IEEE
TÜRKİYE SECTION

UBMK'24

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

9th International Conference on Computer Science and Engineering

26-27-28 Ekim (October) 2024 Antalya - Türkiye

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2024)

9th International Conference on Computer Science and Engineering

26-28 Ekim 2024 Akdeniz Üniversitesi Antalya Türkiye
26-28 October 2024 Akdeniz University Antalya Türkiye

IEEE Catalog Number : CFP24L97-CDR

ISBN : 979-8-3503-6587-0

Additional copies may be ordered from:
Curran Associates, Inc
57 Morehouse Lane Red Hook, NY 12571 USA
Phone: (845) 758 0400
Fax: (845) 758 2633
E-mail: curran@proceeding.com

Algorithm for Aligning Paragraphs and Sentences in Aligner Tool

Elov Botir Boltayevich
Tashkent State University of Uzbek Language and Literature named Alisher Navoi,
Tashkent, Uzbekistan
elov@navoiy-uni.uz

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

Dauletov Adilbek Yusupbayevich
Alfraganus University,
Tashkent, Uzbekistan
davletov--odilbek@mail.ru

Matyakubova Noila Shakirjanovna
Tashkent State Univ. of Uzbek Language and Literature named after Alisher Navoi,
Tashkent, Uzbekistan
matyakubovanoila@navoiy -uni.uz

***Abstract*— A parallel corpus is one of the main resources for training and evaluating machine translation systems. By adapting parallel texts, it is possible to improve the translation quality of machine translators, which allow people to use different languages freely. In addition, parallel corpora play an important role in the efficiency of natural language processing tasks such as searching engines, sentiment analysis, and object recognition. There are several stages in the formation of such corpora, one of them is the alignment process. Once the parallel texts are collected, they need to be aligned at the paragraph, sentence, word or phrase level in order to determine the correspondence between segments in different languages. Today, several Aligner tools are available for these tasks, automating this process by aligning and identifying translation equivalents based on neural or statistical models. But not all available tools are equally effective in different languages. This article provides information about the linguistic and software support of the Uzbek-English "Aligner" system, which aligns parallel texts in Uzbek and English, and the stages of its creation.**

Keywords— Parallel corpus, parallel texts, alignment, Aligner, segmentation, source language, target language.

I. INTRODUCTION

Parallel corpora are a rich source of linguistic information that has far-reaching implications for research, education, and technology development. Sound methodologies are essential for creation and widespread use of parallel corpora because with such corpora we can deepen our understanding of language diversity, advance intercultural communication, and open up new possibilities in the fields ranging from computational linguistics to cross-cultural studies. One of the main steps to ensure that corpora are properly formed is the alignment process, and Aligners are the only tools that correctly distribute texts in parallel. Today there are several aligners and the most commonly used ones are sentence aligner, word aligner and phrase aligner.

II. CONDUCTED SCIENTIFIC RESEARCH

Like many natural language processing tools, alignment tools have gone through several stages of development. There are single-function aligners and hybrid aligners available today, each with its own advantages and disadvantages. **GIZA++,** developed at the University of Aachen in 1999, is a statistical machine translation toolkit that includes word matching tools between parallel corpora[1]. It is widely used in machine translation and natural language processing. **HunAlign,** created at the Budapest University of Technology and Economics in 2002, is a sentence-level aligner that uses heuristics and statistical methods to align parallel corpora. It has been widely used in various machine translation and corpus linguistics projects[2].

Berkeley Aligner [3] was developed by The Berkeley NLP group in 2010 and it is a word aligner based on IBM Model 1 and IBM Model 2. It provides aligning models for different language pairs and mainly used in machine translation research. UCambridge Aligner, developed in 2011, is a tool used to align parallel corpora at the sentence level[4]. It uses a Bayesian model to estimate matching probabilities and is used in research on machine translation and language modeling.

In addition, MGIZA++ (Multidisciplinary GIZA++), which was developed in 2012, provides versatile capabilities to speed up the alignment process, which allows it to be used for large-scale parallel corpora[5]. Fast_align, created in 2014, is an open source word alignment tool[6]. It stands out due to its speed and accuracy in matching large-scale bilingual corpora. In 2017, researchers at Facebook AI Research created MUSE, a toolkit designed for multilingual unsupervised and supervised word alignment[7]. Although it is not a dedicated alignment tool, it includes features for aligning words across languages, which in turn allows for cross-language analysis and transfer studies.

III. UZBEK-ENGLISH "ALIGNER" SYSTEM

Although aligners are used in various areas of NLP, their main task is to match text segments given in the source language(SL) to text segments in the target language(TL) [8]. Aligners are selected depending on what the matching object is. The aligner that we have created is mainly designed for aligning Uzbek-English parallel corpus, and allows to align the parallel corpus in the following stages:

- Paragraph alignment
- Sentence alignment

IV. ALIGNER FORMATION STAGE.

As an input a lexical dataset of Uzbek and English words, parallel texts in Uzbek and English languages which is formed as a parallel corps are used. Creation of the corpus is carried out in several stages:

- First of all, the original sources of Uzbek and English texts are compiled. The Resource guide "Preparing World Heritage Nominations" (Second edition, 2011)Published in November 2011 by the United

Identify applicable funding agency here. If none, delete this text box.

Nations Educational, Scientific and Cultural Organization is taken as a research object [9]. Texts were extracted from it and a corpus of more than 1 million sentences was formed.

- Collected texts for the corpus are also processed in several stages:

1) the text is cleaned of excess noise;
2) if there are abbreviations in the English text, they are identified and rewritten in their full form. For example, the word “**it's”** in the given English text is a contraction of the pronoun it and the verb “**is”** or “**has**”. If we leave them unseparated, these two separate meaning words will be treated as a single token during the tokenization process, leading to large errors in the matching process.

- Processed texts are divided into small segments (sentence form).
- The number of allocated segments is calculated. This process is necessary to know the exact size of aligned texts in the SL and TL, in order to determine whether the number of aligning sentences is the same or how much they differ. Because our main goal is to determine whether there is an identical translation of the SL text in the TL and to highlight the appropriate translation.

If the difference in segments is not so big, we can use as input the texts that have passed all the steps given in the Fig.1.

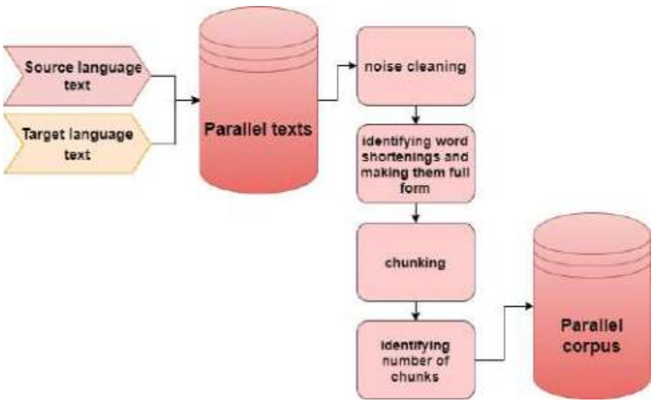

Fig. 1. Stage of preparation of input materials.

B. Paragraph Alignment.

In order to align paragraphs the following steps should be carried:

- Segmentation process: At this stage of alignment, the paragraphs which is going to be aligned are first seperated into sentence.
- Counting number of sentence: The number of sentences in both languages is determined. If the number of sentences is the same, the alignment process goes on much easier, but not always the translated sentences of the SL are not at the same number and same form in TL. There are several reasons for this, sometimes it can be technical error while translating the sentence or othe cases grammatical structure of the TL sentence. To solve grammatical problems, we first studied the sentence structures of the source and target languages (see Table I).

For example, When I got on the coach, the driver had not taken his seat, and I saw him talking to the police. // Avtobusga chiqqanimda haydovchining hali o`z joyini egallamaganini, politsiya xodimi bilan gaplashayotganini ko‘rdim. The example given in English is the compound-complex sentence, which consists of a compound sentence with an adverbial clause and two main clauses. When translated into Uzbek, it becomes a simple extended sentence.

TABLE I. TYPES OF SENTENCES IN UZBEK AND ENGLISH.

Types of the sentence	The simple sentence	The compound sentence	The complex sentence	The compound-complex sentence	The composite sentence
English language	+	+	+	+	+
Uzbek language	+	+	+	-	-

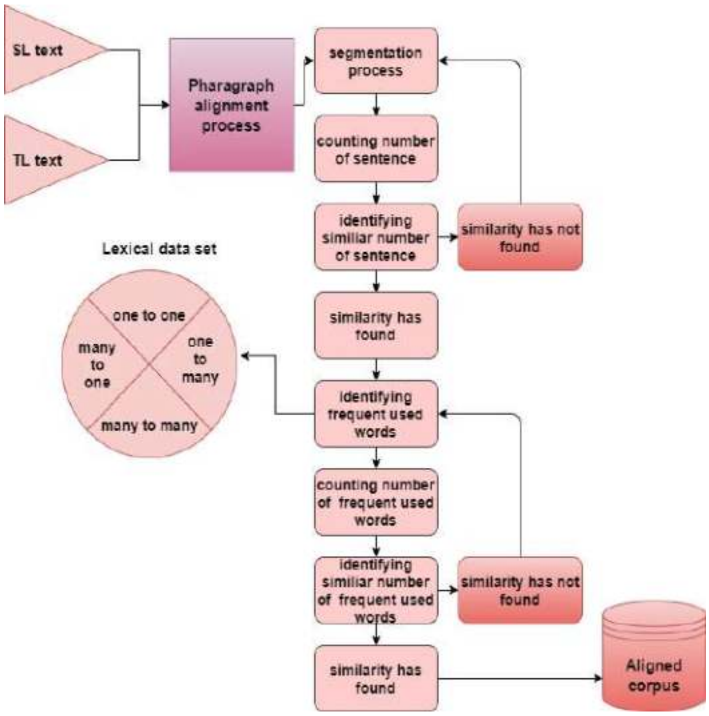

Fig. 2. Paragraph alignment algorithm

Or, One night, as soon as I finished my work at home, I went to get some vegetables from the market. // Bir kuni tunda uydagi ishlarimni yakunlashim bilanoq bozordan biroz sabzavotolishga olishga ketdim. In this example, a compound sentence with a time clause in English is translated into Uzbek as a simple expanded sentence.

During the research, several grammatical forms that do not exist in the Uzbek language or causes of the structural changes of the sentences during the translation process were identified, their linguistic base was formed. We will cover the complete information about this linguistic base in our further articles.

- Identifying frequent used words: At this stage the most frequently used word in SL and their translation in TL are determined. If the number of detected words and their translations are the same or if the degree of proximity is high, it is taken as an aligned pair. (see: Fig 2.).

In the example given in Table II (see: Table II), the word **nomination** is used six times in the SL and the same in the TL. In addition, the number of matching paragraphs is seven in both languages. Therefore, it is possible to accept a paragraph in the TL as corresponding to a paragraph given in the SL.

TABLE II. STEP TO IDENTIFY THE MOST FREQUENTLY USED WORDS.

Source language	Target language
Lack of preparation time is the biggest enemy of successful nominations. Far too many are prepared against unrealistically short timeframes. It can take at least a year to set up appropriate support mechanisms and gather material, and a further year to write the nomination text and consult stakeholders. When research is needed, protection has to be achieved, and new management systems put in place and documented, so the process might take much longer. If the aim is a successful nomination that leads to inscription on the World Heritage List and long-term conservation and presentation of the property, a realistic timeframe should be allowed. Too often, lack of adequate preparation time leads to deferred or referred nominations, which is frustrating for States Parties, the World Heritage Committee and the Advisory Bodies. Sometimes political commitments are made which set an unrealistic timeframe for preparing a nomination, resulting in a nomination dossier which is inadequate and not ready for evaluation.	Tayyorgarlik ko`rish uchun vaqtning yetishmasligi, nomzodlarni muvaffaqiyatli taqdim etishning eng katta dushmani hisoblanadi. Juda qisqa muddat ichida haddan ziyod ko'p nomzodlar tayyorlanadi. Tegishli qo'llab-quvvatlash mexanizmlarini o'rnatish va ma'lumot to'plash uchun kamida bir yil, nomzodlik matnini yozish va manfaatdor tomonlar bilan maslahatlashish uchun yana bir yil kerak bo'ladi. Tadqiqot o'tkazish kerak bo'lganda, yangi boshqaruv tizimlarini himoya qilish, joriy etish, ularni hujjatlashtirish lozim, shuning uchun bu jarayonga ancha uzoq vaqt ketishi mumkin. Agarda maqsad – obyektni Butunjahon merosi ro'yxatiga kiritilishi va obyektni uzoq muddatli muhofaza qilish va uning taqdimotiga olib keluvchi muvaffaqiyatli nomzodlik bo'lsa, buning uchun real muddatlar belgilanishi kerak. Aksariyat hollarda yetarli tayyorgarlik uchun vaqtining yo'qligi nomzodlikka qo'yishning kechiktirilishi yoki qayta ko'rib chiqishga berilishiga olib keladi, bu esa Ishtirokchi-davlatlar, Butunjahon merosi qo'mitasi va Maslahat organlarining umidlarini puchga chiqaradi. Ba'zan nomzodni tayyorlash uchun noreal muddatlarni belgilaydigan siyosiy majburiyatlar olinadi, bu natijada nomuvofiq va baholashga tayyor bo'lmagan nomzodliklarning paydo bo'lishiga olib keladi.

C. Sentence alignment

The sentence alignment process is somewhat more complex than the paragraph alignment process and involves several analytical processes. We will consider them in the order given in the Fig.3 (see Fig.3).

- At the first step the given text is divided into segments in the form of sentences.
- At the next step the length of sentences is determined by counting the number of words in a given sentence. For example: (See Table III.).
- After the length of the sentences of both languages have been identified, the sentence with similar or very close length are determined.

TABLE III. THE STAGE OF DETERMINING THE LENGTH OF SENTENCES.

Sometimes political commitments are made which set an unrealistic timeframe for preparing a nomination, resulting in a nomination dossier which is inadequate and not ready for evaluation	Ba'zan nomzodni tayyorlash uchun noreal muddatlarni belgilaydigan siyosiy majburiyatlar olinadi, bu natijada nomuvofiq va baholashga tayyor bo'lmagan nomzodliklarning paydo bo'lishiga olib keladi.

There are 27 words in the sentence given in English, and 22 words in the sentence with the same translation in Uzbek. The main reason for the difference between the tokens is the non-use of grammatical forms such as articles, prepositions and auxiliary verbs, which do not exist in Uzbek. During the

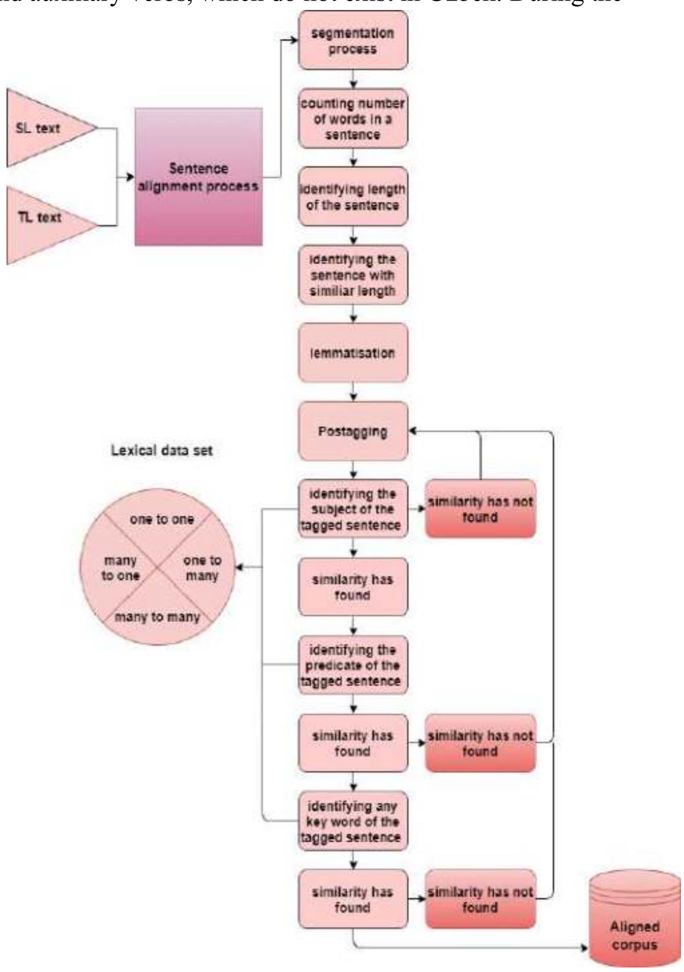

Fig. 3. Sentence alignment algorithm

research process special database of such grammatical forms was formed for Aligner program.

- Before POS Tagging process lemmas of the tokens should be identified[10]. It is very important process in alignment especially when the source language is Uzbek. If the lemma of the word is not identified, some complications arise in the process of POS tagging.

As in the Uzbek language there are sentences with a hidden subject, in which the subject of the sentence can be understood from the suffixes added to the predicate, but in

English, the subject of the sentence must be given separately in order to form the sentence correctly. For example, in the sentence “**Topshiriqlarni tugatdim**”, “topshiriqlarni” is the object, “tugatdim”- is the predicate, subject of the sentence is not given separately on the sentence. But the suffix - **im** added to the verb “tugatmoq (to finish)” indicates that the action was performed by the pronoun “I”, first person singular. So, in English, this sentence is translated as “**I have finished tasks**” using the pronoun “I”, where “I” is the subject of the sentence, “have finished” is the predicate of the sentence, and "tasks" is the object. A database of possessive clauses in such sentences was collected and trained to the Aligner program.

- The subject of the given sentences is determined and matched, if the match is correct, the match is accepted. If the match is not found lemmatization process repeats one more time.
- The predicate of the sentence is identified and compared. At this stage the selected lemma is searched in the dictionary and a match is determined.

This is also a somewhat complicated process, because the structure of the predicate is completely different in both languages. There are simple and complex predicates in the Uzbek language, as well as in English, but their structure is different. Auxiliary verbs that form the complex predicate in English do not exist in Uzbek. Therefore, when translated into Uzbek, the verbs formed in the complex predicate in English becomes the simple one in Uzbek, or complex predicate formed by two independent verbs in Uzbek language is usually translated with a single verb and become simple predicate in English. For example, the complex predicate in English “have been woking” turns into the simple predicate in Uzbek“ishlayotgandi”, or vice versa, the complex predicate “qaytib keldi” in Uzbek becomes the simple predicate in English, “visited”. In order to solve such problems, a lexical data set in the form of “one to one”, “one to many”, “many to one”, “many to one” was formed.

- In the next step, any active word in the tagged sentence in SL is determined and compared with the sentences in the TL. If a match is found, the step ends, if not, the word is compared with another synonym form in the lexical database. If a synonym form is found, the stage ends, if not, another active word is selected and the stage is repeated (see: Fig 3).

IV. CONCLUSION

Aligner tools are the most useful and effective tools for working on parallel texts and determining whether the translation of the source language text into the target language with one-to-one correspondence. Although there are many effective aligners available today, they do not perform equally well and accurately in all languages. This situation is especially common when adapting the translation of texts from languages belonging to different families. When determining their compatibility, if the specially designed aligners for those languages are used, the efficiency indicator will be significantly higher.

A database of many grammatical and lexical rules, which can be separated only by human intervention during the translation process, in Uzbek and English languages has been collected for creation of the Uzbek-English Aligner software and trained. In the future, this software is expected to be used not only for aligning texts in parallel corpus, but also for searching engines and translation tools, and will show effective results.

REFERENCES

[1] F. Och, H. Ney. "Improved Statistical Alignment Models". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October 2000.

[2] F. Och, H. Ney. “A Systematic Comparison of Various Statistical Alignment Models”. Association for Computational Linguistics,2023.

[3] A. Pauls, D. Klein, D. Chiang, K. Knight “Unsupervised Syntactic Alignment with Inversion Transduction Grammars” Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA.

[4] J. Sánchez “A comparative study of Neural Machine Translation frameworks for the automatic translation of open data resources”, Escola Tècnica Superior d’Enginyeria Informàtica Universitat Politècnica de València,2018.

[5] M. Junczys-Dowmunt, A. Szal, “SyMGiza++: Symmetrized Word Alignment Models for Statistical Machine Translation”, International Joint Confererence, SIIS 2011, Warsaw, Poland, June 13-14, 2011

[6] Ch. Dyer, V. Chahuneau, N. Smith, “A Simple, Fast, and Effective Reparameterization of IBM Model 2”, North American Chapter of the Association for Computational Linguistics, 1 June 2013

[7] N. Robinson, N. Carlson, D. Mortensen, E. Vargas, Th. Fackrell, N. Fulda, “Task-dependent Optimal Weight Combinations for Static Embeddings”, Northern European Journal of Language Technology. November 2022

[8] N. Matyakubova, A. Dauletov, Sh. Khamroyeva, B. Mengliyev, E. Adali, “Algorithm of Creating The “Uzbek-English Aligner” Program”, 2023 8th International Conference on Computer Science and Engineering UBMK 2023. Mehmet Akif Ersoy University, Burdur – Turkey.

[9] Preparing World Heritage Nominations. Published in November 2011 by the United Nations Educational, Scientific and Cultural Organization. Second edition, 2011.

[10] Sh. Sirojiddinov, B. Elov, Sh. Khamroeva, E. Adalı, Z. Xusainova. “Pos Taging of Uzbek Text Using Hidden Markov Mode” 8 th International Conference on Computer Science and Engineering UBMK 2023, Mehmet Akif Ersoy University, Burdur – Turkey.