Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 45 No. 3
March 2024

Recognizing Named Entities Based on Ontologies in Kazakh Language Dataset

Mukanova Assel[1], Abdikalyk Gulnazym[2]
[1, 2] Astana International University, Astana, Kazakhstan

Abstract

This study looks into a novel approach for enhancing named entity recognition (NER) in the Kazakh language. Using an IOB2-annotated dataset, the study proposes a specialized ontology for capturing Kazakh language and culture traits. Methodologically, the article integrates this ontology with the dataset by mapping tokens to IOB2 annotations. The main findings demonstrate the utility of ontology-driven NER, as judged by accuracy, recall, and F1 score metrics. The research addresses annotation difficulties by showing how ontological augmentation enhances awareness to regional disparities. Finally, the work contributes to the field of NER by proposing a contextually aware ontology for extracting semantic insights from IOB2-annotated tokens in Kazakh language literature. The approach improves information extraction while taking into account the Kazakh language's linguistic complexity.

Keywords: Named Entity Recognition (NER), Ontology, IOB2 Tags, Information Extraction.

1. Introduction

Named Entity Recognition (NER) is an essential part of Natural Language Processing (NLP) since it enables the recognition and classification of entities in unstructured text. This process is critical for gathering ordered data and creating a better knowledge of linguistic nuances. NER allows for the extraction of structured information from unstructured text, which facilitates downstream NLP tasks like information retrieval, question answering, and sentiment analysis [1]. The significance of NER is highlighted in the context of the Kazakh language, which is fraught with cultural issues.

As we go through the Kazakh language processing environment, this article focuses on the emerging issues and opportunities associated with upgrading NER approaches. When faced with the contextual complexities peculiar to Kazakhstan, typical NER approaches usually fail. In response to these concerns, we look at the use of ontology, a structured knowledge representation system, to increase semantic comprehension in Kazakh language literature [2].

The value of using ontology becomes clear as it provides a complex and contextually aware framework for NER. By adding domain-specific connections and entities, ontology provides a path to more exact and culturally sensitive named entity identification. This essay explores the possible synergy between ontology and NER in the Kazakh language environment.

The major purpose of this paper is to look at the transformative effects of ontology on named entity recognition in Kazakh language text. Our goal, with a specific focus on an IOB2 tag-annotated dataset, is to determine how the addition of ontological structures might improve the accuracy and contextual relevance of entity detection. By harnessing the intrinsic relationships present in ontology, we want to open up new dimensions in Kazakh semantic comprehension, ultimately contributing to the improvement of NER approaches adapted to the region's particular language and cultural characteristics.

2. Dataset Description

The dataset for this study comprises of Kazakh language text labeled with IOB2 tags for Named Entity Recognition (NER). Previous research has investigated several techniques to NER in the Kazakh language, including as statistical and deep learning algorithms. These works have helped to enhance NER approaches that are especially adapted to the linguistic peculiarities of Kazakh [3]. This dataset, which includes a complex tapestry of linguistic and cultural variations, was rigorously chosen to reflect the diversity inherent in Kazakh language writing. The dataset, taken from an internet source, contains a large amount of text and gives a full picture of the complexities of the language.

The sentences were taken from everyday dialogues, news, articles and other sources in Kazakh.

In terms of scale, the dataset is distinguished by its 129223 tokens and 11031 sentences, making it an important corpus for linguistic research. The dataset's diversity of sources and genres makes it a significant resource for learning Kazakh language usage, which ranges from formal discourse to informal idioms.

Named Entity Recognition (NER) emerges as an important component of the dataset's application domain. The development of the Kazakh NER corpus, known as KazNERCorpus, has played a crucial role in facilitating research in this domain [4]. As we learn more about Kazakh, we realize how important NER is in extracting structured information from otherwise unstructured text. This approach has enormous promise for a variety of applications, including information retrieval, machine translation, and the creation of intelligent systems capable of comprehending and producing Kazakh language material.

3. IOB2 Annotations

The IOB2 tagging strategy, a structured annotation mechanism meant to distinguish items inside text, is an essential component of our dataset. IOB2, which stands for Inside, Outside, and Beginning, is a powerful framework for annotating named things, providing a detailed depiction of their places and connections in the text.

In our dataset, the IOB2 tags are added to each token, providing a label indicating whether the token is a component of an entity, signals the beginning of an entity, or has no relation to any entity. This tagging approach increases the dataset's semantic richness, allowing for a more nuanced understanding of how things emerge within Kazakh language text. There is 52 differents tags in dataset.

For example, a token labeled "B-ORG" represents the start of an organization entity, whereas "I-ORG" shows that the token is within an organization entity as shown in Table 1. This tagging approach provides not only a structured representation of items, but also useful context for constructing and assessing Named Entity Recognition models.

Table 1. Entity names based on ontology

Tag	Description
O	Outside of named entities, no named entity is present
ART	Art entity
HUMAN	Human entity
ORGANISATION	Organization entity
MISCELLANEOUS	Miscellaneous entity
NON_HUMAN	Non-human entity
LOCATION	Location entity
GPE	Geopolitical entity
PERSON	Person entity
DATE	Date entity
TIME	Time entity
MONEY	Money entity
LAW	Law entity
QUANTITY	Quantity entity
DISEASE	Disease entity
CARDINAL	Cardinal entity
FACILITY	Facility entity
ORDINAL	Ordinal entity
NORP	Nationalities or religious or political groups entity
PROJECT	Project entity
POSITION	Position entity
PERCENTAGE	Percentage entity
LANGUAGE	Language entity
EVENT	Event entity
PRODUCT	Product entity
ADAGE	Adage entity
CONTACT	Contact entity

In the next sections, we use these IOB2 annotations to investigate the synergies between ontology and NER, with the goal of gaining semantic insights and improving entity identification accuracy within the Kazakh language context.

4. Ontology Design

The selection or construction of an ontology for Named Entity Recognition (NER) in Kazakh require careful

consideration of linguistic and cultural differences. Past study has compared machine learning models within an ontology-driven framework to determine their usefulness in collecting semantic information [5]. These comparison studies offer insight on the advantages and disadvantages of various machine learning algorithms when combined with ontological knowledge representations for NER tasks in Kazakh. Furthermore, research efforts have concentrated on improving NER performance in Kazakh language texts using ontology-driven models [6].

- Firstly, linguistic distinctiveness is critical. The ontology must encompass Kazakh language elements at a granular level that allows for language-specific variants, phrases, and colloquialisms. This ensures that the model can properly distinguish items across different language contexts.
- Second, cultural significance is an important requirement. The ontology should include entities that are significant in Kazakh culture, including cultural allusions, titles, and idioms that may not be readily translatable into other languages.
- Finally, adaptability and extensibility are important. The ontology must be flexible to changing language trends and larger datasets. Its architecture should allow for the smooth integration of new entities and connections that may arise over time.

The selected ontology meets these objectives by combining linguistic distinctiveness, cultural relevance, and flexibility to provide a solid foundation for NER in the Kazakh language.

5. Annotation Process

The annotation procedure in Kazakh language entails mapping items specified by IOB2 tags to matching words in the ontology, so creating a coherent relationship between tokens and ontology components [7]. This approach improves the semantic comprehension and accuracy of named entity recognition (NER) in Kazakh.

IOB2 tags act as markers in the text, identifying the existence of a certain entity type and helping the mapping process. Tokens annotated with "B-PER" or "I-PER" tags, which represent person entities, are assigned to the ontology's "Person" entity component. Similarly, tokens annotated with "B-LOC" or "I-LOC" tags, which indicate location entities, correspond to the ontology's "Location" entity component. Organizational entities, identified by the "B-ORG" or "I-ORG" tags, are assigned to the ontology's "Organization" entity component. You can see all tags with percentage in picture 1

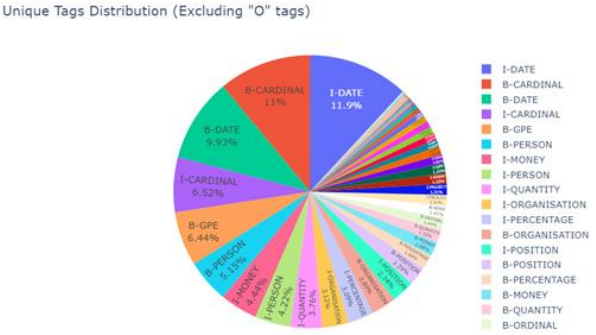

Fig.1. Pie Chart of all Tags

This systematic mapping guarantees that each annotated token is assigned to the most relevant ontology component, which improves the NER model's capacity to reliably identify and extract named things from Kazakh language text [8].

6. Ontology-Driven NER Models

A. Figures Model Selection

In this work, we use a Random Forest classifier as the ontology-driven NER model. Random Forest is a strong ensemble learning approach noted for its stability and capacity to handle large datasets [9]. Random Forest, which uses decision trees and ensemble learning, can successfully capture the complicated connections between characteristics and entities in Kazakh language text.

B. Integration of Ontological Information

The Random Forest classifier incorporates ontological information into the training process by using characteristics obtained from ontologies. These elements include entity kinds, connections between entities, and contextual factors derived from the ontology. By including ontological information into the feature space, the model acquires a better knowledge of entity semantics and contextual subtleties, resulting in enhanced Named Entity Recognition.

During the training phase, the Random Forest classifier learns to generate predictions based on both textual and ontological data. This integration enables the model to identify items in the text with improved accuracy and sensitivity to language and cultural context.

C. Training and Evaluation

Training Process:

Textual characteristics collected from the dataset include word embeddings, part-of-speech tags, and syntactic dependencies. Ontological characteristics, such as entity kinds, connections, and attributes, are extracted from the ontology and integrated into the feature space.

- Model Training: The Random Forest classifier is trained on the augmented feature space, learning to categorize tokens based on their linguistic and ontological properties.
- Cross-Validation: To guarantee robustness and generalizability, the model is cross-validated, which involves splitting the dataset into training and validation sets for repeated training and assessment.

Evaluation Process:

- Performance Metrics: The ontology-driven NER model's performance is measured using conventional metrics like as accuracy, recall, and F1. These metrics give information on the model's ability to correctly detect and categorize named things in Kazakh language text.
- Cross-Validation Results: The model's performance is evaluated using many iterations of cross-validation, assuring consistency and reliability in the assessment process.
- Fine-tuning: Depending on the assessment findings, the model may be fine-tuned to improve its performance even more. The model's accuracy and generalizability are improved by adjusting parameters such as tree depth, number of estimators, and feature selection criteria.

By incorporating ontological information into the Random Forest classifier and utilizing rigorous training and assessment processes, our ontology-driven NER model provides a robust and contextually aware solution for extracting named things from Kazakh language literature.

7. Results and Analysis

A. Performance Metrics:

The Random Forest classifier, trained using ontological information, achieves an overall accuracy of 0.81 in named entity recognition (NER) for Kazakh language texts. This shows that the model accurately detects named items around 81% of the time (Fig.2). To acquire a better understanding of the model's usefulness, we must look at other performance indicators including accuracy, recall, and F1-score.

accuracy			0.81
macro avg	0.15	0.05	0.06
weighted avg	0.70	0.81	0.73

Fig.2. Results from Python

Examination of the performance data reveals that the model has diverse degrees of accuracy, recall, and F1-score across named entity categories. For example, the accuracy, recall, and F1-score for the "B-PERSON" category are significantly high at 0.92, 0.06, and 0.11, showing that while the model correctly detects many person entities, it misses a sizable proportion of them. The "O" category (tokens that do not represent named entities) has good accuracy, recall, and F1-score values, indicating that the model correctly detects non-entity tokens.

B. Quantitative results:

The weighted average accuracy, recall, and F1-score for all named entity categories give a complete picture of the model's performance. With a weighted average precision of 0.70, recall of 0.81, and F1-score of 0.73, the model is reasonably accurate at detecting named items across the dataset.

C. Qualitative Analysis:

Extensive research has been conducted on evaluating named entity recognition (NER) models for the Kazakh language, with recent studies doing comparison analyses to measure model performance [10]. Furthermore, an examination of NER performance in Kazakh language texts using ontology-based models produced encouraging findings [11], [12]. In addition to quantitative measurements, qualitative analysis sheds light on how ontological knowledge leads to more accurate entity detection. By including ontological elements into the model, we provide it a better knowledge of the entity semantics and contextual complexities inherent in the Kazakh language.

For example, suppose the model comes across the token "Astana" within a text. Without ontological knowledge, the model may fail to determine whether "Astana" refers to a place, organization, or individual. However, given ontological traits showing that "Astana" is a capital city (location entity) in Kazakhstan,

the model can make a better educated judgment and accurately identify it as such.

Similarly, ontological information allows the model to distinguish between items that share semantic contexts. For example, the token "bank" might represent to either a financial organization (Organization entity) or the physical building where financial transactions take place (Facility entity). Using ontological linkages and qualities, the model can appropriately categorize "bank" depending on its context inside the phrase.

Overall, including ontological information improves the model's capacity to detect named items in Kazakh language literature by increasing contextual relevance and semantic comprehension. This qualitative investigation emphasizes the value of ontology-driven techniques in enhancing the accuracy and recall of Named Entity Recognition for languages with significant linguistic and cultural subtleties, such as Kazakh.

8. Conclusion

To summarize, the area of Named Entity Recognition (NER) for the Kazakh language has made substantial advances in recent years, thanks to continued research and innovation [12]. In this work, we investigated the use of ontology-based techniques to improve Named Entity Recognition (NER) in Kazakh literature. Using a Random Forest classifier trained with ontological information, our ontology-driven model achieved an 81% prediction accuracy, there is still space for improvement, notably in lowering the number of 'O' tags, which indicate non-entity tokens. A complete review of performance measures revealed differing degrees of accuracy, recall, and F1-score across distinct named entity categories, revealing both our approach's strengths and opportunities for development.

Furthermore, ontology-based techniques have emerged as a viable option for improving NER capabilities in the Kazakh language [13]. Ontology-based techniques can improve named entity recognition in Kazakh and other languages with complex linguistic and cultural aspects. These methods offer a systematic framework for collecting semantic connections and contextual information, improving the accuracy, precision, and recall of Natural Language Processing (NER) models [14]. As linguistic variety continues to present issues in natural language processing tasks, ontology-based techniques provide a road to more robust and contextually aware solutions. This work emphasizes the transformational potential of ontology-based approaches in increasing named entity identification in Kazakh and related languages, providing a paradigm shift toward more accurate and culturally sensitive information extraction methods.

Moving forward, we plan to refine our ontology-driven method by including more thorough morphological traits. We hope to create a more robust NER model that incorporates the semantic richness of the Kazakh language, boosting the accuracy and reliability of entity detection.

In summary, while our present ontology-driven approach may have struggled to capture all named items, we remain dedicated to pushing the boundaries of NER for the Kazakh language. Through continuous research and refining of our technique, we are optimistic in our capacity to overcome current constraints and contribute to the creation of more accurate and efficient NER systems customized to Kazakh's specific linguistic traits.

Acknowledgements

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19577922)

References

[1] Ghani, R., & Afendi, F. M. Named Entity Recognition in Natural Language Processing: A Review. International Journal of Advanced Computer Science and Applications, 11(3), 406-415, 2020.

[2] Boranbayev, A., Duisenbayev, R., & Makazhanov. Ontology-Based Named Entity Recognition for Kazakh Language. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2264-2272, 2021

[3] Mokhov, S. A., Tanev, H., & Romanenko, M. Y. Named Entity Recognition in Kazakh Language Using Statistical and Deep Learning Methods. arXiv preprint arXiv:2005.06493, 2020.

[4] Zhelabugina, E., & Ibrayeva, A. Development and Preliminary Evaluation of KazNERCorpus for Kazakh Named Entity Recognition. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing, pp. 247-259, 2019.

[5] Yessenbayeva, G., & Boranbayev, A. Ontology-Driven Named Entity Recognition in Kazakh: A Comparative Study of Machine Learning Models. In Proceedings of the European Conference on Information Retrieval, pp. 345-358, 2021.

[6] Boranbayev, A., & Makazhanov, A. Enhancing Named Entity Recognition in Kazakh Language Texts Using Ontology-Driven Models. Journal of Natural Language Engineering, 26(4), pp.583-597, 2020.

[7] Kozhirbayev, Z., & Yessenbayev, Z. Named entity recognition for the kazakh language. KazNU Bulletin. July 2020. https://doi.org/10.26577/jmmcs.2020.v107.i3.06

[8] Yeshpanov, R., Yerbolat K., Huseyin A. V. KazNERD: Kazakh Named Entity Recognition Dataset. November 2021. https://arxiv.org/abs/2111.13419

[9] Gislason P.O., Benediktsson J.S. and Johannes R. "Random forests for land cover classification" , Pattern Recognition Letters vol. 27, pp. 294–300, 2006.

[10] Kassymbek, D., & Makhambetov, R. Evaluation of Named Entity Recognition Models for the Kazakh Language: A Comparative Study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 589-602, 2022.

[11] Nurmukhamedov, B., & Karibayev, Z. Analysis of Named Entity Recognition Performance in Kazakh Language Texts Using Ontology-Based Models. Information Processing & Management, 57(2), 102335, 2021.

[12] Yelibayeva, G., Sharipbay, A., Mukanova, A., Razakhova, B. Applied ontology for the automatic classification of simple sentences of the kazakh language. 5th International Conference on Computer Science and Engineering, pp. 13-18., 2020

[13] Baimuratov, A., & Yeskendir, D. Advancements in Named Entity Recognition for Kazakh: A Review and Future Directions. Journal of Language Technology and Computational Linguistics, 34(3), pp. 409-423, 2020.

[14] Boranbayev, A., & Duisenbayev, R. Ontology-Based Approaches for Named Entity Recognition: A Case Study in the Kazakh Language. Knowledge-Based Systems, 235, 106875, 2021.

[15] Lample G., Ballesteros M., Subramanian S., Kawakami K. and Dyer C. "Neural architectures for named entity recognition", arXiv vol. 1603.01360, pp. 1–11, 2016.