Algorithm of Creating the “Uzbek-English Aligner” Program

Matyakubova Noila Shakirjanovna
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, computational linguistics and digital technologies
Tashkent, Uzbekistan
nailya89mm@mail.ru

Dauletov Adilbek Yusupbayevich
Tashkent University of Information Technologies named after Muhammad al-Khorazmi, Department of Basics of Informatics
Tashkent, Uzbekistan
davletov--odilbek@mail.ru

Khamroeva Shahlo Mirdjonovna
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi, computational linguistics and digital technologies
Tashkent, Uzbekistan
shaxlo.xamrayeva@navoiy-uni.uz

Eşref Adalı
Computer Eng. and Informatics Fac. Istanbul Technical University
Istanbul – Türkiye
adali@itu.edu.tr

Mengliyev Bakhtiyor Rajabovich
Tashkent State University of Uzbek Language and Literature named Alisher Navoi, Department of Applied linguistics
Tashkent, Uzbekistan
bakhtiyormengliyev@gmail.com

***Abstract* - Today, the volume of information related to any sphere of human life is increasing. Although it is easy to find the information you need, there are several problems that need to be solved in their use, to overcome linguistic barriers, to help understand information in different languages, and to improve the performance of translation tools. There is a great need for tools that matter. This article provides detailed information about the Aligner software tool aimed at solving the above-mentioned problems, its role and tasks in natural language processing, as well as the stages of developing the algorithm of the "Uzbek-English Aligner" program.**

Keywords: Aligner, alignment, corpus, natural language processing, machine translator, algorithm, source language, target language, token, lemma.

I. INTRODUCTION

Technological progress is one of the main factors behind the rapid development of almost all industries, allowing representatives of any industry to find the necessary information quickly and easily as well as follow the latest news in their field around the world. At the same time with the expansion of the possibilities, people are faced with barriers to understand the language in the assimilation of existing information and their equal use. Although there are automatic translators to solve such problems, they can translate the source text only the limited number of languages and it is the main reason for further development of machine translators.

Today, there are several effective machine translators such as Google Translate, Yandex Translate, DeepL, and the number of languages that can be translated is slightly more than other MTs. Despite the high efficiency and the large number of languages, in some languages there are still some shortcomings in the order, meaning and content of the translated sentences, and in the grammatical structure also. Several "aligner" software tools are being developed by computational linguistics specialists to overcome them.

In addition to ensuring the efficient operation of MTs, [20] Aligner software tools are of great importance in creating linguistic resources such as parallel corpora, multilingual dictionaries, syntactic or semantic databases. It helps to facilitate and speed up the process of conducting research in various fields of computer linguistics by matching linguistic units such as words, phrases, and syntactic structures.

A. Work Stages of Machine Translators and The Role of Aligner in Them.

Alignment in machine translation is the process of establishing correspondences between words or phrases in the source language and their translations in the target language. Alignment is critical for training and improving machine translation models. Below we will consider the steps of the process of translating a sentence using machine translation.

1) Word processing: An input sentence undergoes a preprocessing step such as tokenization, where the sentence is broken down into individual words or phrases. This step helps create a structured representation of the sentence that can be processed by the MT system.

2) Understanding the language: The MT system analyzes the original sentence to understand its linguistic structure, grammar and meaning. This stage includes syntactic analysis, semantic analysis and extraction of relevant features from the source sentence. The system aims to capture the purpose and context of the sentence.

3) Alignment: To match source and target languages, MT system can use aligner software tools to match matching words or phrases between source and target languages. These software tools help create a foundation for the translation and ensure that the translations accurately convey the meaning of the original sentence.

4) Translation model: The translation model is the main component of the MT system. It uses statistical or neural machine translation techniques to generate the translated sentence based on the data from the match and the understanding of the original sentence. The translation model is trained on large parallel corpora of matched sentences in the source and target languages.

5) Post-processing: The translated sentence generated by the translation model can undergo further processing steps to improve its quality and readability. This includes changing word order, correcting grammar or syntax, and manipulating certain linguistic or stylistic rules in the target language. Post-

processing helps improve the result of the translation and makes it more natural and fluent.

6) Evaluation and feedback: It is important to evaluate the quality of the translated sentence. Various metrics and evaluation methods, such as BLEU (Bilingual Evaluation Understudy), are used to measure the similarity between the translated sentence and human-generated reference translations. The evaluation provides feedback on the performance of the MT system, which can be used to improve the translation models.

7) Iterative improvement: MT systems are often trained and improved iteratively. Ratings and user input feedback are used to update and improve translation models. This process helps the system learn from mistakes, eliminate translation errors, and improve overall translation quality.

We summarize these processes in the Fig. 1 below:

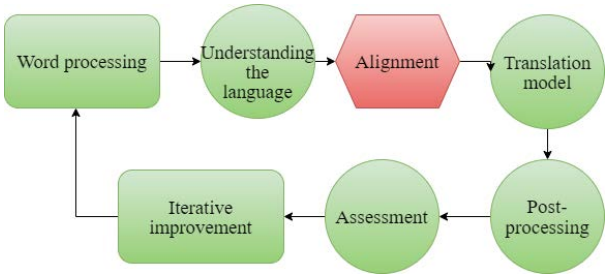

Fig. 1. Stages of the translation process in MT

It should be noted that there may be differences in the specific steps and methods of different MT systems, but the above steps provide an overview of the process of sentence translation using MT.

II. CONDUCTED SCIENTIFIC RESEARCH

Alignment has been an important concept of NLP since the second half of the 20th century. The history of aligners goes back to the early days of MT: development of rule-based alignment systems began in the 1960s.

The Gale and Church algorithm [5], also known as the Hungarian algorithm or Munkres algorithm, is an optimization algorithm used to solve an assignment problem. It was developed by D. Gale and Lloyd S. Shapley in 1960. Although the Gale and Church algorithm was not specifically designed to match texts given in natural language, it can align words or phrases, sentences or documents based on similarity scores, and it is also used to solve other problems in different fields. Therefore, it can be used in aligning objects in different contexts.

As statistical techniques became more popular in NLP in the 1980s, [19] researchers began to develop statistical matching models using probabilities to estimate the probability that a given word/phrase in one language would match a given word or phrase in another language. One of the first statistical alignment models was the IBM Model 1, which used a simple, uniform probability distribution to align words in parallel texts.

In the 1990s, researchers developed more sophisticated statistical matching models, such as the IBM Model 2, which used a more complex hierarchical model to match words and phrases in parallel texts. These models solve relatively complex adaptation problems, such as dealing with non-literal translations, and have been used in various machine translation systems.

Vanilla Aligner [2] was introduced in 1997 by Pernilla Danielson and Daniel Ridings as an improvement of the Gale and Church algorithm. It is considered a "sentence aligner" tool and depends on sentence boundaries. The main advantage of this aligner is compatibility with bitexts in SGML format. One of the advantages of using bitexts in SGML is that a standard form or structure can be established and sentence boundaries can be more easily defined.

With the emergence of the phrase-based machine translation paradigm in the 2000s, researchers developed alignment models specifically designed to match phrases in parallel texts [3]. One of the most popular of these models is the GIZA++ aligner, which used the IBM Model 4 variant to align phrases in parallel texts.

Bleualign, a sentence aligner software tool, was created by Rico Sennrich and Martin Volk [13]. Its main idea is to use the MT framework and the translation estimator BLEU to assist in the alignment process [13]. To better understand the aligner, it is necessary to know about BLEU. BLEU is an algorithm that evaluates the quality of MT translation results. To measure this quality, BLEU MT is evaluated by comparing the translation result with one or more human translations.

Intrative Bleualign, a sentence aligner algorithm, was created in 2011 by Rico Sennrich and Martin Volk [13] as a result of a deeper analysis of the shortcomings of the use of machine translation systems in the alignment process. Sennrich and Volk found that MT-based aligners are strongly dependent on the correct translation of the source text, and given that MT systems are typically fed aligned texts, the existence of a circular dependency is evident. To overcome this dependency, the algorithm provides a bootstrapping method to perform the alignment.

The creation and implementation of an aligner tool suitable for the effective operation of machine translation, natural language processing, information retrieval, and parallel text processing tools is of interest to many field researchers today. In particular, the creation of the software tool "Uzbek-English aligner" is an urgent issue for researchers in today's field, which is mainly to eliminate the shortcomings of existing Uzbek-English (bilingual) translation tools, text analyzers, and programs that work with parallel texts and for their further improvement. Creating the algorithm of the "Uzbek-English aligner program" usually includes several steps. We offer them as follows:

1) Gather information: A parallel corpus of Uzbek and English sentences is collected. This collection should have an English translation corresponding to each Uzbek sentence and should consist of sentences that are aligned with each other.

2) Preprocessing: Sentence cleaning and pre-processing is required to remove unnecessary units, punctuation, and any irrelevant information. Tokenization is used to break sentences into individual words or phrases.

3) Matching words: A "word matching operation" is performed to match the Uzbek and English words in the corresponding sentences. Different approaches can be used to match words, such as statistical matching models like IBM Model 1 or IBM Model 4 or HMM-based model. The

corresponding models study the probability distribution of word matching between the source (Uzbek) and translation (English) languages.

4) Phrase matching: Phrase matching is performed to match longer phrases or phrases between Uzbek and English sentences. Phrase matching can be done using statistical models such as IBM Models or advanced techniques such as statistical machine translation models.

5) Improvement of the aligner: Post-processing methods are used to improve alignment: alignment errors or inconsistencies are addressed. This step may involve manual alignment or the use of alignment quality metrics.

6) Matching the result: Corresponding sentence pairs are stored in a parallel corpus format or an alignment-friendly format. This result can be used in various fields such as machine translation, natural language processing or language learning for their various applications.

We summarize these processes in the Fig. 2 below:

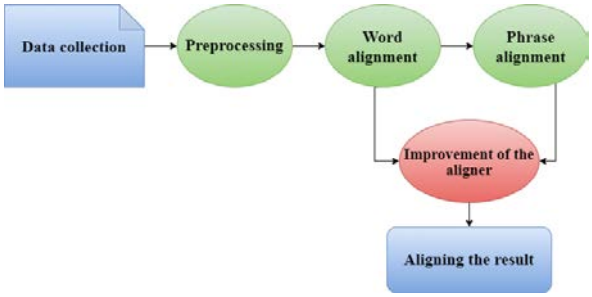

Fig. 2. Steps of creating the algorithm of the "Uzbek-English aligner program"

The accuracy and quality of the Aligner program largely depends on the quality and size of the parallel corpus used for training. The larger and more varied is the corpus, the better the aligner will work. In addition, the specific implementation details and algorithms used for alignment may vary depending on available resources and expertise.

III. ALGORITHM OF DEVELOPMENT OF "UZBEK-ENGLISH ALIGNER PROGRAM".

A. First step: collecting texts for the corpus.

As we have seen above, when creating the "Uzbek-English aligner program", a parallel corpus of Uzbek and English sentences is first collected. Today, the fact that there is no parallel corpus in Uzbek and English, and the limited number of texts with translations in two languages makes this stage somewhat difficult and takes more time. We use the primary Uzbek-English corpus in the aligner software tool that we plan to develop. It mainly consists of the original English text and its Uzbek translations, as well as the original Uzbek text and its English translations, in a word, it consists of a collection of English/Uzbek and Uzbek/English translations.

TABLE I. UZBEK-ENGLISH CORPUS BLOCKS

Uzbek-English corpus blocks			
English →	Uzbek	Uzbek →	English

The used information is taken from textbooks, various works, epics or short stories, legal documents and newspaper articles. It is important that these sources have been translated or checked by experienced translators. The text collected for corpus is given in Fig.3.

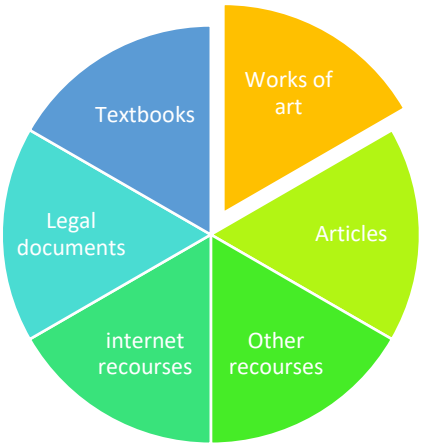

Fig. 3. The text collected for corpus

In this article, we will use several of Abdulla Qadiri's and Utkir Hashimov's bitext stories as a source corpus for programming and examine their alignment process.

B. Second step: "Uzbek-English aligner program" operation stage.

Step 1: the bilingual text is divided into sentences based on punctuation marks and sentence length. In this case, the sentences given in the original language are separated on the basis of conditional signs such as S1, S2, ...S10; Sentences in the translated language are separated by symbols S1, S2, ...S10.

Stage 2: it is verified that the given text is correctly divided into sentences based on human intervention. In this case, the texts in both languages based on S1, S2, ...S10 and similar symbols given in both languages are compared and it is determined whether they are correctly/incorrectly separated.

Step 3: the degree of compatibility of the compared sentences is assessed. We can evaluate the assessment stage by human intervention or by Batch Processing method. Considering that the parallel corpora in Uzbek and English are not fully formed, a human intervention-based evaluation system will be more effective [11].

Step 4: In the evaluation step, the differences between the divided sentences of the source text and the translated text are studied. After determining the proportion of the given texts, the reason for quantitative differences between them are studied and analyzed.

TABLE II. PROPORTION OF THE UZBEK-ENGLISH PARALLEL TEXTS

Text in Uzbek language	Text in English language
Лекин онамнинг оёқ оғриғи фақат кексаликдан эмас. Буни бошқалар билмаса ҳам, мен биламан. Яхши биламан.	But I know well how she got that pain.

Step 5: changes in the structure of the sentence(s) are studied. It studies and analyzes the phenomena arising from the fact that the Uzbek and English languages are languages with different systems (English belongs to the Roman-Germanic language family, Uzbek belongs to the Turkish language family). These languages have different syntactic structures. Therefore, there are differences in the translation. For example, three S1, S2, S3 in the source language given in Table I became only S1 when translated, and the main reason for this is that the simple sentences given in the source language became compound sentences in the translated language.

TABLE III. QUANTITATIVE DIFFERENCE IN SENTENCES OF THE TEXT

S1	Лекин онамнинг оёқ оғриғи фақат кексаликдан эмас.	S1	But I know well how she got that pain.
S2	Буни бошқалар билмаса ҳам, мен биламан.	S2	
S3	Яхши биламан.	S3	

Step 6: sentences are divided into tokens. The tokenization process is important mainly for matching the given words in sentences. It determines the number of tokens in the matching sentences. If there is a difference between them, the reason is investigated.

Step 7: the resulting differences are analyzed based on the grammatical and semantic structure of the language. At this stage, the lemmatization process is carried out: it helps to determine the main reason for the increase or decrease in the number of words in the original and translated languages during the translation process. For example, if there are five tokens in the sentence **“Dilshod tez orada qaytib keldi”** in Uzbek, when translated into English, **“Dilshod returned soon”** will become only three tokens. The reason for this situation is that tokens “**tez orada**” in the Uzbek language are considered as one lemma and correspond to the "Soon" token in the English language. The matching of the lemma “**qaytib keldi**” with the token **“returned”** also creates a difference in the number of tokens. Therefore, lemming sentences divided into tokens ensures a fast and efficient process of finding matches in sentences.

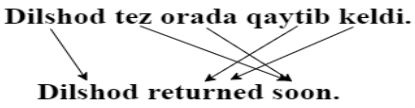

Fig. 4. Alignment process

Step 8: in the word alignment step, the quantitative aspect of the words is re-compared after the lemma step. In this, sentences given in both languages are separated and compared based on symbols such as W1 W2 W3.

TABLE IV. MATCHING WORDS BASED ON SYMBOLS.

W_1	there	W_6	Kavkaz
W_2	were	W_4	Tomonlarda
W_3	many	W_3	Ko‘p
W_4	in	W_1	Bo ‘ladi
W_5	the		
W_6	Caucuses		

Words that do not exist in the grammatical structure of the translated language are separated separately. For example, in Uzbek grammar, there are no word groups such as articles, prepositions, and auxiliary verbs, but they are taken into account in the translation process in order to ensure the grammatical correctness of the language being translated. We will consider this situation in the above given Table IV in the analysis of one sentence taken from Utkir Hashimov's story "Socks".

As can be seen from the table, the number of words in English is six, while in Uzbek they are reduced to four, the main reason for this is that some forms that do not exist in the grammatical structure of the Uzbek language are used in the English sentence.

Step 9: Identify the equivalence of words using a dictionary. If an equivalent is found, the step is terminated, if no equivalent is found, the reason is determined. In the case of grammatical inequivalence, the stage is terminated, in the case of lexical inequivalence, realia or lacuna dictionaries are referred to [11]. We summarize these processes in the Fig 4 below

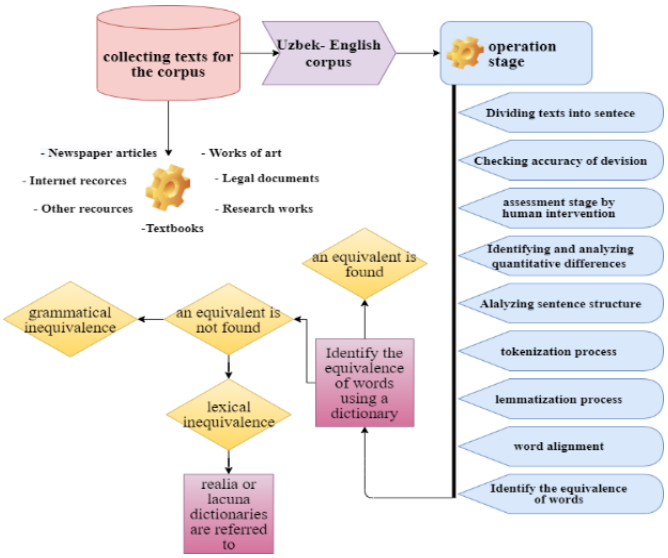

Fig. 4. Algorithm of development of "Uzbek-English aligner program"

Although the English vocabulary is 1,022,000 [18] according to research conducted in 2010, today the number of words that have a translation in Uzbek is 70,000. For this purpose, this step is very important to determine whether the translation of some words encountered during the matching process is available or not in the vocabulary we have.

IV. CONCLUSION

In this article, the important factors for the creation of the "Uzbek-English aligner program" and some of the stages of its creation have been mentioned. The non-existence of the Uzbek-English corpus and the limited resources that can be used as the primary form of the corpus were noted as the most important and problematic part of the creation stages. Along with the full development and improvement of the steps of the Aligner software tool, the formation of the corpus is equally important, and we have taken into account the solution of this issue in our research. Because the size and diversity of the corpus is one of the most important factors that increase the effectiveness of the alignment tools.

Although not all stages of the "Uzbek-English aligner program" have been fully developed, it is intended to improve the quality and efficiency of several areas of Uzbek computer linguistics, including machine translators and parallel text processing tools

REFERENCES

[1]. P.F.Brown, and J.C.Lai, R.L. “Mercer Aligning sentences in parallel corpora”. Proceedings of the 29th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics. 1991. – p. 169-176.

[2] P.Danielsson, and D. Ridings “Practical presentation of a “vanilla” aligner”. TELRI Workshop in alignment and exploitation of texts. 1997.

[3] S.M.Abdul-Rauf, P.Fishel, S.Lambert, and R.Sennrich.” Evaluation of Sentence Alignment Systems”, Project at the Fifth Machine Translation Marathon. 2010.

[4] P.Fung, and K.McKeown, Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. – 1994.

[5] W.A.Gale, and K.W.Church, “A program for aligning sentences in bilingual corpora”. Computational linguistics, 19(1) – pp. 75-102.1993

[6] A.Gelbukh, G.Sidorov, and J.Á.Vera-Félix, “A bilingual corpus of novels aligned at paragraph level”, Advances in Natural Language Processing. Springer Berlin Heidelberg, pp. 16-23. 2006.

[7] V.Aleksic, and G.Thurmair. Rule-based MT system adjusted for narrow domain (ACCURAT Deliverable D4.4.). Technical report. – 2012.

[8] C.Kit, J.J.Webster, K.K.Sin, H.Pan, and H.Li, ” Clause alignment for Hong Kong legal texts: A lexical-based approach”, International Journal of Corpus Linguistics, – p. 29-51, 2004.

[9] A.McEnery, and R.Z.Xiao, ” Paralell and comparable corpora: what are they up to?”, Incorporating Corpora: Translation and the Linguist. Translating Europe. Clevendon: Multilingual Matters, 2008.

[10] E.Macklovitch, M.L.Hannan “Line ‘em up: advances in alignment technology and their impact on translation support tools.” Machine Translation, 13(1). pp. 41-57. 1998.

[11] A.Meyers, M.Kosaka, and R.Grishman, “A multilingual procedure for dictionary-based sentence alignment”, Conference of the Association for Machine Translation in the Americas. Springer Berlin Heidelberg, pp 187-198. October, 1998.

[12] R.C. Moore, “ Association-based bilingual word alignment”, Proceedings of the ACL Workshop on Building and Using Parallel Texts. Association for Computational Linguistics. 2005.

[13] R.Sennrich, and M.Volk ,“MT-based sentence alignment for OCR generated parallel texts”, The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado. 2010.

[14] R.Sennrich, and M.Volk. Iterative, “MT-based sentence alignment of parallel texts”, 18th Nordic Conference of Computational Linguistics, NODALIDA 2011.

[15] A.Simões,” Parallel corpora word alignment and applications (master thesis)”, Universidade do Minho, Braga. 2004.

[16] Natural Language Toolkit (NLTK): https://www.nltk.org/

[17] British National Corpus (BNC): https://www.natcorp.ox.ac.uk/

[18] https://englishlive.ef.com/blog/language-lab/many-words-english-language/

[19] T. Nguyen and T. Nguyen, “Heavyweight Statistical Alignment to Guide Neural Translation” , Hindawi, Computational Intelligence and Neuroscience, 2022.

[20] Z. Zong, and C. Hong,”Research on Alignment in the Construction of Parallel Corpus”, Journal of Physics: Conf. Series 1213. 2019.