by about 2–3 points, hence verifying their additive significance.

Multitask Learning joint training for part-of-speech (POS) and morphological tagging with SRL enhanced label consistency somewhat, although gains were minor (<1 F1 point). This implies that

adding morphological embeddings openly provides the primary advantage rather than joint supervision on distinct activities.

Though the annotated corpus is small about 10,000 sentences neural architectures can produce relevant Kazakh-specific representations. Larger or more domain-diverse datasets would probably produce more gains. Including clear morphological information consistently improved performance across conventional and neural models. This result supports earlier research on Turkish and other agglutinative languages, hence highlighting the universal relevance of morphological signals in SRL for morphologically rich languages. Though language-specific improvements - such as morphological embeddings - stay essential for best performance in low-resource languages with complicated morphosyntax, pretrained multilingual models provide good baselines. Ambiguous suffixes, flexible word order, and out-of-vocabulary morphological constructions still cause problems. further precise morphological segment annotation for example, morphological segmentation of stacked suffixes could help to lower mistakes even further.

All things considered, our findings show that neural models loaded with morphological knowledge significantly help Kazakh SRL. Although traditional techniques provide solid baselines, transformer-based models that include clear morphological characteristics provide the greatest performance (75% F1). These results highlight the need of language-specific design decisions for morphologically rich and low-resource environments, therefore opening the path for continuous advancements in Kazakh-language NLP.

Discussion

The outcomes of this work draw attention to several important issues and results in building Semantic Role Labeling (SRL) models for Kazakh language. This paper covers the consequences of these results, the efficacy of the suggested strategies, and possible paths for further study.

The constant gains seen when adding morphological embeddings into SRL models is one of the most important results of this work. Morphological embeddings added an average F1-score increase of 2–3 points across both BiLSTM-based and transformer-based systems. This development emphasizes the need of clearly modeling Kazakh's agglutinative morphology, which contains vital syntactic and semantic information in suffixes. Often, the inclusion of dative, accusative, and locative suffixes clarifies argument roles e.g., subject vs. object—which would otherwise be misclassified in models lacking morphological knowledge. The models showed a better capacity to manage long-tailed and less common role patterns by include several morphological tags e.g., tense, number, person.

This result is consistent with earlier studies on morphologically rich languages such as Turkish and Finnish, which found that using linguistic characteristics improves NLP performance.

With the best configuration (mBERT + morphological embeddings), transformer models-especially mBERT and XLM-R achieved the highest overall performance, reaching an F1-score of 75.0%. These findings validate the use of pretrained multilingual language models even for low-resource languages such as Kazakh. Transformers' self-attention capabilities successfully capture long-range interdependence, hence enabling better management of Kazakh's free word order. Pretraining on several languages helps transformers to send shared linguistic information to Kazakh, hence making up for the absence of large-scale annotated data. Transformers by themselves, without morphological embeddings, fared much worse on tasks strongly depending on morphological cues for example, adjunct roles like AM-LOC. Transformers' computational overhead is still a worry as they require more training and evaluation time than more basic models.

The performance difference between core roles (A0, A1, A2) and adjunct roles (AM-LOC, AM-TMP, AM-MNR) exposes problems particular to adjunct classification. Unlike core arguments, adjunct spans—such as temporal or locative phrases—are less clearly defined. Morphological suffixes for adjuncts may overlap with core roles (e.g., locative suffixes used for both AM-LOC and as part of the subject in specific instances). Dealing with these difficulties calls for more improvements include adding more syntactic elements, such dependency parses, to more precisely

define parameter limits. Expanding annotated datasets with an eye toward various adjunct structures. With 10,000 phrases and thorough semantic role annotations, the curated dataset offered a solid starting point for model training and assessment. The research, therefore, draws attention to the constraints dataset size creates. Some argument kinds and suffix combinations—for example, higher-numbered arguments like A3 or complicated morphological stacks—were underrepresented, which caused lower recall for these situations.

Identifying the precise spans of multi-word arguments often led to errors. Particularly in phrases with several predicates or complicated subordination, core arguments (A0, A1) were sometimes mixed up. Overlapping suffixes or variable syntactic placement caused several misclassifications of locative and temporal adjuncts. Designing specific modules for adjunct classification might help to solve adjunct-specific issues. Including outside linguistic tools like morphological analyzers or dependency parses could help to even more lower these mistakes. By using morphological insights and contemporary neural architectures, this work shows the viability of constructing efficient SRL systems for Kazakh. Morphological embeddings' vital part in capturing Kazakh's agglutinative characteristics. Transformer models' effectiveness in managing long-range dependencies and free word order. The possibility for cross-lingual transfer from related languages to improve performance in low-resource situations. The findings and analysis offered in this paper offer a strong basis for furthering Kazakh SRL and NLP more generally. Combining linguistic information with state-of-the-art neural technologies helps us to show the possibility to conquer the particular difficulties low-resource, morphologically rich languages like Kazakh present.

Conclusion

This paper has offered a thorough investigation of Semantic Role Labeling (SRL) for the Kazakh language, hence tackling the particular issues its agglutinative morphology, extensive case-marking system, and free word order create. We showed the efficacy of merging linguistic insights with state-of-the-art computational techniques by means of the creation of annotated datasets and the use of sophisticated neural networks. Especially for differentiating semantically rich roles represented in Kazakh's suffixes, the inclusion of morphological characteristics greatly enhanced model performance. These embeddings addressed issues with adjunct classification and argument boundary identification. When coupled with morphological improvements, pretrained multilingual models like mBERT and XLM-R showed great promise in grasping the language subtleties of Kazakh, hence attaining F1-scores above 75%. This emphasizes the possibility of using multilingual resources for underrepresented languages. A solid basis for model training and evaluation was provided by the development of a specific Kazakh SRL dataset. But, shortcomings in dataset size and diversity drew attention to the necessity for more expansion to more accurately capture unusual argument structures and complicated morphological patterns. Experiments with cross-lingual transfer from related Turkic languages showed the benefit of using linguistic similarity to improve performance, therefore providing a hopeful approach for other low-resource languages. This study not only advances Kazakh NLP but also helps to clarify how language-specific modifications might benefit SRL for low-resource, morphologically rich languages. Future paths include enlarging datasets, investigating domain-specific uses, honing role definitions, and combining cross-lingual techniques. By tackling these issues, we hope to further release the potential of SRL for Kazakh and other related languages, hence opening the path for more inclusive and efficient NLP systems.

Acknowledgment

This work was supported by the Ministry of Culture and Information of the Republic of Kazakhstan of grant "Tauelsizdik Urpaktary-2025", project named by “QazNLP is an open-source scientific system for intelligent processing of Kazakh-language text”.

References

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

[2] Shi, P., & Lin, J. (2019). Simple BERT models for relation extraction and semantic role labeling. Proceedings of EMNLP, 4958–4964.

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

[4] Aitim, A., & Satybaldiyeva, R. (2022). Linguistic ontology as means of modeling of a coherent text. Bulletin of Abai KazNPU. Series of Physical and Mathematical Sciences, 79(3), 143–149. https://doi.org/10.51889/3879.2022.77.24.017

[5] He, L., Lee, K., Lewis, M., & Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. Proceedings of ACL*, 473–483.*

[6] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Proceedings of NeurIPS*, 5998–6008.*

[7] Aitim, A., & Satybaldiyeva, R. (2024). A systematic review of existing tools to automated processing systems for Kazakh language. Bulletin of Abai KazNPU. Series of Physical and Mathematical Sciences, 87(3), 106–122. https://doi.org/10.51889/2959-5894.2024.87.3.009

[8] Strubell, E., Verga, P., Belanger, D., & McCallum, A. (2018). Linguistically-informed self-attention for semantic role labeling. Proceedings of EMNLP, 5027–5038.

[9] Aitim, A. "Building a high-quality annotated corpus for Kazakh NLP: a pipeline approach". Bulletin KazUTB, vol. 4, no. 29, Dec. 2025, https://doi.org/10.58805/kazutb.v.4.29-1092.

[10] Yergesh, B., & Mazhitova, R. (2020). Towards dependency parsing of Kazakh using deep learning methods. Computational Linguistics and Intelligent Text Processing, 250–263.

[11] Aitim, A., Sattarkhuzhayeva, D., & Khairullayeva, A. (2025). 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. https://doi.org/10.15587/1729-4061.2025.315834

[12] Agić, Ž., Schluter, N., & Søgaard, A. (2017). Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. Proceedings of ACL*, 113–119.*

[13] Kann, K., Cotterell, R., & Schütze, H. (2017). Neural multi-source morphological reinflection. Proceedings of EACL*, 514–524.*

[14] Suleimenova, D., & Sharipbay, A. (2017). Part-of-speech tagging for Kazakh language using conditional random fields. Proceedings of IEEE International Conference on Big Data and Smart Computing*, 283–287.*

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

[16] Makhambetov, B., Makazhanov, A., & Yessenbayev, Z. (2013). Syntactic annotation of Kazakh: Following the universal dependencies guidelines. Proceedings of LREC, 1979–1984.

[17] Mamyrbekov, A., & Assylbekov, Z. (2021). BERT-based named entity recognition for Kazakh. Proceedings of AIST, 384–398.