CREATING A TRANSLATION APPLICATION BASED ON CORPUS.

Moksadbek Sapaev

Masters student in Computational Linguistics, Urgench State University.

Аннотация. *Услуги перевода пользуются большим спросом, чем когда-либо, в современном глобализированном мире, особенно в цифровом секторе. Приложение для перевода на основе корпуса может обеспечить точные, естественно звучащие переводы с быстрым развитием технологий.*

Ключевые. *Параллельный корпус, приложение для перевода, нейронный машинный перевод (NMT).*

Abstract. *Translation services are in greater demand than ever in today's globalized world, particularly in the digital sector. A corpus-based translation application can provide accurate, natural-sounding translations with the rapid advancement of technology.*

**Key words.** *Parallel corpus, translation app, Neural machine translation (NMT)*

1.Introduction.

A machine translation software program uses artificial intelligence to translate texts from one language to another, for example from Uzbek to English. Over time technology has enabled machine translation software to become more accurate, despite the fact that it was initially not very accurate. In this article , we will explore the process of creating a corpus-based translation app, including the steps involved in building a corpus, the techniques used to train the translation models, and the evaluation of the system's performance.

2.Building the Corpus.

A corpus-based translation application requires a corpus of text samples from the source and target languages that contain a wide range of genres and topics.[1] This corpus should be large enough to train the translation models with adequate data. [2] In addition to using online databases, websites, and publications as sources for collecting the corpus, it should be preprocessed to extract meaningful linguistic features, such as part-of-speech tags, syntactic, and structures.

Matching sentences with each others can be done automatically by computer programs, Several such programs are available for free. LF Aligner, Twente-Aligner is one of the above programs and they match texts cross section:
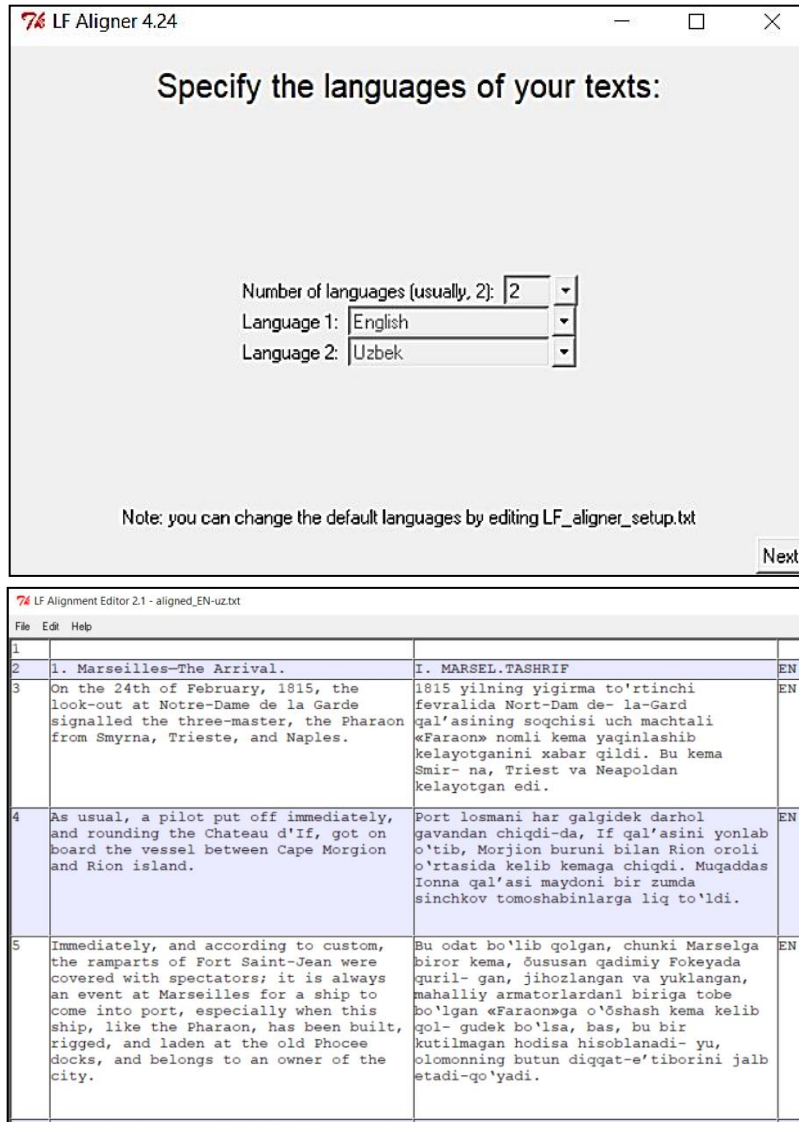
An example of an Uzbek-English parallel corpus created using the LF Aligner program

2. Training the Translation Models.

As soon as the corpus is built and preprocessed, training the translation models takes place. This involves using machine learning techniques to learn patterns and relationships between source and target languages and generating translations based on those patterns. A number of machine translation techniques can be used for training the models, including neural machine translation (NMT) and statistical machine translation (SMT)

A neural network is used in the NMT models to learn a mapping between a source language and a target language by layering the input data.[3] The models can be trained supervised or unsupervised, and they can be optimized using a variety of techniques, such as backpropagation and gradient descent.

3. Evaluating the System's Performance.

In order to create a corpus-based translation app, the last step is to evaluate its accuracy, fluency, effectiveness. It is possible to do this using BLEU, METEOR, and human evaluation scores, among others.[4] By comparing n-grams in the two sets of translations, the BLEU score measures how similar the system's translations are to the reference translations [5]. A METEOR score compares the semantic and syntactic features

of two sets of translations to determine the overall quality of the system's translations. By requesting feedback from human evaluators, the human evaluation score measures the subjective quality of the system's translations[6].

4. Conclusion.

To conclude, building a corpus-based translation app requires several steps, including building a corpus, training the translation models, and evaluating the results. Machine learning and advanced algorithms can provide accurate, natural-sounding translations tailored to the users' specific needs through such apps. As technology continues to improve, we can expect to see even more innovative and effective translation tools in the years to come.

REFERENCES:

[1]   Alexandr Rosen, Martin Vavrin.2012. "Building a multilingual parallel corpus for human users." Conference: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).

[2]   Alberto Simoes, Jose Joao Almeida.2007. "Parallel Corpora based Translation Resources Extraction." Procesamiento del Lenguaje Natural, nº39 (2007), pp. 265-272.

[3]   Joint Training for Neural Machine Translation Models with Monolingual Data March 2018 Proceedings of the AAAI Conference on Artificial Intelligence 32(1) DOI:10.1609/aaai.v32i1.11248.

[4]   Michael Denkowski, Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language

[5]   K. Madatov, S. Bekchanov, and J. Vičič, "Accuracy of the Uzbek stop words detection: a case study on" School corpus"," arXiv preprint arXiv:2209.07053, 2022.

[6]   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.