THE DEVELOPMENT OF AN APPLICATION FOR COLLECTING AND FORMING CORPORA OF THE KAZAKH LANGUAGE

Karyukin Vladislav
PhD, Senior lecturer of
Al-Farabi Kazakh National University, Almaty, Kazakhstan.

Abdurakhmonova Nilufar
Doctor of science, Professor,
National University of Uzbekistan
named after Mirzo Ulugbek,Tashkent, Uzbekistan.

Annotation: Today, text processing and analysis tasks are crucial in developing the NLP field. Most of the tasks in this field, like text generation, sentiment analysis, and machine translation, require many language resources. While multi-resource languages, such as English, German, French, Chinese, and others, benefit from large, diverse corpora that facilitate the development of robust language models, low-resource languages face significant challenges related to data scarcity. This paper focuses on the lack of resources for the Kazakh language, which is one of the languages in the Turkic group, for high-quality model training in various NLP tasks. In order to increase the size of the available corpora, the research proposes the approach of parsing the Adilet legislative website, which includes a very large collection of well-structured and error-free texts. The dataset was gathered in two phases. In the first step, the links to the Adilet website were collected using the Content Downloader program. Then, the parser based on the Selenium WebDriver was utilized to extract data and form a dataset with *the title, the date of the law article, its text, and a URL link. The total corpora included 9575 texts.*

Keywords: data parsing, text processing, low-resource languages, Selenium WebDriver, Adilet website, Kazakh.

Аннотация: На сегодняшний день задачи обработки и анализа текста имеют решающее значение в развитии NLP области. Большинство задач в этой области, такие как генерация текста, анализ настроений и машинный перевод, требуют большого количества языковых ресурсов. В то время как многоресурсные языки, такие как английский, немецкий, французский, китайский и другие, получают развитие благодаря большим и разнообразным корпусам, которые способствуют разработке надежных языковых моделей, малоресурсные языки сталкиваются с серьезными проблемами, связанными с нехваткой данных. В данной статье основное внимание уделяется нехватке ресурсов казахского языка, который является одним из языков тюркской группы, для качественного обучения моделей различных NLP задач. Для увеличения размера доступных корпусов в исследовании предлагается подход к анализу законодательного сайта Адилет, который включает в себя очень большую коллекцию хорошо структурированных и не содержащих ошибки текстов. Из данного сайта датасет формировался в два этапа. На первом этапе ссылки сайта Адилет были собраны с помощью программы Content Downloader. Затем парсер на базе Selenium WebDriver был использован для извлечения данных и формирования датасета,

включающего заголовок, дату статьи, ее текст и URL адрес. Всего корпус включал 9575 текстов.

Ключевые слова: парсинг данных, предобработка данных, малоресурсные языки, Selenium WebDriver, веб-сайт Адилет, казахский язык.

Introduction

Currently, text data processing tasks play a significant role. The rapid development of machine learning [Mokhamed, 2024; 2] and large language models [Choi, 2024; 3] requires an increasing volume of corpora for text generation tasks [Jaesub, 2024; 5], sentiment analysis [Karyukin, 2022; 20], machine translation [Guo, 2023; 6], development of question-answer systems [Xu, 2024; 7] and other areas of data analysis. All these tasks require a large amount of language resources [Zhong, 2024; 10]. Multi-resource languages such as English, German, French, Spanish, Portuguese, and Chinese have extensive collections of text data, including scientific articles [Dagdelen, 2024; 7], news resources [Abd El-Mageed, 2024; 11], web pages [Modi, 2022; 15], literary works, and much more. Large corpora allow you to train more powerful and accurate language models. A wide range of available data allows developers to create custom NLP tools for various systems, including personalized recommendation and question-answer systems. For low-resource languages, this situation changes significantly and requires a serious challenge. Without enough data, machine learning algorithms cannot learn effectively, which hinders the creation of high-quality models for machine translation, sentiment analysis, speech recognition, and question-answering systems. This problem is especially relevant for systems processing the Kazakh language. There is a great lack of quality resources for the Kazakh language. The texts are only in Kazakh, Russian, and English on some sites. Such sites include the KazNU website (http://www.kaznu.kz), the Bolashak International Scholarship (http://www.bolashak.gov.kz), the Eurasian National University (http://www.enu.kz), Kazakhstan Post (http://www.kazpost.kz), news portals (http://inform.kz, http://tengrinews.kz) and others. In the work [Karyukin, 2023; 3], multilingual corpora from these sources were used to prepare a Kazakh-English machine translation system. This paper presents an expansion of the Kazakh language corpus by parsing the legislative website https://adilet.zan.kz/kaz/. This website contains a large number of laws of the Republic of Kazakhstan. The texts are of a formal business nature, are grammatically correct in structure, do not contain errors, and include many sections, such as Constitution, Constitutional law, Code, Law, Order, Decree, Agreement, Amendments, Article, Article of agreement, Changes, Charter, Classification, Concept, Conclusion, Conditions, Convention, Methods, Norms, etc.

Related works

In the field of data analytics, many works explore various aspects of data processing. The work [Wei, 2024; 5] analyzes the sentiment of comments in social networks of 22 different domains. The dataset includes more than 200 thousand reviews in Chinese. In the article [Afli, 2016; 10], the construction of statistical machine translation systems, which is demanding regarding the volume of corpora, is being studied. To train the system, texts in the source and target languages are used, obtained from Euronews websites and TED. The model's performance was assessed by analyzing the English-French translation, which showed significant improvements compared to the baseline. Multiclass classification of dialogue texts from the game Fallout labeled by emotional categories New Vegas presented in an article [Hämäläinen, 2022; 2]. Here, the texts are presented in 5 languages: English, Spanish, German, French, and Italian. During

the experiments, classification accuracy values of 54%-55% were obtained for these corpora using multilingual pre-trained BERT and XLMRoBERTa models.

A number of scientific articles also represent research on low-resource languages. The work [Allaberdiev, 2024; 3] presents the creation of a Uzbek-Kazakh corpus for machine translation. This corpus was created in several stages: at the first stage, corpora in Uzbek were collected from publicly available resources (books, websites), then the corpora were translated into Kazakh by a group of experts, and at the next stage, texts in two languages were combined into a parallel corpus. In work [Shymbayev, 2023; 3], a study of an extractive question-answer system using the BERT model is presented for the Kazakh language. Since the development of such a system requires tens of thousands of question-answer pairs, to solve this problem, a machine translation of the Stanford corpus was performed using Google Translate API. In the article [Tolegen, 2024; 4], an empirical study is presented that evaluates text generation models for the resource-poor and morphologically complex Kazakh language. This work trained a large transformer-based language model using a large text corpus of the Kazakh language from different areas. The second dataset contained question-answer problems in Kazakh and Russian languages.

Methodology

A Python web parser was developed to collect Kazakh legislative texts corpora. The first step was to collect links from Adilet (https://adilet.zan.kz/kaz/), the legislative website of the Republic of Kazakhstan. In order to implement this, we used the Content Downloader program. This program was designed to collect links and content from various websites. It is required to set the URL address of the website, the number of threads that determine the speed of data collection, and filtering parameters to gather links. The URL parameters were specified in the filtering window addresses, which must have indicated that a search was done for links to documents in the Kazakh language – *kaz/docs/*. Also excluded from the sample were links that did not contain documents, such as *search/*, *docs/search*, *index/docs*.

The window for collecting links and filtering in the Content Downloader program is presented in Fig. 1 and 2.

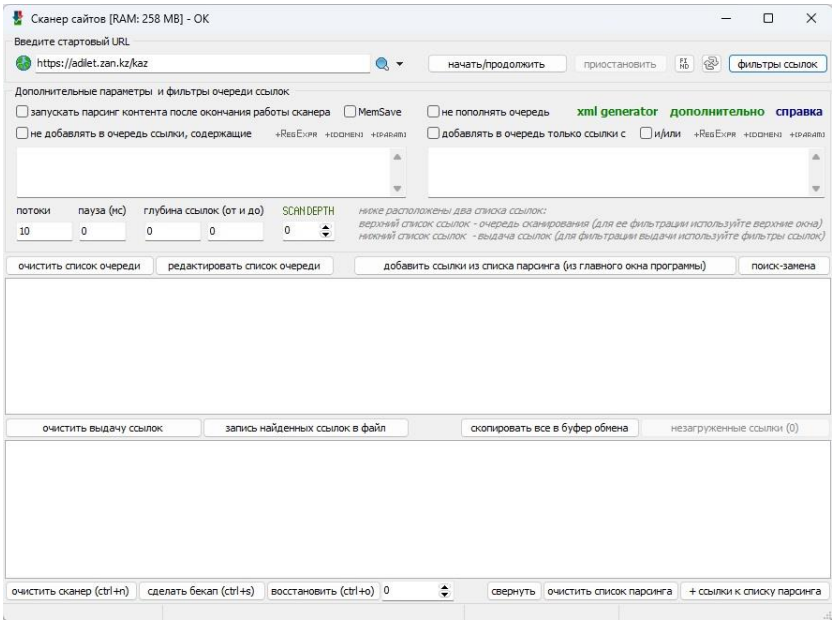

Figure 1 – The program windows of link collection

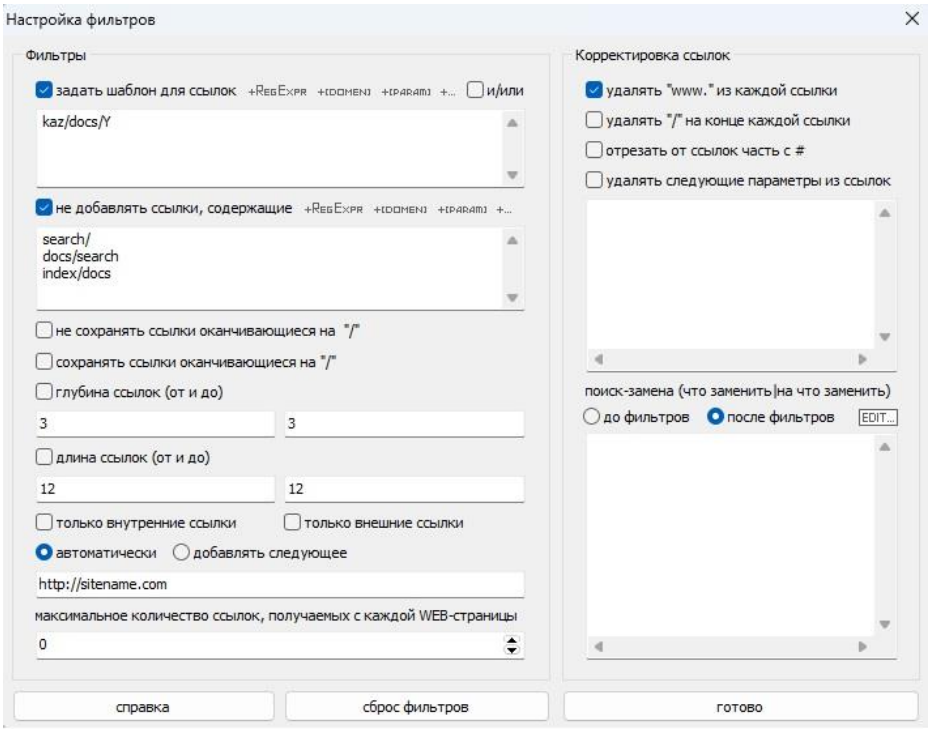

Figure 2 – The program windows of link filtering

Obtained using the Content Downloader program, links were saved in corresponding files. Next, the parser script adds the links extracted from the files to the list. The Selenium WebDriver was used to extract the content of web pages. This tool interacts with the browser at a lower level and uses browser-specific drivers that access it directly. This web driver receives information about the state of the web page and checks the presence of certain elements in it, their properties, and content. Thus, it is very effective for extracting data from web pages.

Results

The parser based on the Selenium WebDriver was utilized to extract the title, the date of the law article, and its text. This data, along with the URL address, was saved in JSON files. Then, the data was read from all JSON files, combined into one dataset, and saved in a CSV file. As a result, the dataset consisting of 9575 texts was obtained. It is shown in Fig. 3.

	header	date	text	source
0	Ветеринариялық бақылауға (қадағалауға) жататын...	2019 - 24 желтоқсан	2014 жылғы 29 мамырдағы Еуразиялық экономикалы...	https://adilet.zan.kz/kaz/docs/H19EK000237
1	Еуразиялық экономикалық одақтың кедендік аумағ...	2019 - 24 желтоқсан	Санитариялық, ветеринариялық-санитариялық және...	https://adilet.zan.kz/kaz/docs/H19EK000238
2	Дәрілік препараттың сапасы бойынша нормативтік...	2018 - 7 қыркүйек	2014 жылғы 29 мамырдағы Еуразиялық экономикалы...	https://adilet.zan.kz/kaz/docs/H18EK000151
3	Дәрілік заттарды сынақтан өткізудің аналитикал...	2018 - 17 шілде	2014 жылғы 29 мамырдағы Еуразиялық экономикалы...	https://adilet.zan.kz/kaz/docs/H18EK000113
4	Өсімдіктен алынатын дәрілік препараттардың сап...	2018 - 10 мамыр	Еуразиялық экономикалық комиссия Алқасы Еуразі...	https://adilet.zan.kz/kaz/docs/H18RK000006
...	...	...	...	...
9570	Қазақстан Республикасының кейбір заңнамалық ақ...	2017 - 3 шілде	РҚАО-ның ескертпесі!Қолданысқа енгізілу тәртіб...	https://adilet.zan.kz/kaz/docs/Z1700000086
9571	Қазақстан Республикасының кейбір заңнамалық ақ...	2021 - 3 қаңтардағы	1-бап. Қазақстан Республикасының мына заңнамал...	https://adilet.zan.kz/kaz/docs/Z2100000406
9572	Қазақстан Республикасының кейбір заңнамалық ақ...	2016 - 6 сәуір	1-бап. Қазақстан Республикасының мына заңнамал...	https://adilet.zan.kz/kaz/docs/Z1600000481
9573	Әділет органдары туралы	2002 - 18 наурыз	Ескерту. Орыс тіліндегі мәтінге өзгеріс енгізі...	https://adilet.zan.kz/kaz/docs/Z020000304_
9574	Өзбекстан Республикасының Еуразиялық экономика...	2006 - 4 шілде	РҚАО-ның ескертпесі!Келісім қолданысын тоқтатт...	https://adilet.zan.kz/kaz/docs/Z060000152_

9575 rows × 4 columns

Figure 3 – The dataset obtained with the use of the parser

Conclusion

This paper presents a study on creating text corpora for the low-resource Kazakh language, which is important for tasks in the field of NLP, such as machine translation, question-answer systems, text generation, etc. The main source of resources for quality texts in the Kazakh language was the legislative website Adilet, which is characterized by a high level of structure and grammatical correctness of the texts. An automated parsing script was used to obtain texts from this site, significantly simplifying and speeding up the process of collecting text data while maintaining high-quality information. The list of site links was generated with the use of the Content Downloader program, which returned a list of URLs for a given site. When setting up the program, the number of threads that determine the speed of data collection and filtering parameters was also set. The resulting links are added to the list and used in the parser program, where Selenium WebDriver is used to extract the content of web pages. During the experiment, a corpus was obtained that included 9575 legislative texts. In the future, the resulting corpus will be used to create a model of a question-answer system for the Kazakh language. It is also important to continue developing parsing methods to improve the accuracy and efficiency of data extraction, which will significantly contribute to the development of digital technologies in Kazakhstan.

References:

1. Mokhamed T., Harous S., Hussein N. *et al.* Comparative analysis of Deep Learning and Machine Learning algorithms for emoji prediction from Arabic text. *Social Network Analysis and Mining*, 14, 67, 2024. https://doi.org/10.1007/s13278-024-01217-w
2. Choi J., Lee B. Accelerating materials language processing with large language models. *Communication Materials*, 5, 13, 2024. https://doi.org/10.1038/s43246-024-00449-9
3. Jaesub Y., Jong-Seok L. Learning from class-imbalanced data using misclassification-focusing generative adversarial networks. *Expert Systems with Applications*, vol. 240, 2024. https://doi.org/10.1016/j.eswa.2023.122288
4. Karyukin V., Mutanov G., Mamykova Z. *et al.* On the development of an information system for monitoring user opinion and its role for the public. *Journal of Big Data*, 9, 110, 2022. https://doi.org/10.1186/s40537-022-00660-w
5. Guo S., Deng N., He Y. ISTIC's Neural Machine Translation Systems for CCMT' 2023. *Communications in Computer and Information Science*, vol. 1922, 2023. Springer, Singapore. https://doi.org/10.1007/978-981-99-7894-6_9
6. Xu L., Lu L., Liu M. *et al.* Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. *Heritage Science*, 12, 118, 2024. https://doi.org/10.1186/s40494-024-01231-3
7. Zhong Y., Goodfellow Sebastian D. Domain-specific language models pre-trained on construction management systems corpora. *Automation in Construction*, vol. 160, 105316, 2024. https://doi.org/10.1016/j.autcon.2024.105316
8. Dagdelen J., Dunn A., Lee S. *et al.* Structured information extraction from scientific text with large language models. *Nature Communications*, 15, 1418, 2024. https://doi.org/10.1038/s41467-024-45563-x
9. Abd El-Mageed A.A., Abohany A.A., Ali A.H. *et al.* An adaptive hybrid African vultures-aquila optimizer with Xgb-Tree algorithm for fake news detection. *Journal of Big Data*, 11, 41, 2024. https://doi.org/10.1186/s40537-024-00895-9

10. Modi A., Shah K., Shah S. *et al.* Sentiment Analysis of Twitter Feeds Using Flask Environment: A Superior Application of Data Analysis. *Annals of Data Science*, 11, 159–180, 2024. https://doi.org/10.1007/s40745-022-00445-1
11. Karyukin V, Rakhimova D, Karibayeva A, Turganbayeva A, Turarbek A. The neural machine translation models for the low-resource Kazakh–English language pair. PeerJ Computer Science 9: e1224, 2023. https://doi.org/10.7717/peerj-cs.1224
12. Wei Zh., Zhang Sh. A structured sentiment analysis dataset based on public comments from various domains. Data in Brief, vol. 53, 110232, 2024. https://doi.org/10.1016/j.dib.2024.110232
13. Afli H., Barrault L., Schwenk H. Building and using multimodal comparable corpora for machine translation, *Natural Language Engineering*, 22(4), pp. 603–625, 2016. https://doi.org/10.1017/S1351324916000152
14. Hämäläinen M., Alnajjar K. , Poibeau T. Video Games as a Corpus: Sentiment Analysis using Fallout New Vegas Dialog. In Proceedings of the 17th International Conference on the Foundations of Digital Games (FDG '22). Association for Computing Machinery, New York, NY, USA, Article 56, 1–4, 2022. https://doi.org/10.1145/3555858.3555930
15. Allaberdiev B., Matlatipov G., Kuriyozov E., Rakhmonov Z. Parallel texts dataset for Uzbek-Kazakh machine translation. Data in Brief, vol. 53, 110194, 2024. https://doi.org/10.1016/j.dib.2024.110194
16. Shymbayev M., Alimzhanov Y. Extractive Question Answering for Kazakh Language. *IEEE International Conference on Smart Information Systems and Technologies (SIST)*, Astana, Kazakhstan, pp. 401-405, 2023. https://doi.org/10.1109/SIST58284.2023.10223508
17. Tolegen G., Toleu A., Mussabayev R., Zhumazhanov B., Ziyatbekova G. Generative Pre-Trained Transformer for Kazakh Text Generation Tasks, *19th International Asian School-Seminar on Optimization Problems of Complex Systems (OPCS)*, Novosibirsk, Moscow, Russian Federation, pp. 144-118, 2023. https://doi.org/10.1109/OPCS59592.2023.10275765
18. Ismailov, A. S., & Abdurakhmonova, N. (2022). The development of Alisher stemmer for Uzbek Language. Science and Education, 3(4), 187-213.
19. Abdurakhmonova, N., & Tuliyev, U. (2018). Morphological analysis by finite state transducer for Uzbek-English machine translation/Foreign Philology: Language. Literature, Education, 3, 68.
20. Abdurakhmonova, N., & Urdishev, K. (2019). Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL), 6(1-2019), 131-7.
21. Abduraxmonova, N. Z. (2018). Linguistic support of the program for translating English texts into Uzbek (on the example of simple sentences): Doctor of Philosophy (PhD) il dis. aftoref.