"BUILDING AND OPTIMIZING UZBEK–ENGLISH TERMINOLOGICAL DATABASES USING DATA SCIENCE METHODS"

Eshpo'latova Nafisa
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi,
Student, Faculty of Translation Theory and Practice
Phone: +998934877764
Nafisaeshpolatova840@gmail.com
https://doi.org/10.5281/zenodo.21305856

Abstract. This article examines the construction of Uzbek–English terminological databases and the optimization of existing ones through Data Science methods. The aim of the study is to develop a model for a digital terminological database that ensures terminological correspondence between the two languages, can be automatically expanded, and is subject to quality control. The study employed corpus linguistics, natural language processing (NLP), automatic text tagging, and machine learning algorithms. A parallel corpus comprising technical, scientific, and socio-political texts was compiled as the data source, from which term candidates were extracted using statistical and neural methods. The results show that a hybrid approach — combining rule-based filtering with machine learning models — significantly improves the accuracy of term identification and reduces the time required for manual database population. The structure of the resulting database, its metadata schema, and the quality assessment criteria are also presented. In conclusion, Data Science tools represent an effective and promising direction for developing terminological resources for the Uzbek language.

Keywords: Data Science, terminology, terminological database, artificial intelligence, natural language processing, machine translation, corpus linguistics, Uzbek language, English language.

Introduction

In contemporary science, technology, and education, the precision, consistency, and standardization of terminology are of critical importance. The English equivalents of Uzbek scientific and technical terms are often inconsistent, sometimes appearing in multiple variants or being adopted without translation at all. This situation adversely affects the quality of scientific texts, educational literature, technical documentation, and machine translation systems. In recent years, advances in Data Science and artificial intelligence technologies have made it possible to automatically analyze large volumes of textual data, identify recurring linguistic units within them, and systematize these units. This, in turn, considerably accelerates the process of creating and updating terminological databases compared with traditional manual methods. The relevance of this study lies in the fact that existing Uzbek terminological dictionaries and databases are largely static in nature and fail to timely capture newly emerging terms. This creates a need for a dynamic, automated terminological database grounded in Data Science methods. The purpose of the study is to develop a methodology for creating an Uzbek–English terminological database using Data Science methods, automatically enriching it with new terms, and qualitatively optimizing existing databases.

To achieve this aim, the following tasks were defined:

- compiling an Uzbek–English parallel text corpus;
- selecting and testing methods for automatically extracting term candidates from the corpus;
- developing a metadata schema for classifying terms and incorporating them into the database;

● defining criteria for evaluating database quality and analyzing the results.

The theoretical and practical significance of the study lies in the fact that the developed methodology can be applied not only to the creation of terminological databases but also to the advancement of machine translation, electronic dictionaries, and language technologies more broadly.

Literature Review

Considerable experience has been accumulated in foreign linguistics in the field of automatic terminology recognition. Within the framework of corpus-based terminology theory, statistical indicators — such as frequency, mutual information, and TF-IDF — are widely used to extract term candidates. Although such methods have yielded good results for languages such as English, Russian, and German, research in this direction remains insufficient for the Uzbek language. Machine learning-based approaches, including Conditional Random Fields (CRF) and, more recently, transformer-based neural networks (such as multilingual models like BERT and mBERT), have demonstrated high accuracy in extracting and classifying terms from text. These models operate with contextual awareness, which gives them an advantage in identifying polysemous and multi-component terms. In Uzbek linguistics, terminology has traditionally been studied through manually compiled dictionaries and classification systems. While such work has provided an in-depth analysis of the lexical richness of the language, the use of modern computational linguistics tools remains limited. At the same time, the number of projects aimed at creating electronic corpora and NLP resources for the Uzbek language has been increasing in recent years, laying the groundwork for the automation of terminological databases. Unlike previous studies, which rely on either purely statistical or purely neural methods, the present study proposes a hybrid approach combining both, and validates the results through practical testing on the Uzbek–English language pair. In addition, an extensible metadata schema for storing and managing terms is developed — an aspect that has not been sufficiently addressed in prior research.

Materials and Methods

The study is experimental and applied in nature and was carried out in several stages. In the first stage, a corpus consisting of parallel and monolingual texts in Uzbek and English was compiled. The corpus included scientific and technical articles, educational literature, official documents, and mass media texts. The texts first underwent noise removal, tokenization, and morphological analysis. In the second stage, two different methods were applied to extract term candidates: a statistical method (based on frequency and TF-IDF indicators) and a neural method (contextual extraction using multilingual language models). The results of both methods were compared and combined through a hybrid model, in which the statistical method proposed high-frequency candidates while the neural model excelled at identifying context-appropriate terms. In the third stage, the extracted term candidates were manually reviewed and validated by linguist experts, providing the quality control necessary to increase the reliability of the automatically generated results. In the fourth stage, a metadata schema was developed for the database of validated terms. Each entry includes the following fields: term (Uzbek and English variants), domain, definition, example sentence, source, and confidence score. Evaluation was conducted using precision, recall, and F1-score, calculated by comparing the automatically identified terms against an expert-validated “gold standard” list.

4. Results and Discussion

According to the results of the study, the accuracy of term candidates extracted using the hybrid method was significantly higher than that obtained using the statistical method alone. Whereas the statistical method on its own tended to propose more noisy (incorrect) candidates, this rate decreased when combined with the neural model.

The table below presents the comparative results of the three approaches:

Method	Precision	Recall	F1-score
Statistical method (TF-IDF)	0.68	0.61	0.64

Neural method	0.79	0.74	0.76
Hybrid method	0.87	0.82	0.84

As the table shows, the hybrid approach achieved the best results across all indicators. This indicates that combining the strengths of statistical and neural methods constitutes an effective strategy for building terminological databases. The study also found that while the neural model performed better in identifying multi-word terms, the statistical method proved sufficiently effective for detecting single-word, high-frequency technical terms. This finding empirically confirms the theoretical rationale behind the hybrid approach.The structure of the resulting terminological database is extensible: as new texts are added to the corpus, the database can be automatically populated with new term candidates. This constitutes the primary advantage of the database over static dictionaries. At the same time, it was noted as a limitation of the study that a portion of the automatically identified terms requires expert review, particularly in the case of polysemous or domain-specific terms.

5. Conclusion

The study demonstrates that Data Science methods are highly effective for creating and optimizing Uzbek–English terminological databases. The hybrid model, which combines statistical and neural methods, proved superior to traditional manual processing methods in terms of both time efficiency and quality.The objectives of the study were successfully achieved: a parallel corpus was compiled, a methodology for extracting term candidates was developed and tested, a metadata schema for the terminological database was proposed, and the database was evaluated on the basis of quality indicators. The following directions are recommended for future research: extending the database to other language pairs (e.g., Uzbek–Russian), adapting deep learning models to the morphology of the Uzbek language, and making the database available as an open resource to the scientific and educational community. In addition, introducing a continuous update mechanism based on user feedback could further enhance the quality of the database.

References

1. Cabré, M. T. (1999). Terminology: Theory, Methods and Applications. John Benjamins Publishing.
2. Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.
3. Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries.
4. Kageura, K., Umino, B. (1996). Methods of automatic term recognition: A review. Terminology, 3(2).
5. Jurafsky, D., Martin, J. H. (2023). Speech and Language Processing (3rd ed. draft). Stanford University.
6. Sultonov, M. (2021). Problems of Creating an Uzbek Language Corpus. Journal of Language and Literature Education.
7. Manning, C. D., Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
8. Vaswani, A. et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS).