TRANSFORMERS FOR NAMED ENTITY RECOGNITION IN ZERO AND FEW SHOT CONTEXTS

Safarov Ruzmat Abdiqayum o‘g‘li
Junior Researcher, V.I. Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Tashkent.
Davronov Rifqat Rahimovich
Ph.D., Senior Researcher V.I. Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Tashkent.

Аннотация. Крупные языковые модели Large Language Models набирают популярность в последнее время в связи со своими способностями решать многочисленные задачи обработки естественного языка. Настоящее исследование обсуждает применение трансформерных моделей, таких как RoBERTa и T5, для проблемы именованных сущностей NER на узбекском языке при условиях обучения без примеров Zero-shot и малым числом примеров Few-shot. Мы используем новые подходы для набора данных и активное обучение для улучшения модели (качество) при минимальном количестве размеченных примеров. Это новаторское по сравнению с предыдущими исследованиями, в основном нацеленными на контролируемое обучение при больших размеченных корпусах. Мы демонстрируем, что даже при ограниченной разметке можно достичь заметных результатов в качестве предсказания. Как триггер к развитию определенных областей обработки естественного языка для языков с ограниченным языковым ресурсом, наша работа существенно привнесла свой вклад в эти направления.

Ключевые слова: распознавание именованных сущностей, обучение без примеров, обучение с малым числом примеров, узбекский язык, трансформеры, RoBERTa, T5, отбор данных, низкоресурсные языки

Abstract. Large Language Models (LLMs) have gained popularity recently due to their ability to solve numerous natural language processing tasks. This study discusses the application of transformer models such as RoBERTa and T5 to the Uzbek NER problem under zero-shot and few-shot training conditions. We use novel dataset approaches and active learning to improve the model performance with a minimal number of labeled examples. This is groundbreaking compared to previous studies that mainly focused on supervised learning with large labeled corpora. We demonstrate that even with limited labeling, remarkable results in prediction quality can be achieved. As a trigger for developments in certain areas

ISSN 3030-3907

of natural language processing for languages with limited language resource, our work makes significant contributions to these directions.

Keywords: Named Entity Recognition, Zero-shot Learning, Few-shot Learning, Uzbek Language, Transformers, RoBERTa, T5, Data Targeting, Low-resource NLP

Annotatsiya. Large Language modellarining yirik til modellari so'nggi paytlarda tabiiy tilni qayta ishlashning ko'plab muammolarini hal qilish qobiliyati tufayli mashhurlikka erishmoqda. Ushbu tadqiqot RoBERTa va T5 kabi transformator modellarni Zero-shot va kam sonli Few-shot misollarisiz o'qitish sharoitida o'zbek tilidagi ner nomli ob'ektlar muammosi uchun qo'llashni muhokama qiladi. Biz ma'lumotlar to'plami uchun yangi yondashuvlardan foydalanamiz va modelni (sifatni) minimal belgilangan misollar bilan yaxshilash uchun faol treningdan foydalanamiz. Bu avvalgi tadqiqotlar bilan taqqoslaganda, asosan, katta belgilangan holatlarda nazorat ostida o'rganishga qaratilgan. Biz hatto cheklangan belgilar bilan ham bashorat sifatida sezilarli natijalarga erishish mumkinligini ko'rsatamiz. Til resurslari cheklangan tillar uchun tabiiy tilni qayta ishlashning muayyan sohalarini rivojlantirishga turtki sifatida bizning ishimiz ushbu yo'nalishlarga sezilarli darajada hissa qo'shdi.

Kalit so'zlar: nomlangan ob'ektlarni tanib olish, misollarsiz o'qitish, kam sonli misollar bilan o'qitish, o'zbek tili, transformatorlar, RoBERTa, T5, ma'lumotlarni tanlash, kam resursli tillar

1. Introduction

Natural language processing (NLP) has advanced significantly in recent years, mostly as a result of the emergence of transformer models like BERT, RoBERT'a, and T5. These models have shown exceptional performance in Named Entity Recognition (NER) and other NLP tasks.

The analysis of these research can be substantially aided by automating the information extraction process. By recognizing entities like persons, places, businesses, dates, and currencies in the text, NLP and NER techniques are crucial in this situation [1]. After that, these entities are given the proper labels; for instance, "London" can be recognized and designated as a place.

Furthermore, NER is a crucial technique for extracting information and is frequently combined with other NLP tasks like question-answering and summarization [1]. Its universality spans a number of domains, such as the legal environment [4], automatic voice recognition [3], and the biomedical and financial industries [2]. Automating processes and improving productivity across a range of domains are made possible by accurately identifying and classifying textual items.

ISSN 3030-3907

Examining several methods for entity recognition is essential given the significance of NER. There are several ways to recognize entities in texts [1, 5]. Manually created dictionaries and rules that rely on the amount and quality of specific words and regulations are the foundation of traditional methods. However, despite the fact that they need a great deal of feature engineering and domain expertise, machine learning techniques employ pattern learning to identify items in unseen texts [5]. On the other hand, because deep learning techniques may automatically uncover hidden characteristics, they have gained popularity for entity recognition [1, 5].

The BERTbek [6], a pre-trained language model based on the BERT architecture, was recently trained for the Uzbek language by a number of scientists. Numerous NLP tasks, such as named entity identification, multi-label topic categorization, and sentiment analysis, have been used to evaluate this model. The outcomes demonstrated that BERTbek outperforms other basic models and multilingual BERT models on these tasks.

NER is the process of automatically recognizing and categorizing named entities in a text, including names, locations, organizations, and other important phrases. Because of their capacity to consider the context and links between words in a phrase, models like RoBERTa and T5 have demonstrated notable progress in this area. For instance, the RoBERTa model, which uses a head to classify tokens, consistently and effectively processes text and has high accuracy in NER tasks.

In this paper, we explore the application of transformer-based models, specifically RoBERTa and T5, for Named Entity Recognition (NER) in the Uzbek language under zero- and few-shot learning settings. We conduct a comparative analysis to evaluate their performance when limited or no annotated data is available. The outcomes of this study are expected to provide valuable insights for advancing NLP tools for Uzbek and other low-resource languages, particularly in scenarios where labeled data is scarce.

2. Background

Pre-learning language models has resulted in notable advancements in a number of natural language processing tasks in recent years. Although models like ELMo, GPT, BERT, XLM, and XLNet have significantly advanced this sector, it is still difficult to compare various strategies because of the disparities in hyperparameters and training data sets.

One of the most widely used models that has demonstrated notable gains in a variety of applications is BERT (Bidirectional Encoder Representations from Transformers). This model was under-trained, though, as the study shown, and it can match or surpass the performance of later models with the right hyperparameter

ISSN 3030-3907

adjustments and more training data. RoBERTa [7] (Robustly Optimized BERT Approach), a model that incorporates a number of significant modifications, was developed as a result of this research. These modifications include lengthening the training period, employing a lot of data, dropping the next sentence prediction task, learning on longer sequences, and dynamically altering masking patterns. The T5 (Text-to-Text Transfer Transformer) model [8] is designed to unify all text processing tasks into a text-to-text format. This means that each task, whether it's translating, summarizing text, answering questions, or classifying, is presented as a task of converting one text fragment into another. This approach allows you to use a single modeling and training process for a wide range of tasks.

The study examines different pre-learning objectives, architectures, untagged data sets, and transfer strategies, contrasting them on a range of language comprehension tasks. Advanced outcomes are obtained on tasks including text categorization, question answering, and summarization by utilizing the new Colossal Clean Crawled Corpus dataset and the model's scalability. Data, trained models, and source code have been released to support further studies in the area of learning transfer for NLP.

3. Methods: RoBERTa and T5 models.

A brief overview of the Uzbek language and the procedures used to train the RoBERTa and T5 models for it are provided in this section.

3.1. Uzbek language.

To distinguish it from southern Uzbek, another dialect of Uzbek spoken by the ethnic Uzbek minority in Afghanistan, Uzbek (native: O‘zbek tili) is a member of the Eastern Turkic or Karluk branch of the Turkic languages. Together, they form a macro-language. It is Uzbekistan's sole national language and its first official language. More over 30 million people in Uzbekistan alone speak Uzbek, and over 10 million more in neighboring Central Asian nations, the southern region of the Russian Federation, and northeastern China.

3.2 RoBERTa-based model.

Two years ago, we trained the RoBERTa model in Uzbek and posted it on the Hugging Face platform [9]. The process of teaching our RoBERTa model in Uzbek included several stages. First, we have collected a large corpus of texts in Uzbek, consisting of 5.4 million sentences. This corpus covers a wide range of topics and styles, which allows the model to learn from a variety of material and improve her ability to generalize (see section 4.1). The model was then trained for 25 epochs. Each epoch represents a complete passage through the entire data set, which allows the model to gradually improve its representations and adapt to the specifics of the Uzbek language. During the training, we used various optimization

ISSN 3030-3907

and regularization techniques such as gradient clipping, random zeroing, and early stopping to prevent overfitting and ensure the stability of the learning process. Our RoBERTa model in Uzbek was placed on the Hugging Face platform, which allows researchers and developers to easily integrate it into their applications and conduct additional experiments.

3.3. T5-based model.

In the past few years, interest in creating and utilizing machine learning models for the purposes of word processing in a variety of languages has risen rapidly. Recently, we made the decision to adapt and train the T5 model in Uzbek language to contribute to creating new resources and tools for word processing in Uzbek. We trained our model using a fully populated dataset encompassing 18.1 million sentences, which provides extensive coverage of a variety of topics and language types. This model training occurred over 10 epochs, which is appropriate for achieving performance and quality of text generation. After training we placed our model on the Hugging Face [10] platform, which allows researchers and developers to freely access and test our model. We see this access to our model as an important step towards increasing access to good quality language models for the Uzbek language and furthering the research directed at this under researched area.

4. Experiments and Results

This section describes the results of evaluating pre-trained RoBERTa and T5 models adapted to perform Zero-shot and Few-shot learning tasks in the field of natural language processing. We have completed the models previously trained in the previous stage for our target tasks. At this stage, we again used special classes provided by the transformer's library (unless explicitly stated otherwise, default parameters were used), as well as training and validation datasets for further training.

4.1. Data Transformation Method

We used a prompt-based data transformation to support zero- and few-shot learning in the task of Named Entity Recognition (NER) with general-purpose text. While uzbek language datasets can have complex entity taxonomies, our work focused on standard NER categories, such as:

- **ORG** (organizations)
- **PER** (persons)
- **LOC** (locations)
- **DATE** (dates)
- **TIME** (time expressions)
- **NUMBER** (numerical quantities)

ISSN 3030-3907

Each sentence in the dataset was converted to many training instances, one for each entity type present and this allowed us to accomplish the transformation of a multi-class sequence labeling task into a set of binary token classification tasks. [11]

Here, LabelName refers to the entity class (i.e., "LOC") of interest, Sentence is the natural language input, and the <**SEP**> token is a delimiter between the prompt and the text being annotated.

Example:

Uzbek: Bugun Toshkent shahridagi Amir Temur xiyobonida dam oldik

English: Today we rested on Amir Temur Avenue in Tashkent

This sentence includes two named entities:

• "Amir Temur xiyoboni" — LOC

• "Toshkent" — LOC

So the sentence is processed twice; once for each entity class. The input for LOC would read:

LOC Bugun Toshkent shahridagi Amir Temur xiyobonida dam oldik

A corresponding output label vector would indicate the tokens "Amir Temur xiyobonida" and "Toshkent" as having indicator 1s, with everything else receiving 0s.

Output Encoding:

• Tokens matching the specified **LabelName** are labeled as **1**

• All other tokens are labeled as **0**

This binary encoding simplifies the learning objective and allows the model to focus on detecting one entity type at a time. The same approach is applied across all entity classes.

The final corpus of data consists of 1,200 sentences and more than 20,000 words from the uza.uz website and were tagged with six categories (ORG, PER, LOC, DATE, TIME, and NUMBER) with the help of the Label Studio web tool [12]. Each file was checked at least twice by native speakers and native Uzbek language expert, allowing for good quality data. The developed corpus is an important resource for recognizing named entities in the Uzbek language as well as to accommodate other Turkic languages, including for example, the Karakalpak language or the Oguz dialect of the Uzbek language.

4.2. Fine tuning models.

Named Entity recognition. To solve the problem the RoBERTa and T5 models were used, each of which was set up in much the same fashion, which includes the following key steps.For each of the models the data was split into 80%

ISSN 3030-3907

training, 10% validation and 10% test samples. After splitting the data, the text was tokenized and aligned to the NER labels.

The transformers library was used to import the RoBERTa model with the addition of a classification layer for predicting NER labels. Hyperparameters were configured, including the learning rate (from 2e-5 to 5e-5), the pocket size (16), the number of epochs (3) and the maximum sequence length (128). An AdamW optimizer and a linear learning rate shader with warm-up are used. Metrics such as accuracy, completeness, precision and F1 measures were used to evaluate the model, and performance was monitored on a validation sample to adjust hyperparameters

The NER task was presented as a sequence generation task, where the source text provides the input and the output is a sequence of entity labels. T5 uses a "span corruption" retraining strategy, where segments of random text are replaced with a masking token and then the model is trained to predict the masked segments. We selected model parameters based on computer process capacity and accuracy (model size: small, base, large, 3B, 11B). Hyper-parameters included learning rate (1e-5 to 1e-4), batch size (8 to 64), number of epochs (3 to 10), AdamW optimizer, warm-up steps (about 10% of total steps), and weight decay (0.01).

4.3. Results

Named Entity recognition. Table 3 presents a comparative assessment of the two uzRoBERTa and uzT5 models by several metrics, namely: accuracy (precision), completeness (recall) and F1-measure (f1-score) in various categories such as ORG (organizations), PER (persons), LOC (locations), DATE (dates), TIME (time) and NUMBER (numbers).

Table 3.

Performance of the uzRoBERTa and uzT5 models by metrics for various categories of named entities.

Model	Class	Epoch	Number of supporting examples			
			0-shot	1-shot	10-shot	100-shot
UzRoberta	ORG	2	40.55	59.81	75.03	78.25
	PER	2	63.84	67.21	76.92	84.02
	LOC	2	51.36	66.35	76.01	82.58
	DATE	2	26.92	35.46	50.27	71.98
	TIME	2	61.18	78.58	79.07	85.15
	NUMBER	2	30.08	47.23	57.86	65.98

ISSN 3030-3907

uzT5	ORG	2	38.17	57.86	74.99	77.87
	PER	2	60.82	67.25	75.91	82.56
	LOC	2	49.38	62.97	75.22	81.15
	DATE	2	24.99	34.86	49.96	69.95
	TIME	2	58.28	75.68	77.24	83.22
	NUMBER	2	30.02	45.03	55.96	64.92

From the results,we can observe that uzRoBERTa yields better results over almost all entity classes and setting than uzT5. Especially in the ORG category, uzRoBERTa performs with the F1-score 40.55% in the 0-shot setting, and it grows 78.25% at 100-shot; it is slightly higher than uzT5 as its 77.87%. In PER, uzRoBERTa's range is 63.84% to 84.02% while that of uzT5 is 60.82% to 82.56%. The same trend can be seen on the LOC class, where uzRoBERTa reaches 51.36% in 0-shot and 82.63% in 100-shot, again with a narrow lead over uzT5

This distinction is even more pronounced in the DATE category, where it has always been more complex due to its contextual viability. Here, uzRoBERTa improves from 26.92% in 0-shot setting, to 71.98% at 100-shot, always performing better than uzT5 at every level and significantly better in the earlier levels (for example; 0-shot = 24.99% in uzT5). In the TIME class, uzRoBERTa consistently scores higher, eventually finishing with a F1-score of 85.15% against uzT5's 83.22%. And all of NUMBER began near 30% in the 0-shot setting and both completed 100-shot (~65%) with uzRoBERTa slightly ahead.

To sum up, these results demonstrate the effectiveness of uzRoBERTa in low-resource conditions. The model performs better in a zero-shot one-shot setting—meaning that it has not seen examples during training—and scales effectively as the data increases. Given this, uzRoBERTa is a viable option for realistic real-world NER applications in the Uzbek language in situations wherein labeled data is either difficult and/or expensive to collect.

5. Conclusion

As a result of this research, we have shown that transformer-based models, especially uzRoBERTa and uzT5 for Named Entity Recognition (NER) in Uzbek, can learn from small amounts of annotated data, zero- and few-shot learning methods made this exploration possible. Our careful experimental design indicates that using very few annotated data, uzRoBERTa, in particular, can successfully complete NER tasks with I-precision and high generalizability. The contribution to our NER research was primarily the application of the targeted data selection type that included high-entity-density sampling and active learning which promoted model-learned NER development in 6 entity categories. This enabled the model to maximally train and learn from limited data, generally increasing the effectiveness

ISSN 3030-3907

of the uzRoBERTa and uzT5 models across all 6 category entities. The input prompt-based approach we tested with the models also helped facilitate few-shot learning processes and aided the uzRoBERTa and uzT5 models in generalizing to unseen entity categories. Results from comparative analysis indicated the uzRoBERTa-Ts model outperformed the uzT5 across all configurations of zero/few-shot learning. In fact, the biggest improvement between uzRoBERTa and uzT5 was with ORG and DATE entities, where uzRoBERTa demonstrated larger recall and performance with F1-scores than uzT5. This difference was greater for low-shot settings with uzRoBERTa being more resilient in undeveloped resource conditions.These results are especially important for the development of natural language processing technologies for Uzbek and other low-resource languages. The ability to train effective models with limited annotation allows for novel applications of NER systems in the real world (e.g., information extraction, document classification, conversational AI).

In future work, we would like to further improve model performance through additional data augmentation, cross-lingual transfer, and syntactic signals. We would also like to continue expanding the annotated datasets and apply these approaches to other types of linguistic tasks, to contribute to the larger goal of making NLP available for low-resource languages.

REFERENCES

1. J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Trans. on Knowl. and Data Eng. 34 (1) (2022) 50–70. doi:10.1109/TKDE.2020.2981314.

2. J. C. Salinas Alvarado, K. Verspoor, T. Baldwin, Domain adaption of named entity recognition to support credit risk assessment, in: B. Hachey, K. Webster (Eds.), Proceedings of the Australasian Language Technology Association Workshop 2015, Parramatta, Australia, 2015, pp. 84–90.

3. M. Del Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong, P. Żelasko, M. Jetté, Earnings-21: A Practical Benchmark for ASR in the Wild, in: Proceedings of Interspeech 2021, 2021, pp. 3465–3469. doi:10.21437/Interspeech.2021-1915.

4. P. H. Luz de Araujo, T. E. de Campos, R. R. R. de Oliveira, M. Stauffer, S. Couto, P. Bermejo, LeNER-Br: a dataset for named entity recognition in Brazilian legal text, in: International Conference on the Computational Processing of Portuguese (PROPOR), Lecture Notes on Computer Science (LNCS), Springer, Canela, RS, Brazil, 2018, pp. 313–323. doi:10.1007/978-3-319-99722-3_32.

ISSN 3030-3907

5. B. Song, F. Li, Y. Liu, X. Zeng, Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison, Briefings in Bioinformatics 22 (6) (2021) bbab282. doi:10.1093/bib/bbab282.
6. Kuriyozov, E., Vilares, D., & Gómez-Rodríguez, C. (2024). BERTbek: A Pretrained Language Model for Uzbek. SIGUL2024 Workshop, pages 33–44
7. Yinhan Liu, Myle Ott, Naman Goyal et. RoBERTa: A Robustly Optimized BERT Pretraining Approach, https://arxiv.org/abs/1907.11692v1
8. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://arxiv.org/abs/1910.10683
9. Adilova, F., Davronov, R., Kushmuratov, S., Safarov, R. (2022). Uztext-3Gb-BPE-Roberta (Revision 0c87494). Hugging Face. https://huggingface.co/rifkat/uztext-3Gb-BPE-Roberta.
10. Davronov, R. (2022). T5 Base Uzbek. Hugging Face. Retrieved from https://huggingface.co/rifkat/t5-base-uzbek
11. Safarov, R., Kushmuratov, S. (2024). NER dataset for Uzbek language [Dataset]. GitHub. https://github.com/rsafarov1995/NER-dataset
12. 12. Košprdić, M., Prodanović, N., Ljajić, A., Bašaragin, B., & Milošević, N. (2024). *From Zero to Hero: Harnessing Transformers for Biomedical Named Entity Recognition in Zero- and Few-shot Contexts*. arXiv. https://arxiv.org/abs/2305.04928

ISSN 3030-3907