THE IMPORTANCE OF LARGE LANGUAGE MODEL IN UZBEK-ENGLISH TRANSLATION SYSTEM

Elmira Nazirova[1], Nodira Samigova[2], Kamola Usmonova[3], Mamura Uzakova[4]

[1]Professor of Tashkent University of Information Technologies named after Mukhammad Al-Khwrazmi, Tashkent, Uzbekistan

[2]Senior lecturer of the Pharmaceutical Education and Research Institute

[3]PhD of the the Department of Multimedia,Technologies, Tashkent University of Information Technologies, named after Al-Khwrazmi Tashkent, Uzbekistan

[4]Senior lecturer of the Department of Artificial Intelligence, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

Corresponding author: kamolausmonova93@gmail.com

Annotation: The recent years there are a significant rise appling artificial intelligence across different domains, it is used to improve performance as well to get more accurate results. In this article delve into study carried on Uzbek-English and vice verce translation system which include models also. Initially, in the study rule-based approach was used and get translation results. Yet, to achieve translation quality and accuracy it is decided to do hybride approach. We used 142838 paired corpus and did normalization before training it different 3 models. The result adduces satisfactory result as translation system can translate various topics in Simple and complex sentences.

Keywords: rule-based approach, model fine-tuning, corpus-based traning, hybrid approaches.

Introduction

In order to exchange scientific knowledge across different communities language plays important role in human communication. Yet, there are developing technology and science so some new concepts and specialized terms make difficult to understand the information. Especially, translating words or text from Uzbek (agglutinative language) to English (analytic language) creates challenges as Uzbek language has different grammatical structures and linguistic features when comparing English. Hence, it is important to appropriate computational approaches in both languages' morphology as Uzbek language word formed attaching suffixes to root while English follows limited inflectional patterns.

Traditionally, communication relied on human, although human translation has high-quality results, it requires time and effort. As increasing volume digit around the world machine translation is practical alternative for supporting communication in different areas [4]. In the field of natural

langugae processing, reliable translation systems is one of the essential task. It is not deniable, recently artificial intellegence and neural language models have improved signaficantly the quality of machine translation systems. These days Uzbek language also hold its position around the word. Considering the fact that more than 34 million people speak Uzbek language. That is why, national initiatives emphasize the importance of developing translation system that is necessary and timely and can translate various texts from Uzbek and English or vice verse [2]. Hence, we developed Uzbek- English machine translation which is named UzMarine translation system. Creating it several neural machine translation models which based on large language model architechture were implmented and identified the most effective solution. We prepaired dtaset and trained different 3 models namely, T5, Helsinki and no language left behind models and achieved satisfactory result the third model respectively.

In this study, we present the development of an Uzbek–English machine translation system called the UzMarine translation system. Initially, a rule-based translation approach was implemented to establish baseline results. Although this method provided preliminary translations, further improvements were required to increase translation accuracy and linguistic flexibility. Therefore, several neural machine translation models based on large language model architectures were evaluated and compared in order to identify the most effective solution. To support the training and evaluation process, paired parallel corpus consists of 142848 sentence, so doing hybride aproaches (rule-based and model) comprises quality translation with 3% comparing with Google translator. The effectiveness of combining analysis and neural model improve machine translation system.

Methodology

Translation system consists of several phases as preprocessing, rule-based and model training process and evaluation and output stages [1].Linguistic preprocessing procedures, we collect datasets and normalize it such as omitting unneccessary symbols, formating inconsistencies and duplicated enteries [3]. The second phase rule-based translation as Uzbek languages is agglutinative language with complex suffix- based formation we decided to implement using

mathematical representations and programming procedures and we also experimented models training and we chose among three models, satisfactory one and used. This helped us constructing the proposed UzMarine translation system. We evaluated output with Meteor matrix[5] and got 3 % higher accuracy comparing with Google translating system.

Discussion

Developing translation system, creating proper mathematical model is very essential [7]. After analyzing both languages morphology we created the formulas for Uzbek and English languages. For uzbek language: If the root is nominal this matematical modal for uzbek language [6].

$$W = \oplus\downarrow \sum_{i=1}^{\tau} P_i \oplus \sum_{p=1}^{n} R_i \oplus\downarrow \sum_{j=1}^{c} I_j \oplus\downarrow \sum_{k=1}^{\iota} S_k \oplus\downarrow \sum_{r=1}^{q} K_r$$

Here, W-word, P- refix, R-root, I- inflectional affixes, S-syntactic affixes, K- indicates predicativity. If the root is verbal then this matematical model for this types of word for uzbek language

Here, W-word, R-root, I- inflectional affixes, T-tense, M- mood, Sh-personal indication.

$$W = \oplus \sum_{p=1}^{n} R_i \oplus\downarrow \sum_{j=1}^{c} I_j \oplus\downarrow \sum_{k=1}^{l} T_k \oplus\downarrow \sum_{r=1}^{q} M_r \oplus\downarrow \sum_{r=1}^{r} SH_r$$

For English language:

lS тy *dd*

$$W = \sum C_j \left(\sum Pm\right)\left(\oplus\downarrow \sum Pri \oplus \sum Rk \oplus\downarrow \sum Ij \oplus\downarrow \sum Dp\right)$$

j=1 *m* *i*=1 *k* *j*= *p*=1

=1 =1 1

Here, C- grammatical case, P-possessive**,** Pr- refix, R-root, I- inflectional affixes, D-derivational affix.lSyd

$$W = \left(\sum Augj\right)\left(\sum Im\right)\left(\oplus \sum Ri \oplus\downarrow \sum Dp\right)$$

j=1 m= p=1 p=1

1

Here, Aug- auxiliary, I-inflectional affixes, Pr- refix,D-derivational affix.

For example: Uzbek: talabalarimizdan =talaba(root)+ lar (Inflectional affix)+imiz(syntactic affix- person)+ dan(grammatical case). This is translated from our students.

From(grammatical case), our(possessive),student(root)+s(inflectional affix).We constructed simple and comlex sentence mathematical model and have used them to translate. In rule based approach we have analyzed 32 sentences Uzbek and corresponds Engish sentences. Furthermore, experimented all with model and develop UzMarine translation system(photo 1). Training pipline(photo 3) followed preprocessing and data preparetion workflow to improve translation quality. Firstly parallel corpus was done and examine carefully to record single aligned sentence pair. This step is very essential before training. After this step, normalization procedures should be done so unnecessary apostrophes and formatting inconsistencies and symbols. Following this, dataset and validation were divided and models are experimented as well as Language identifiers[8] such as *eng_Latn* and *uzn_Latn*. Finally padding and truncation techniques were also applied. At the end, we achivied successful outcome. In the future size of parallel corpus will be expanded.

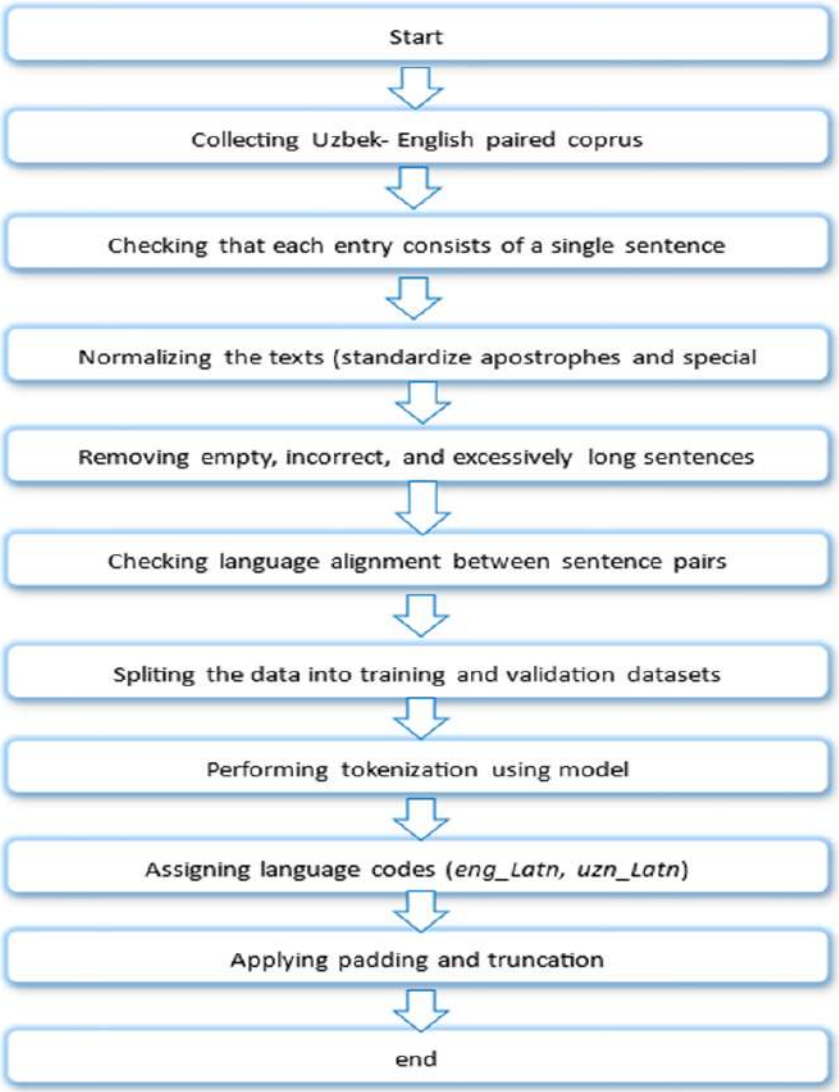

Photo 1 preprocessing model fine-tuning

Experiment results of UzMarine model produced stable and reliable performance across dataset sizes as shown (photo 2) 2000 step it improved progressively from 71% to 74% in 6000 step and in fact the model performance reached a plateau between 8,000 and 10,000 training steps without signs of overfitting. As is shown T5 model produced negative accuracy values (-22% and -4%). Yet, it improved gradually reaching 6%,13% and 18% accuracy. Despite improvement it adduced significant lower than UzMarine translation system. In Helsinki transformation model the result was not low as T5, however comparing UzMarine it indicates 46% which is still law resut.

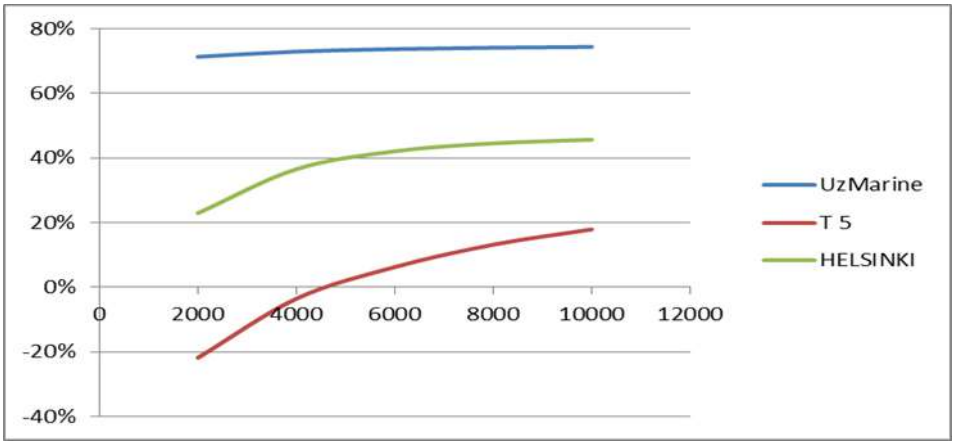

Photo 2. Models results

```
Train dataset size: 114271
Validation dataset size: 28568
/tmp/ipython-input-3781522964.py:147: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for `Seq2SeqTrainer.__init__`. Use `processing_class` instead.
  trainer = Seq2SeqTrainer(
Starting training with NLLB-200 model...
[10713/10713 5:36:08, Epoch 3/3]
```

Step	Training Loss	Validation Loss	Bleu	Meteor
2000	0.313300	0.286979	0.089819	0.280091
4000	0.287000	0.270674	0.110450	0.296190
6000	0.271500	0.263302	0.114834	0.302684
8000	0.258500	0.259068	0.116537	0.307581
10000	0.253000	0.256576	0.120600	0.309901

```
BLEU: 0.0898, METEOR: 0.2801
BLEU: 0.1105, METEOR: 0.2962
/usr/local/lib/python3.12/dist-packages/transformers/modeling_utils.py:3918: UserWarning: Moving the following attributes in the config to the generation config: {'max_length': 200
  warnings.warn(
BLEU: 0.1148, METEOR: 0.3027
BLEU: 0.1165, METEOR: 0.3076
BLEU: 0.1206, METEOR: 0.3099
There were missing keys in the checkpoint model loaded: ['model.encoder.embed_tokens.weight', 'model.decoder.embed_tokens.weight', 'lm_head.weight'].
Training completed!
```

Photo 3. UzMarine model training process

Conclusion

In this study, UzMarine translation system was developed by combining both rule- based and model. Since Uzbek and English are different morphological language types and structure this method was the best way to improve translation system. In the future size of our parallel corpus will be increased and additional linguistic features incorporated.

References

[1] Alawneh, M. F., & Sembok, T. M. (2011). Rule-based and example-based machine translation from English to Arabic. In: Proceedings of the 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), pp. 343–345.

https://doi.org/10.1109/BIC-TA.2011.76

[2] Mengliev, D., Barakhnin, V., & Abdurakhmonova, N. (2021). Development of intellectual web system for morph analyzing of uzbek words. Applied Sciences, 11(19), 9117.

[3] Nazirova, E., Abdurakhmonova, N., & Usmonova, K. (2024). Challenges in corpus linguistics and machine translation. In: Proceedings of the International Scientific-Practical Conference "Contemporary Technologies of Computational Linguistics" (CTCL 2024), Vol. 2, No. 22.04, MyScience Proceedings.

[4] Nagineni, V. S., Idamakanti, K., Bhoopal, N., Kumar, D. G., Rao, D. S. N. M., & Bacha, B.T. (2023). Quality evaluation and assessment of machine translations for English to Turkish (Google Translations vs DeepL Translations). In: Proceedings of the 2023 Global Conference on Information Technologies and Communications (GCITC). https://doi.org/10.1109/GCITC60406.2023.10426024

[5] Nazirova, E., & Usmonova, K. (2024). Algorithm of identifying adjectives from participles in translation from Uzbek to English for machine translation. In: Proceedings of the 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE), pp. 1740–1744.

[6] Nazirova, E., & Usmonova, K. (2024). Algorithmic strategies of complex word analyzing from Uzbek to English for machine translation. AIP Conference Proceedings, 3244, 030054. https://doi.org/10.1063/5.0242104

[7] Nazirova, E. Sh., Abidova, Sh. B., Kxolmuminov, Y. X., Boymurodov, F., Medetbayeva, N. B., & Usmonova, K. A. (2024). Mathematical model and algorithm for grammatical editing of given words in Uzbek language. Springer Proceedings in Mathematics & Statistics, Mathematics for Sustainable Industry (ISMI 2024), Kuala Lumpur, Malaysia.

[8] Tutar, K., & Yıldız, O. T. (2024). Turkish question-answer dataset evaluated with deep learning. In: Proceedings of the 2024 9th International Conference on Computer Science and Engineering (UBMK). https://doi.org/10.1109/UBMK63289.2024.10773453