КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 004.89
IRSTI 50.01
https://doi.org/10.53364/24138614_2026_41_2_13

A. Aidynkyzy
Astana IT University, Astana, Kazakhstan

E-mail: aidana.aidynkyzy@astanait.edu.kz

FROM NAMED ENTITY RECOGNITION TO RELATION EXTRACTION: LARGE LANGUAGE MODEL ASSISTED CONSTRUCTION OF THE KAZAKH RELATION EXTRACTION DATASET

Abstract. *Relation Extraction is a fundamental bridge between unstructured text and formal knowledge representations. Its development for the Kazakh language has been hindered by the scarcity of high-quality annotated semantic resources. While the KazNERD dataset established a robust baseline for entity identification, the transition toward modeling complex interactions between entities remains a critical challenge. This study addresses this bottleneck by introducing the Kazakh Relation Extraction Dataset (KRED), a high-fidelity benchmark constructed through a scalable pipeline that leverages the synergistic capabilities of Large Language Models (LLMs) and human expertise. The annotation workflow used the KazNERD corpus's verified entity boundaries as a structural base. It included candidate pair generation, zero-shot prompting using GPT-4o-mini, and iterative semantic refinement. Schema-driven normalization and targeted re-annotation with Gemini-3-flash, followed by manual verification, ensured linguistic accuracy. The resulting KRED dataset contains 16,149 relation instances across ten semantic categories. Experiments using transformer architectures such as multilingual BERT, XLM-RoBERTa, and Kaz-RoBERTa show the dataset's effectiveness. Multilingual BERT performed best, achieving a micro-F1 score of 0.8832 and a macro-F1 score of 0.8113, which provides a solid baseline for future work. This hybrid approach, which uses LLMs, offers a cost-effective alternative to manual labeling. It provides a methodological framework for quickly expanding information extraction resources in low-resource and Turkic languages.*

Keywords: *natural language processing, information retrieval, low-resource languages, large language models, dataset construction, relation extraction.*

Introduction

Relation Extraction (RE) is a central task in natural language processing (NLP), aiming to identify and categorize semantic relationships between entities mentioned in text. Accurate relation modelling is the bridge between raw text and structured intelligence, enabling critical downstream applications such as knowledge base construction, complex question answering, advanced information retrieval, and reasoning over structured knowledge [1]. Over the past decade, the development of large, annotated benchmarks has played a decisive role in advancing RE systems. Large-scale resources, such as DocRED [2], introduced document-level reasoning challenges, while subsequent analytical evaluations like TACRED revisited [3] demonstrated how dataset noise and annotation quality substantially affect model evaluation. These efforts have highlighted a crucial insight: progress in relation extraction is increasingly driven not merely by model architectures, but by the quality and methodology of data annotation.

Recent surveys confirm that RE research has entered a new phase shaped by Large Language Models (LLMs), which enable stronger contextual reasoning and reduce reliance on handcrafted features [4]. However, despite these rapid methodological advances, most available datasets remain concentrated in high-resource languages, particularly English and Chinese. For low-resource languages, the scarcity of high-quality annotated corpora continues to be the primary bottleneck limiting progress [5, 6]. Traditional manual annotation pipelines are expensive, slow, and notoriously difficult to scale, especially when dealing with the complex linguistic structures of morphologically rich languages. This has motivated recent work exploring cross-lingual transfer and collaborative data augmentation using a synergy between large and small models [7].

The Kazakh language represents a particularly challenging and illustrative case within the low-resource NLP landscape. Although foundational processing pipelines such as KazNLP [8] have enabled basic tokenization and linguistic preprocessing, the bulk of research has historically focused on lower-level tasks such as lexicon-free stemming [9], keyword extraction [10], or general information retrieval [11], rather than deep semantic relation understanding. While there is a growing interest in integrating AI techniques for Kazakh and developing small-scale language models adapted to limited data conditions [12, 13], structured semantic datasets for relation extraction have remained extremely scarce.

A recent investigation by [14] introduces a method for extracting information from scientific documents spanning multiple domains, thereby contributing to the growing field of Kazakh information extraction. That study stands out due to its careful manual annotation and its concentration on highly specialized scientific texts, offering meaningful insights into domain-restricted information extraction. Nonetheless, such an approach is notably resource-heavy and generally suffers from limitations in terms of scalability and breadth of domain coverage. By contrast, the creation of general-purpose relation extraction resources for Kazakh remains significantly underexplored.

The appearance of the KazNERD dataset [15] constituted an important step forward by supplying large-scale annotated entities for Kazakh texts, drawing on earlier efforts in Kazakh named entity recognition [16]. Even so, NER by itself only delineates entity boundaries and types, leaving the semantic connections among those entities unaddressed. Many practical applications – for instance, event understanding, knowledge graph construction [17], and reasoning at the document level – depend on explicit relational data rather than on isolated entity mentions. Consequently, repurposing existing NER resources to create relation extraction datasets represents a promising avenue, albeit one that remains largely untapped for low-resource languages.

The movement toward annotation assisted by large language models marks a fundamental break from conventional crowdsourcing approaches. Recent studies have demonstrated that frontier models can achieve accuracy levels that rival human crowd-workers [18]. In the specific context of RE, LLMs act as collaborative annotators, facilitating a “human-in-the-loop” workflow where the model handles initial semantic mapping while humans resolve high-ambiguity cases [19]. This approach effectively mitigates the “annotation tax” for Kazakh, leveraging the cross-lingual reasoning of models like GPT-4 to infer relationships even from limited data [20].

In this work, we present KRED (Kazakh Relation Extraction Dataset), a comprehensive dataset constructed through an iterative, LLM-assisted annotation framework built upon the scaffolding of the KazNERD corpus. We design a multi-stage pipeline that combines LLM-assisted candidate generation, structured prompting, and expert manual verification. Our contributions are threefold: (1) we introduce a linguistically grounded relation schema for Kazakh; (2) we propose a reproducible LLM-guided NER-to-RE conversion pipeline; and (3) we release KRED along with benchmark experiments to advance the state of Kazakh NLP.

By framing dataset construction as a data-centric and collaborative human-LLM process, this work aims to advance not only Kazakh NLP resources but also practical methodologies for building relation extraction datasets in other low-resource languages.

Materials and research methods.

The KRED dataset was constructed by transforming the KazNERD named entity recognition (NER) corpus into a relation extraction (RE) dataset through a multi-stage annotation pipeline. KazNERD contains 112,702 sentences extracted from Kazakh television news transcripts and 136,333 named entity annotations across 25 entity classes in IOB2 format [15]. While the corpus provides detailed entity annotations, it does not include explicit relations between entities.

To convert the NER annotations into relation annotations, we developed a pipeline that generates candidate entity pairs within each sentence and assigns relation labels using large language models. Initial relation predictions are produced using zero-shot prompting with GPT-4o-mini and subsequently refined through schema-driven normalization rules and targeted re-annotation using Gemini-3-flash. The choice of these lightweight LLMs was motivated by efficiency and scalability considerations. Since the annotation pipeline processes a large number of candidate entity pairs, using smaller models significantly reduces computational cost and latency while still providing sufficiently accurate semantic predictions. This trade-off is particularly important for low-resource settings, where large-scale annotation must remain cost-effective. The process concludes with a manual inspection of ambiguous cases to guarantee final semantic and structural integrity. The overall dataset construction workflow is illustrated in Figure 1.

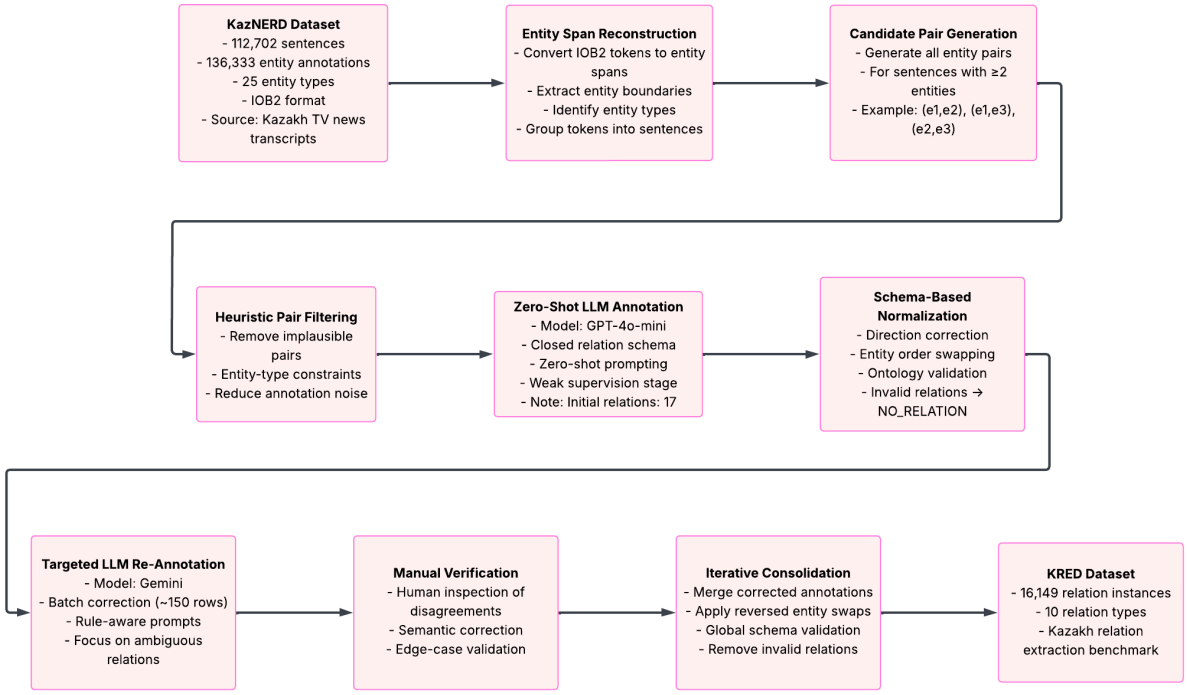

Figure 1 – Overview of the KRED dataset construction process, including entity span reconstruction, candidate pair generation, LLM-based relation annotation, schema normalization, and targeted re-annotation.

The following subsections describe the relation schema design, annotation guidelines, NER-to-RE conversion pipeline, normalization and re-annotation procedures, and dataset validation process.

Relation Schema Design

The relation schema defines the set of semantic relationships that can occur between entity pairs in the KRED dataset. The schema was designed to capture common relational patterns found in Kazakh news texts while remaining compact enough to support reliable automatic annotation.

The initial set of candidate relations consisted of 17 relation types, derived from common information extraction tasks and patterns observed in the KazNERD corpus. These relations

included POSITION_HELD, EMPLOYMENT, LOCATED_IN, PART_OF, PRODUCED_BY, PERSONAL_RELATION, SAME_GROUP, RELATED_TO, PARTICIPATED_IN, ORGANIZED_BY, FOUNDED_BY, CITIZEN_OF, BORN_IN, and FINANCED_BY, among others. During the initial annotation stage, relation labels were generated using zero-shot prompting with GPT-4o-mini.

After the first annotation pass, the resulting relation distribution was analysed. Several relations appeared extremely rarely or overlapped semantically with broader relation categories. For example, relations such as FOUNDED_BY and EMPLOYMENT occurred only a small number of times, while relations like CITIZEN_OF and BORN_IN were often semantically similar to other geographic or affiliation relations. To improve dataset consistency and avoid sparsity issues during model training, these relations were removed or merged with more general relation categories during schema refinement.

As a result, the final KRED dataset contains ten relation types, including the negative class NO_RELATION. The final schema includes the following relations: SAME_GROUP, POSITION_HELD, PART_OF, RELATED_TO, PERSONAL_RELATION, LOCATED_IN, PRODUCED_BY, PARTICIPATED_IN, ORGANIZED_BY, NO_RELATION.

These relations capture several types of entity interactions frequently observed in news text, including organizational affiliation, institutional roles, geographic containment, interpersonal relationships, and event participation. Examples of annotated relations from the dataset are shown in Figure 2.

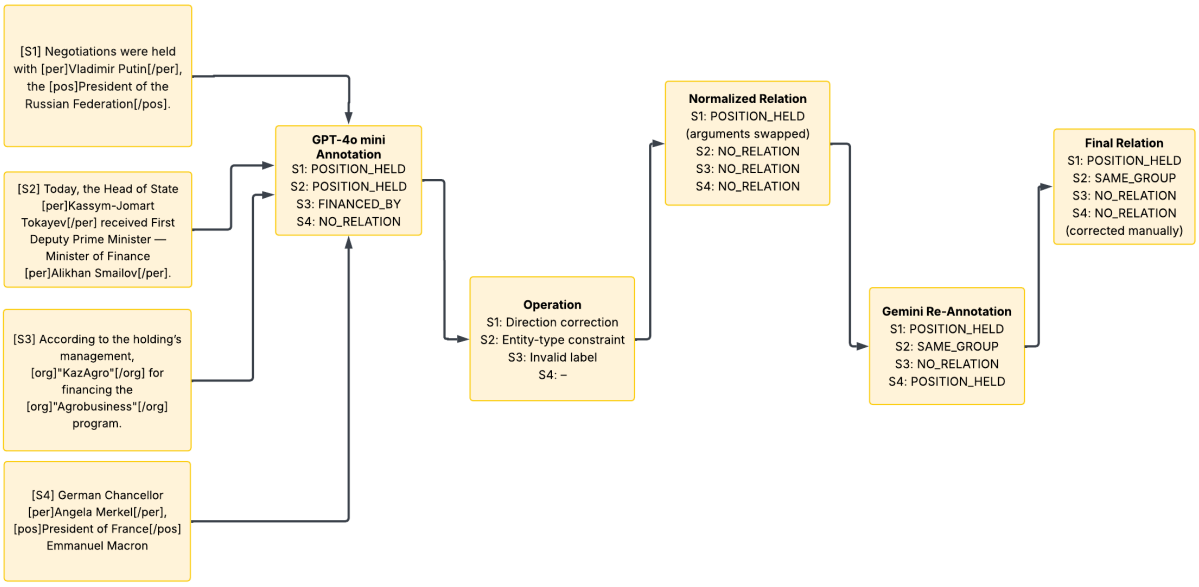

Figure 2 – Examples of relation annotations and schema-driven normalization and re-annotation in the KRED dataset.

Annotation Guidelines

To ensure consistent interpretation of relations, a set of annotation guidelines was established to define how relations should be assigned between entity pairs. These guidelines were incorporated into the prompts used during LLM-based annotation and served as reference rules during later validation stages.

In news text, one frequent cause of ambiguity stems from entities appearing together without any clear semantic link between them. It is common for a single sentence to include several named entities that, although present in the same surrounding context, lack any genuine relationship. When such situations arise, the corresponding entity pair receives the NO_RELATION label. This design choice stops the model from inferring relations based merely on how close entities are to one another or how often they happen to appear together.

For the SAME_GROUP relation, the team defined additional special constraints. This particular label comes into play when two entities fall under one shared conceptual category and show up together as part of a coordinated grouping – for instance, several organizations involved in a joint agreement or activity. That said, entities occupying different rungs of a hierarchy, like a city paired with a country, do not receive this label regardless of their appearance within the same phrase.

Turning to geographic relations, these operate according to a containment-based hierarchy. The interpretation of locations assumes nested spatial arrangements, meaning smaller administrative units sit inside larger geographical ones. To illustrate, a given city might lie within a region, which in turn lies within a country. Should any relations violate this hierarchical logic, corrections get applied later during the normalization step.

A further important distinction separates institutional roles from actual individuals. Certain entities point not to specific people but to positions or titles. Take the phrase “President of the Russian Federation” – this refers to a formal office. By contrast, “Vladimir Putin” names the particular person occupying that office. In cases like these, the proper relation label becomes POSITION_HELD rather than PERSONAL_RELATION.

Elsewhere, relations that capture who produced something or authored a work receive the PRODUCED_BY label. Relations describing someone's participation in events or activities take the PARTICIPATED_IN label. When the focus falls on who coordinated or organized an event, the relation ORGANIZED_BY applies.

Overall, these guidelines work to ensure that relation labels genuinely capture semantic connections rather than merely reflecting how often words co-occur in the text. By weaving these constraints into both the annotation instructions and the later validation process, the pipeline for building the dataset cuts down on annotation noise and boosts overall consistency.

NER-to-RE Conversion Pipeline

The conversion of the KazNERD corpus into a relation extraction dataset involved transforming token-level named entity annotations into structured relation instances through a multi-stage pipeline, including entity span reconstruction, candidate pair generation, heuristic filtering, and LLM-assisted annotation. Specifically, the original IOB2 annotations were first converted into complete entity spans by reconstructing continuous entity mentions along with their corresponding types and character positions within each sentence, providing the foundation for subsequent relation extraction steps.

Once entity spans were reconstructed, candidate relation pairs were generated. For each sentence containing two or more entities, all possible ordered pairs of entities were created. If a sentence contains n entities, this process produces $n(n-1)/2$ candidate pairs. This exhaustive pairing strategy ensures that all potential relations within a sentence are considered during annotation.

To reduce noise and computational cost, a set of heuristic filtering rules was applied before relation annotation. These rules removed entity pairs that were unlikely to form meaningful relations, such as pairs involving extremely distant entities within long sentences or pairs that violate obvious semantic constraints. The remaining candidate pairs were then prepared for LLM-assisted relation labelling.

Initial relation annotations were generated using GPT-4o-mini in a zero-shot prompting setting. Each candidate pair was presented to the model together with its sentence context, entity spans, and a list of possible relation labels defined by the schema. The model was instructed to assign exactly one relation label from the predefined schema or return NO_RELATION if no meaningful relation exists between the entities. The overall pipeline used to transform KazNERD into the KRED dataset is illustrated in Figure 1.

The output of this stage provides an initial set of relation predictions that serve as weakly supervised annotations. These predictions are further refined through schema normalization and targeted re-annotation, which are described in the following subsection.

Schema-Driven Normalization and LLM Re-Annotation

The initial relation predictions generated by the LLM may contain invalid labels, incorrect argument order, or relations that violate entity-type constraints. To ensure consistency with the predefined relation schema, all predictions were processed through a rule-based normalization stage.

The normalization procedure consists of three checks: schema validation, relation direction correction, and entity-type compatibility verification. First, predicted labels were validated against the predefined relation schema. Predictions outside the allowed relation set were reassigned to NO_RELATION.

Second, relations requiring a canonical argument order were normalized. For example, POSITION_HELD must connect a position entity with the person occupying that role. If the LLM produced the correct relation but reversed the arguments, the entity order was automatically swapped.

Third, entity-type compatibility constraints were applied. Certain relations are only valid between specific types of entities. For instance, LOCATED_IN typically connects geographic entities, while PRODUCED_BY links a product or media entity with an organization. Predictions violating these constraints were reassigned to NO_RELATION. Examples of how initial LLM predictions are normalized using schema constraints and subsequently refined through Gemini-3-flash based re-annotation and manual verification are shown in Figure 2. The overall normalization logic is illustrated in Figure 3.

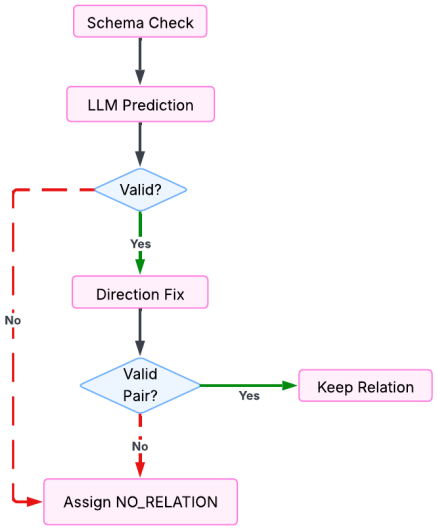

Figure 3 – Schema-Driven Normalization Workflow

After normalization, the dataset was re-evaluated using the Gemini-3-flash model to further refine relation annotations. All instances were processed in small batches using prompts that included explicit definitions of relation semantics to reduce ambiguity.

After the re-annotation phase concluded, additional quality control procedures were conducted to verify the dataset's structural integrity. These validation steps confirmed that every relation adhered to the established schema and that each entity pair met the necessary type compatibility requirements.

To further validate annotation quality, corrected instances were manually inspected. This inspection focused primarily on cases where Gemini-3-flash predictions differed from the initial GPT-4o-mini annotations or where relations involved potentially ambiguous entity pairs. Relations that remained inconsistent after normalization and re-annotation were also inspected and corrected manually.

Manual validation was performed by two annotators from Astana IT University with expertise in natural language processing and native proficiency in Kazakh. The annotators reviewed instances where LLM predictions disagreed, relations violated schema constraints, or semantic ambiguity was detected. Each instance was independently verified and corrected when necessary. To ensure consistency, annotators followed the predefined annotation guidelines described earlier. In cases of disagreement, a consensus-based resolution strategy was applied.

Finally, the corrected annotations were consolidated into the final KRED dataset. Each relation instance was associated with its corresponding sentence, entity spans, and normalized relation label. The resulting dataset provides a structurally consistent set of relation annotations suitable for training and evaluating relation extraction models.

Prompt Design for LLM-based Annotation

To ensure the reproducibility and internal consistency of the annotation pipeline, we developed a two-stage structured prompting strategy designed to guide LLMs toward schema-compliant outputs. In the initial phase, GPT-4o-mini was employed in a zero-shot setting to generate preliminary relation labels. The prompt was engineered to simulate the behavior of an expert linguist, incorporating the full sentence context, surface forms of the target entities, their respective types, and the predefined relation schema. To maintain data integrity, the model was restricted to outputting only the selected label (or NO_RELATION) without supplementary explanation. Crucially, the instructions required the LLM to infer the semantic link even when the relational direction appeared reversed in the text, thereby establishing a consistent "silver-standard" baseline.

Following initial annotation and normalization, a second stage of refinement was conducted using Gemini-3-flash. This stage functioned as a structured validation process, where the model was tasked with simultaneously verifying the assigned label and identifying directionality errors. This design is particularly vital for relations with strict hierarchical dependencies, such as POSITION_HELD (which requires a Position-Person order) and PART_OF. To reduce ambiguity, this re-annotation prompt incorporated detailed guidelines on entity-type compatibility and hierarchical logic, such as the distinction between SAME_GROUP and geographic PART_OF relations. Outputs were returned in a structured tabular format, allowing for systematic correction of errors introduced in the first stage.

Ultimately, this prompting architecture was governed by three core principles: schema alignment, which enforced a closed-set taxonomy; output constraint, which eliminated conversational verbosity for efficient post-processing; and progressive refinement, where a secondary model served as an automated consistency checker. To facilitate reproducibility and future research, the full prompt templates will be made available in the official KRED repository upon the publication of this work.

Experimental Setup

A total of 16,149 annotated relation instances are included in the KRED dataset, covering ten distinct relation categories. For experimental purposes, the data was split into training, development, and test subsets according to a 70/15/15 proportion. The relation extraction task was framed as a problem of multi-class classification. In this setup, the model must determine which semantic link connects a pair of entities occurring within the boundaries of a single sentence.

To ensure that the two target entities stand out clearly to the model, special delimiter tokens were inserted around each one in the input text. Specifically, the first entity gets wrapped in [E1] and [/E1], while the second receives [E2] and [/E2]. Through this marking approach, the model can direct its attention to the entity pair in question while still preserving the surrounding sentence context. For example: “Negotiations were held with [E2]Vladimir Putin[/E2], the [E1]President of the Russian Federation[/E1].”

All experiments were implemented using the HuggingFace Transformers framework. Models were fine-tuned using a learning rate of 2e-5, batch size 16, and five training epochs.

To establish baseline performance for Kazakh relation extraction, we evaluate three transformer architectures representing different pretraining strategies. First, we consider the multilingual variant of BERT, which was pretrained on more than one hundred languages and is commonly used as a baseline for multilingual NLP tasks. We further evaluate XLM-RoBERTa, a large multilingual model trained on large-scale CommonCrawl data covering a wide range of languages. Finally, experiments include Kaz-RoBERTa, a RoBERTa-based architecture pretrained specifically on Kazakh text.

Results and their discussion.

Dataset Statistics

This subsection provides an overview of the KRED dataset, including its overall size, relation distribution, and entity-type interaction patterns.

The final dataset contains 16,149 relation instances across 10 relation types, derived from entity pairs extracted from the KazNERD corpus. The dataset includes 9 entity types for both argument positions, allowing a wide range of entity interactions. An overview of the dataset statistics is presented in Table 1.

Table 1 – Summary statistics of the KRED dataset.

Metric	Value
Total relation instances	16,149
Relation types	10
Entity1 types	9
Entity2 types	9
Imbalance score	173.51

Table 2 presents how the various relations are distributed across the dataset. Among all categories, SAME_GROUP appears most often, making up a significant share of the annotated instances. This high frequency stems from the way news texts regularly include coordinated entity mentions – for example, lists of organizations or political figures all taking part in a shared event.

In addition, the dataset includes a considerable number of POSITION_HELD and PART_OF relations. These two categories capture institutional roles and nested geographic hierarchies, both of which are commonly encountered in news reporting.

Table 2 – Distribution of relations in the KRED dataset.

Relation	Count
SAME_GROUP	8,155
NO_RELATION	2,358
POSITION_HELD	2,255
PART_OF	1,127
RELATED_TO	829
PERSONAL_RELATION	824

LOCATED_IN	265
PRODUCED_BY	213
PARTICIPATED_IN	76
ORGANIZED_BY	47

A long-tail distribution characterizes the dataset, meaning that certain relations occur only sparingly. Specifically, both PARTICIPATED_IN and ORGANIZED_BY make up fewer than one percent of all annotated instances each. Despite their low frequency, these two relation types capture meaningful event-oriented interactions that appear regularly in news stories.

To better understand how entities interact within the dataset, we analyse the distribution of entity-type pairs. As presented in Figure 4, the most frequent interactions occur between GPE–GPE, PERSON–PERSON, and POSITION–PERSON entity pairs. The large number of GPE–GPE interactions reflects the prevalence of geographic containment relations such as PART_OF and LOCATED_IN. Similarly, PERSON–PERSON interactions commonly correspond to PERSONAL_RELATION or SAME_GROUP relations in political or social contexts.

Another prominent pattern is the POSITION–PERSON pair, which primarily corresponds to the POSITION_HELD relation, reflecting institutional role assignments frequently reported in political news.

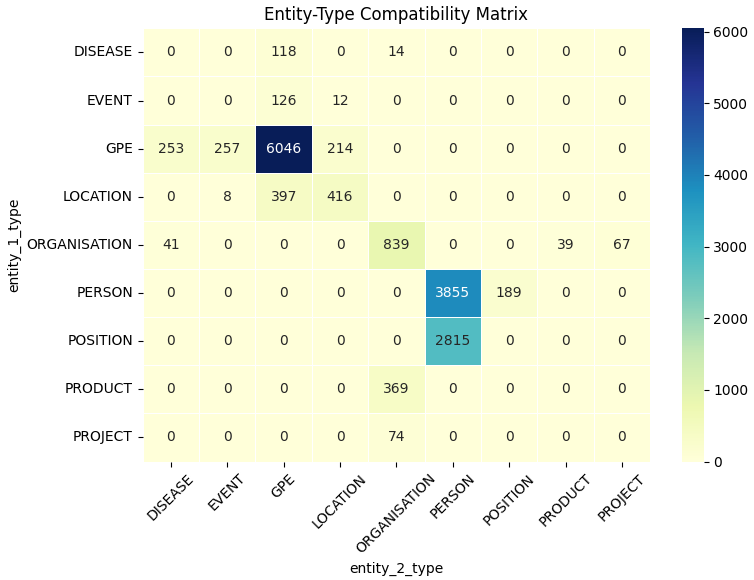

Figure – 4. Entity-type interaction matrix for relation instances in the KRED dataset.

Before relation annotation, candidate entity pairs were filtered using a whitelist of valid entity-type combinations to reduce annotation noise. This filtering stage significantly reduced the number of candidate pairs.

Initially, 93,908 entity pairs were generated from the KazNERD corpus. After applying entity-type constraints, the number of candidate pairs was reduced to 16,151, corresponding to an 82.8% reduction in potential pairs. This filtering stage helps eliminate semantically implausible entity combinations while preserving meaningful relation candidates.

During the final validation stage, the dataset was checked for schema violations and structurally inconsistent relations. After Gemini-3-flash re-annotation and manual inspection, 72 instances were identified as potentially invalid according to the relation schema. The majority of these cases occurred within the ORGANIZED_BY relation. These cases were again manually reviewed, and corrections were applied when necessary.

The statistics presented in this section demonstrate that KRED contains a diverse set of entity interactions and relation types derived from real-world news text. Despite the inherent class imbalance typical for relation extraction datasets, the dataset provides a broad range of semantic relations suitable for evaluating modern relation extraction models.

Benchmark Results

In this section, we establish baseline performance metrics for Kazakh relation extraction using several transformer-based architectures. Our experiments evaluate the capacity of both multilingual and Kazakh-specific pretrained language models to capture relational patterns within a low-resource linguistic framework. Performance was measured using precision, recall, and F1 scores. We report both micro-averaged F1, which provides an overall accuracy across all instances, and macro-averaged F1, which offers a more stringent assessment of model performance across imbalanced classes. The experimental results on the KRED test set are summarized in Table 3.

Table 3 – Performance of the evaluated models on the KRED test set.

Model	Micro-F1	Macro-F1
mBERT	0.8832	0.8113
XLM-RoBERTa	0.8523	0.5921
Kaz-RoBERTa	0.8494	0.6970

The evaluation reveals that mBERT delivers the most robust performance, significantly outperforming its counterparts in both metrics. While XLM-RoBERTa is often regarded as a superior architecture for massive multilingual benchmarks, our results show that mBERT exhibits better adaptability to the specific relational demands of the KRED dataset. Most notably, mBERT's high micro-F1 and macro-F1 score (0.8832 and 0.8113, respectivly) suggests a superior ability to generalize across minority classes, whereas the sharp drop in macro-F1 for XLM-RoBERTa (0.5921) indicates a tendency to favor dominant relations at the expense of less frequent ones.

Interestingly, Kaz-RoBERTa demonstrated competitive performance, particularly in terms of its macro-F1 score (0.6970), which noticeably surpassed that of the larger XLM-RoBERTa. This parity suggests that language-specific pretraining – even on a more modest corpus – can produce contextual representations that are better tuned to the nuances of Kazakh than those derived from general-purpose multilingual models. Furthermore, the high micro-F1 scores across all three architectures indicate that KRED contains clear and consistent relational signals that modern transformer models can successfully decode.

Overall, these findings confirm that modern pretrained frameworks are well-equipped to decode relational patterns in Kazakh text. By establishing these strong initial baselines, this study provides a foundation for more advanced information extraction research in low-resource linguistic environments.

Error Analysis

To gain deeper insights into model behavior, we examined the per-class performance metrics (Table 4) and the confusion matrix for the top-performing mBERT model (Figure 5). The analysis of the results shows a clear connection between the per-class performance in Table 4 and the specific errors shown in the confusion matrix in Figure 5. The mBERT model performs exceptionally well on relations that have rigid and predictable structures. For example, POSITION_HELD and PRODUCED_BY achieved the highest scores, both exceeding 0.97 F1. The confusion matrix confirms this precision, showing that professional roles were correctly identified in 334 cases with almost no interference from other labels. Similarly, the model handled SAME_GROUP very effectively, correctly labeling 1,164 instances. This high accuracy is likely

due to the frequent use of list-like structures in Kazakh news, where multiple people or organizations sharing the same role are easily recognized by the model.

In contrast, the model struggles when the semantic boundaries between relations are less clear. The lower F1-scores for RELATED_TO (0.53) and PERSONAL_RELATION (0.55) reflect this difficulty. According to the confusion matrix, 45 personal links were mistakenly categorized as group memberships, and vague semantic ties were often confused with no relation at all. This suggests that the model often defaults to a broader or simpler category when the specific type of connection is subtle or ambiguous.

Table 4 – Per-class performance metrics for the best-performing model (mBERT).

Relation	F1-score
SAME_GROUP	0.9290
NO_RELATION	0.8355
POSITION_HELD	0.9766
PART_OF	0.8968
RELATED_TO	0.5316
PERSONAL_RELATION	0.5572
LOCATED_IN	0.8471
PRODUCED_BY	0.9841
PARTICIPATED_IN	0.6316
ORGANIZED_BY	0.9231

A notable source of confusion also appears between PART_OF and LOCATED_IN, despite both categories performing reasonably well with scores above 0.84. In many Kazakh sentences, the distinction between a geographic location and an administrative belonging is often subtle, which leads the model to occasionally mix up the two labels. When it comes to event-based relations like PARTICIPATED_IN, the weaker performance can be traced back primarily to data sparsity. These relation types appear far less frequently in the dataset, which naturally makes it harder for the model to acquire stable patterns than it does for more common categories. Overall, the model handles clear and explicit relational patterns with a high degree of reliability. But

improvement is still needed – especially for relation classes that either share overlapping meanings or show up only rarely during training.

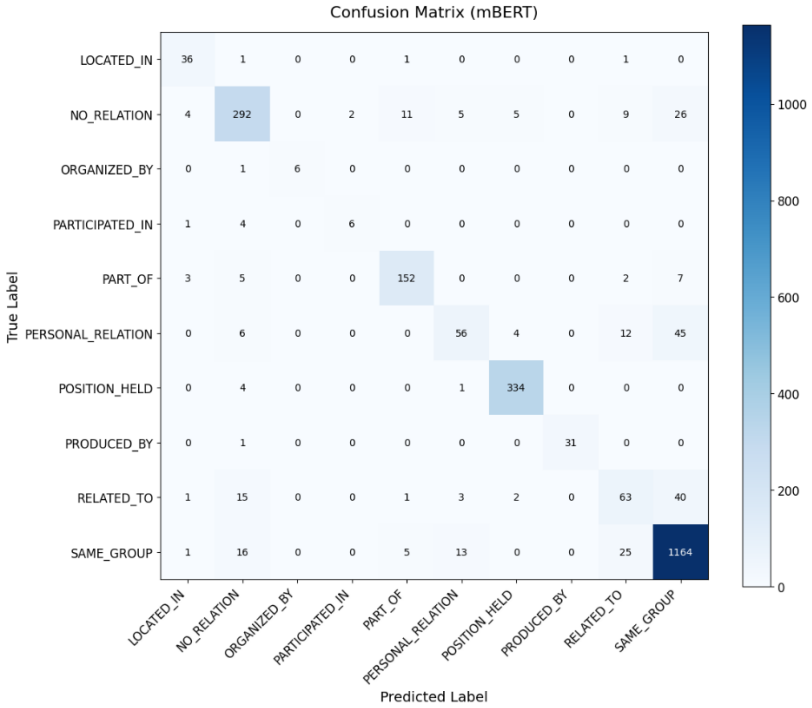

Figure 5 – Confusion matrix of the best-performing model (mBERT).

Discussion

The construction of KRED highlights both the opportunities and challenges of developing relation extraction resources for low-resource languages. When placed alongside widely adopted benchmarks like DocRED and TACRED, KRED stands apart in a few keyways. One major difference is its focus on Kazakh, which fills an important gap in the multilingual NLP space. Another distinguishing feature lies in the annotation approach. Whereas conventional datasets depend heavily on labor-intensive manual curation, KRED introduces a scalable pipeline assisted by LLMs. This design brings down annotation costs considerably while still maintaining high data quality. Finally, KRED captures relational patterns unique to Kazakh news discourse – such as morphologically rich structures and flexible word order – which are often underrepresented in existing benchmarks. These distinctions justify the necessity of a dedicated resource, as global datasets cannot be directly transferred to the Kazakh context without a substantial loss in performance.

Our findings also emphasize the practical effectiveness of schema-driven normalization and targeted re-annotation. Initial LLM predictions often contained inconsistencies, including reversed argument orders and invalid labels. However, these issues were significantly mitigated through rule-based corrections and selective re-evaluation using a secondary model. This confirms that combining raw LLM outputs with structured validation layers is essential for achieving high annotation precision.

More broadly, KRED demonstrates that pretrained models can achieve strong overall performance, even though subtle contextual cues – such as the boundary between RELATED_TO and SAME_GROUP – remain difficult to capture. This work demonstrates that hybrid annotation approaches, combining the speed of AI with controlled human oversight, provide a viable and scalable alternative to fully manual dataset construction. Such frameworks are particularly valuable for expanding the frontier of information extraction in low-resource and Turkic languages.

Conclusion.

This study introduces KRED, a new relation extraction dataset for Kazakh built using an innovative LLM-assisted annotation pipeline. The dataset contains 16,149 annotated instances distributed across ten relation types and covers a wide range of entity categories, including people, organizations, places, and job titles. To ensure structural consistency, the development process relied on a schema-guided framework, complemented by automated normalization steps and targeted LLM-assisted re-annotation aimed at resolving unclear semantic cases.

Baseline experiments confirm that KRED functions as both a learnable and suitably challenging benchmark for information extraction. Among the architectures tested, multilingual BERT proved most effective, reaching a top micro-F1 score of 0.8832. Although this level of performance is strong, a closer look at the errors reveals ongoing difficulties with infrequent relation types and semantically overlapping categories. These shortcomings point toward clear directions for future model improvements.

In summary, KRED stands as one of the first publicly available benchmarks for Kazakh relation extraction. It also offers a reproducible blueprint for building linguistic resources in low-resource language settings. Looking ahead, we plan to expand the dataset with broader domain coverage and to explore cross-lingual transfer learning scenarios. Both KRED and its associated benchmarks will be released publicly to support further research in Turkic information extraction.

The KRED dataset will be made publicly available to the research community upon publication of this work. It will be released through an open-access repository, along with annotation guidelines and preprocessing scripts to ensure reproducibility.

References

1. Zhao, X., Deng, Y., Yang, M., Wang, L., Zhang, R., Cheng, H. & Xu, R. (2024). A comprehensive survey on relation extraction: Recent advances and new frontiers. ACM Computing Surveys, 56(11), 1-39. https://doi.org/10.1145/3674501

2. Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z. & Sun, M. (2019, July). DocRED: A large-scale document-level relation extraction dataset. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 764-777).

3. Alt, C., Gabryszak, A. & Hennig, L. (2020, July). TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 1558-1569).

4. Diaz-Garcia, J. A. & Lopez, J. A. D. (2025). A survey on cutting-edge relation extraction techniques based on language models. Artificial Intelligence Review, 58(9), 287. https://doi.org/10.1007/s10462-025-11280-0

5. Pakray, P., Gelbukh, A. & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. Natural Language Processing, 31(2), 183-197. doi:10.1017/nlp.2024.33

6. Gharagozlou, H., Mohammadzadeh, J., Bastanfard, A., & Ghidary S. S. (2023). Semantic relation extraction: a review of approaches, datasets, and evaluation methods with looking at the methods and datasets in the Persian language. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(7), 1-29. https://dl.acm.org/doi/10.1145/3592601

7. Wang, Y., Lin, M., Hu, Q., Bao, L., Bai, S. & Li, Y. (2025). Large and Small models for collaborative cross-lingual data augmentation in entity relationship extraction for low-resource languages. Journal of King Saud University Computer and Information Sciences, 37(4), 56. https://doi.org/10.1007/s44443-025-00055-w

8. Yessenbayev, Z., Kozhirbayev, Z. & Makazhanov, A. (2020, September). KazNLP: A pipeline for automated processing of texts written in Kazakh language. In International Conference on Speech and Computer (pp. 657-666). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-60276-5_63

9. Tukeyev, U., Turganbayeva, A., Abduali, B., Rakhimova, D., Amirova, D. & Karibayeva, A. (2018, October). Lexicon-free stemming for Kazakh language information retrieval. In 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE. doi: 10.1109/ICAICT.2018.8747021

10. Abibullayeva, A. & Çetin, A. (2022). Keyword extraction from kazakh news dataset with bert. El-Cezeri, 9(4), 1193-1200. doi: 10.31202/ecjse.1131826

11. Diana, R. & Assem, S. (2019, August). Problems of semantics of words of the Kazakh language in the information retrieval. In International Conference on Computational Collective Intelligence (pp. 70-81). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28374-2_7

12. Kamshat, A., Auyeskhan, U., Zarina, N., Alen, S. & Yeskazina, M. (2024, April). Integration AI techniques in Low-Resource Language: the case of Kazakh language. In 2024 IEEE AITU: Digital Generation (pp. 7-13). IEEE. doi: 10.1109/IEEECONF61558.2024.10585350

13. Kadyrbek, N., Tuimebayev, Z., Mansurova, M. & Viegas, V. (2025). The development of small-scale language models for low-resource languages, with a focus on kazakh and direct preference optimization. Big Data and Cognitive Computing, 9(5), 137. https://doi.org/10.3390/bdcc9050137

14. Batura, T., Yerimbetova, A., Mukazhanov, N., Shvarts, N., Sakenov, B. & Turdalyuly, M. (2025). Information Extraction from Multi-Domain Scientific Documents: Methods and Insights. Applied Sciences, 15(16), 9086. https://doi.org/10.3390/app15169086

15. Yeshpanov, R., Khassanov, Y. & Varol, H. A. (2022, June). KazNERD: Kazakh named entity recognition dataset. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 417-426).

16. Hayinaer, A., Ye, Z. & Xu, L. (2025, July). A Study of Low-Resource Kazakh Named Entity Recognition. In International Conference on Computational Linguistics and Natural Language Processing (pp. 13-27). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-95-4788-3_2

17. Khairova, N., Petrasova, S., Mamyrbayev, O. & Mukhsina, K. (2020, March). Open Information Extraction as Additional Source for Kazakh Ontology Generation. In Asian Conference on Intelligent Information and Database Systems (pp. 86-96). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-41964-6_8

18. Gilardi, F., Alizadeh, M. & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

19. He, X., Lin, Z., Gong, Y., Jin, A. L., Zhang, H., Lin, C. & Chen, W. (2024, June). Annollm: Making large language models to be better crowdsourced annotators. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track) (pp. 165-190).

20. Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P. & Sitaram, S. (2023, December). Mega: Multilingual evaluation of generative ai. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 4232-4267).

АТАУЛЫ МӘНДЕРДІ ТАНУДАН ҚАТЫНАСТАРДЫ АНЫҚТАУҒА ДЕЙІН: ҮЛКЕН ТІЛДІК МОДЕЛЬДЕР КӨМЕГІМЕН ҚАЗАҚ ТІЛІНДЕГІ ҚАТЫНАСТАРДЫ АНЫҚТАУ ДЕРЕКТЕР ЖИЫНЫН ҚҰРАСТЫРУ

Аңдатпа. *Қатынастарды анықтау – құрылымданбаған мәтіннен формальды білім құрылымдарын қалыптастырудағы негізгі кезеңдердің бірі. Қазақ тілі үшін бұл бағыттың дамуы сапалы аннотацияланған семантикалық ресурстардың жетіспеушілігімен шектеліп келді. KazNERD деректер жиыны атаулы мәндерді тану міндеті үшін берік негіз қалыптастырғанымен, мәндер арасындағы күрделі семантикалық байланыстарды*

модельдеу әлі де өзекті мәселе болып отыр. Осы зерттеуде аталған мәселені шешу мақсатында KRED (Kazakh Relation Extraction Dataset) деректер жиыны ұсынылады. Бұл деректер жиыны үлкен тілдік модельдер мен сараптамалық тексеруді біріктіретін көпкезеңді және ауқымды аннотациялау үдерісі арқылы құрылған. Аннотациялау барысында KazNERD корпусындағы тексерілген мән шекаралары негіз ретінде алынып, мән жұптарын генерациялау, GPT-4o-mini моделі арқылы zero-shot белгілеу және семантикалық нақтыландыру кезеңдері жүзеге асырылды. Құрылымдық бірізділікті қамтамасыз ету үшін схемаға негізделген нормализация қолданылып, кейін Gemini-3-flash моделі арқылы қайта аннотациялау және қолмен тексеру жүргізілді. Нәтижесінде 10 түрлі қатынас түрін қамтитын 16 149 аннотацияланған байланыстан тұратын деректер жиыны алынды. Жүргізілген эксперименттер mBERT, XLM-RoBERTa және Kaz-RoBERTa модельдерін қолдана отырып бағаланды. Ең жоғары нәтижені mBERT көрсетіп, micro-F1 метрикасы бойынша 0.8832 және macro-F1 метрикасы бойынша 0.8113 мәніне жетті. Ұсынылған тәсіл толықтай қолмен аннотациялауға балама ретінде тиімді әрі үнемді шешім ұсынады және ресурсы шектеулі тілдер, соның ішінде түркі тілдері үшін ақпаратты автоматты түрде өңдеу ресурстарын кеңейтуге мүмкіндік береді.

Түйін сөздер: *табиғи тілдерді өңдеу, ақпаратты іздеу, ресурсы шектеулі тілдер, үлкен тілдік модельдер, деректер жиынтығын құрастыру, қатынастарды шығару.*

ОТ РАСПОЗНАВАНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ К ИЗВЛЕЧЕНИЮ ОТНОШЕНИЙ: ПОСТРОЕНИЕ НАБОРА ДАННЫХ ДЛЯ КАЗАХСКОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Абстракт. *Извлечение отношений является ключевым этапом преобразования неструктурированного текста в формализованные представления знаний. Для казахского языка развитие данного направления существенно ограничено недостатком качественных размеченных семантических ресурсов. Несмотря на то, что корпус KazNERD сформировал надежную основу для задачи распознавания именованных сущностей, моделирование сложных связей между ними по-прежнему остается актуальной проблемой. В данной работе предлагается новый набор данных KRED (Kazakh Relation Extraction Dataset), направленный на решение этой задачи. Датасет создан с использованием масштабируемого пайплайна аннотирования, объединяющего возможности больших языковых моделей (LLM) и экспертной валидации. В качестве основы использовались размеченные границы сущностей из корпуса KazNERD, после чего выполнялись генерация пар сущностей, zero-shot аннотирование с применением GPT-4o-mini и последовательное уточнение семантики. Для обеспечения согласованности применялась схема-ориентированная нормализация, дополненная повторной аннотацией с использованием модели Gemini-3-flash и ручной проверкой спорных случаев. В результате был сформирован набор данных, включающий 16 149 примеров отношений, распределенных по 10 семантическим категориям. Экспериментальная оценка с использованием моделей multilingual BERT, XLM-RoBERTa и Kaz-RoBERTa показала высокую пригодность датасета для задач извлечения отношений. Наилучший результат продемонстрировала модель mBERT с показателем micro-F1 = 0.8832 и macro-F1 = 0.8113. Предложенный гибридный подход представляет собой эффективную и экономичную альтернативу полностью ручной разметке и может служить основой для расширения ресурсов информационного извлечения в языках с ограниченными ресурсами, включая тюркские языки.*

Ключевые слова: *обработка естественного языка, информационный поиск, языки с ограниченными ресурсами, большие языковые модели, построение набора данных, извлечение отношений.*

Сведение об авторах

Айдынкызы Айдана	Магистр компьютерной инженерии, преподаватель Astana IT University, Астана, Казахстан, E-mail: aidana.aidynkyzy@astanait.edu.kz

Авторлар туралы мәлімет

Айдынқызы Айдана	Компьютерлік инженерия магистрі, Astana IT University оқытушысы, Астана қ., Қазақстан, E-mail: aidana.aidynkyzy@astanait.edu.kz

Information about the authors

Aidynkyzy Aidana	MSs in Computer Engineering, Teacher at Astana IT University, Astana, Kazakhstan E-mail: aidana.aidynkyzy@astanait.edu.kz