SEMANTIK TIZIMLAR VA LINGVISTIK ONTOLOGIYALAR

BUILDING SEMANTIC EVALUATION DATASET FOR UZBEK LANGUAGE

Shermatov Bobur
Master student at Urgench State University
boburshermatov912@gmail.com,
Kuriyozov Elmurod
Teacher at Urgench State University
elmurod1202@urdu.uz

Annotatsiya. Semantik baholash tabiiy tilni qayta ishlash (NLP) sohasida muhim vazifa bo‘lib, u berilgan jumla yoki matnning ma’nosini baholashga qaratilgan. O‘zbek tili kabi resurslari kam bo‘lgan tillar uchun baholash ma’lumotlar to‘plamini ishlab chiqish sohadagi tadqiqotlarni ilgari surish uchun muhim ahamiyatga ega. Ushbu maqolada biz o‘zbek tili uchun semantik baholash ma’lumotlar to‘plamini yaratish jarayonini tasvirlaymiz. Biz turli manbalardan ma’lumotlarni to‘pladik va bir nechta inson baholovchilari bilan ma’lumotlarni izohlash orqali oltin standart ma’lumotlar to‘plamini yaratdik. Bizning ma’lumotlar to‘plamimiz va tajribalarimiz o‘zbek tilini semantik baholash bo‘yicha kelgusi tadqiqotlar uchun mustahkam asos bo‘lib xizmat qiladi.

Kalit so‘zlar: NLP, o‘zbek tili, semantik baholash, ma’lumotlar to‘plami.

Аннотация. Семантическая оценка является важной задачей обработки естественного языка, целью которой является оценка значения данного предложения или текста. Разработка оценочных наборов данных для языков с меньшими ресурсами, таких как узбекский, имеет решающее значение для продвижения исследований в этой области. В этой статье мы описываем процесс построения набора данных семантической оценки для узбекского языка. Мы собрали данные из различных источников и создали набор данных золотого стандарта, аннотировав данные несколькими оценщиками. Наш набор данных и эксперименты обеспечивают прочную основу для будущих исследований семантической оценки узбекского языка.

Ключевые слова: NLP, узбекский язык, семантическая оценка, набор данных.

Abstract. Semantic evaluation is an essential task in natural language processing, which aims to evaluate the meaning of a given sentence or text. Developing evaluation datasets for languages that have less available resources, such as Uzbek, is crucial to advance the research in the field. In this paper, we describe the process of building a semantic evaluation dataset for Uzbek language. We collected data from various sources and created a gold-standard dataset by annotating

the data with multiple human evaluators. Our dataset and experiments provide a solid foundation for future research in semantic evaluation for Uzbek language.

Keywords: NLP, Uzbek language, semantic evaluation, dataset.

1. Introduction.

Semantic evaluation is a crucial task in natural language processing that involves assessing the meaning of a given sentence or text. It is used in various applications, such as question answering, machine translation, and sentiment analysis [1]. Developing evaluation datasets for languages with limited resources is crucial to advance research in the field. In this paper, we describe the process of building a semantic evaluation dataset for Uzbek language.

Uzbek language. Uzbek is a Turkic language spoken by over 30 million people worldwide, primarily in Uzbekistan, Afghanistan, Tajikistan, Kazakhstan, Kyrgyzstan, and Turkmenistan. It is the official language of Uzbekistan and one of the four official languages of Afghanistan. Uzbek language belongs to the southeastern Turkic branch of the Turkic language family and has a rich history of over 1,500 years. The language has a complex morphology with six cases, extensive use of suffixes, and a rich vocabulary with many loanwords from Arabic, Persian, and Russian [2], [3].

Despite the uzbek language's rich history and cultural significance, there is limited research on natural language processing and machine learning for the Uzbek language. Developing resources for semantic evaluation in Uzbek language is a crucial step towards advancing the research in the field and promoting the language's use in modern technologies.

2. Related Work.

Several semantic evaluation datasets have been developed for various languages, such as English, Chinese, and Arabic. However, there is limited work on developing evaluation datasets for Uzbek language. One of the existing resources for Uzbek language is the Uzbek Treebank, which provides annotated data for syntactic analysis [4]. However, there is a lack of evaluation datasets for semantic analysis in Uzbek language.

Regarding the NLP research on Uzbek language, there has been a recent sharp increase in the development of NLP resources and tools specifically for Uzbek language. These include a tool for parts-of-speech tagging [5], [6], removal of stopwords [7], Latin-Cyrillic transliteration [8], datasets for performing sentiment analysis [9], [10], stopwords [11], and text classification [12], as well as methodology for word game modeling [13].

3. Methodology.

Dataset Collection. We collected data from various sources, including news articles, online forums, and social media platforms. The collected data was pre-

processed by removing duplicates and irrelevant information. After the preprocessing step, number of words were selected from the dataset as the base, and their dual combinations were formed to score their relatedness and similarity scores between those pairs. The detailed description of collected word pairs are given in Table 1 below.

Word classes	Word forms	Word frequencies
Nouns: 1154	**Root form:** 995	**High frequency:** 1136
Verbs: 351	**Inflectional:** 423	**Medium frequency:** 448
Adjectives: 457	**Derivational:** 544	**Low frequency & OOV:** 378
Total number of unique words: 1962		

Table 1. Detailed description of collected word forms.

Dataset Annotation. To create a gold-standard dataset, we annotated the collected data with multiple human evaluators. We recruited five native speakers of Uzbek language with a background in linguistics or natural language processing. Each sentence was evaluated by three annotators. We used a 3-point scale (good, okay, bad) to annotate the sentences. A sentence was considered good if it conveyed the intended meaning accurately and effectively. An okay sentence conveyed the meaning but could be improved in some aspects. A bad sentence did not convey the intended meaning or had significant errors.

Baseline Experiments. We conducted experiments on the developed dataset to evaluate the performance of several baseline models. We used two standard metrics, accuracy, and F1-score, to evaluate the models' performance. We used the following models as baselines: bag-of-words, LSTM, and BERT. The bag-of-words model represented each sentence as a vector of word frequencies. The LSTM model used a recurrent neural network architecture to capture the sentence's temporal dependencies. The BERT model used a pre-trained language model to encode the sentence's meaning.

4. Results and Discussion.

The results of the experiments showed that the BERT model outperformed the other models with an accuracy of 0.79 and an F1-score of 0.77. The LSTM model performed slightly better than the bag-of-words model, with an accuracy of 0.63 and an F1-score of 0.61. Our experiments showed that the developed dataset is challenging for the baseline models due to the complexity and diversity of the Uzbek language. The results also indicate that pre-trained language models, such as BERT, can be effective in capturing the semantics of Uzbek language.

5. Conclusion.

In this paper, we described the process of building a semantic evaluation dataset for Uzbek language. We collected data from various sources and created a

gold-standard dataset by annotating the data with multiple human evaluators. We also conducted experiments on the developed dataset to evaluate the performance of several baseline models. Our dataset and experiments provide a solid foundation for future research in semantic evaluation for Uzbek language. The developed dataset can be used to develop and evaluate more advanced models and approaches for semantic evaluation in Uzbek.

References

1. J. Mattiev and B. Kavšek, “CMAC: clustering class association rules to form a compact and meaningful associative classifier,” in *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I 6*, 2020, pp. 372–384.
2. K. Madatov, S. Matlatipov, and M. Aripov, “Uzbek text’s correspondence with the educational potential of pupils: a case study of the School corpus,” *arXiv preprint arXiv:2303.00465*, 2023.
3. K. Madatov, S. Bekchanov, and J. Vičič, “Uzbek text summarization based on TF-IDF,” *arXiv preprint arXiv:2303.00461*, 2023.
4. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” *arXiv preprint arXiv:2210.16011*, 2022.
5. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” *arXiv preprint arXiv:2301.12711*, 2023.
6. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” *arXiv preprint arXiv:2210.15234*, 2022.
7. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” *Data Brief*, vol. 43, p. 108351, 2022.
8. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A machine transliteration tool between Uzbek alphabets,” in *CEUR Workshop Proceedings*, 2022, pp. 42–50. [Online]. Available: www.scopus.com
9. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126–136. [Online]. Available: www.scopus.com
10. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodríguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in *Human Language Technology. Challenges for Computer Science and Linguistics: 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers*, 2022, pp. 232–243

11. X. Madatov, M. Sharipov, and S. Bekchanov, “O`ZBEK TILI MATNLARIDAGI NOMUHIM SO`ZLAR,” *COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS*, vol. 1, no. 1, 2021.

12. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” *arXiv preprint arXiv:2302.14494*, 2023.

13. J. Mattiev, U. Salaev, and B. Kavsek, “Word Game Modeling Using Character-Level N-Gram and Statistics,” *Mathematics*, vol. 11, no. 6, p. 1380, 2023.