The paper presents the two-stage alignment and extending methods of parallel corpora for the Kazakh language. The Kazakh language is agglutinative with rich morphology and related to the Turkic language group. So, the traditional alignment methods for similar languages do not work for the Kazakh language. The alignment is used primarily to ensure that the fragment corresponding to the original is found in the translation. After that, identical fragments of parallel texts are compared with each other. At the initial stage, the question is what needs to be leveled. It is possible to align word by word, but this often becomes almost impossible for several reasons: sets of lexemes and expressions do not match in different languages. Considering the linguistic peculiarities of languages, the developed technologies and ways of universal alignment of parallel text may not work in languages with agglutination. It means that the form of the word is formed by additional affixes and auxiliary words that carry semantic and morphological information. The approach presented in this paper is to use a two-stage alignment, which uses a bilingual dictionary of synonyms. The evaluation with the use of the English-Kazakh corpus verifies that our method shows an average of 89 % correct alignment. The second method is designed to expand the parallel corpus due to the lack of natural parallel corpora of the Kazakh-English language pair with good quality. The developed method uses a combinatorial method taking into account the semantic and grammatical features of the Kazakh language. Different tenses of the Kazakh language are used for sentence generation, and different endings for parts of speech are also considered

Keywords: parallel corpora, aligning, Kazakh, English, sentence generation, extending technology

UDC 519.766
DOI: 10.15587/1729-4061.2022.259452

ALIGNING AND EXTENDING TECHNOLOGIES OF PARALLEL CORPORA FOR THE KAZAKH LANGUAGE

Diana Rakhimova
PhD*
Aidana Karibayeva
Corresponding author
Master*
E-mail: a.s.karibayeva@gmail.com
*Department of Information Systems
Al-Farabi Kazakh National University
Al-Farabi ave., 71, Almaty,
Republic of Kazakhstan, 050040

Received date 03.06.2022
Accepted date 02.08.2022
Published date 31.08.2022

***How to Cite**: Rakhimova, D., Karibayeva, A. (2022). Aligning and extending technologies of parallel corpora for the kazakh language. Eastern-European Journal of Enterprise Technologies, 4 (2 (118)), 32–39. doi: https://doi.org/10.15587/1729-4061.2022.259452*

1. Introduction

Corpus formation is an urgent task of many modern world communities, as codified languages and their implementation in all styles and genres can correspond to the state status of the language.

The parallel corpora is an important tool for scientific research, for example, in the fields of typology, constructive linguistics, intralinguistic variation, and also for research on the theory and implementation in practice of statistical and neural machine translation. It is actively used in search engines, for conducting various analyses and tonality of the text, and in compiling various dictionaries for certain language pairs. The development of high-quality linguistic data (dictionaries, corpora, etc.) for low-resource languages will allow active development in the modern world of information technology.

Parallel corpora alignment, the automatic matching of sentences or words in the exact text to their translation equivalents, is an essential preprocessing step for many applications.

The volume and quality of the parallel corpus are the main factors in obtaining high-quality machine translation. However, aligning a parallel corpus of two texts is not as easy as it seems for several reasons. First, translators often do not translate the text from one sentence to one. It is especially noticeable when translating into hieroglyphic texts (Chinese, Japanese, etc.), where complex sentences, as a rule, will be divided into several simple ones. In translations into other languages, this is also quite common, and in corpora containing the Kazakh language.

Some sentences or paragraphs may simply be missing, and sometimes the translator adds something of his own.

Methods of alignment of sentences are divided into about three categories:

– based on length, assuming that the duration of the sentence in the original and the translation is approximately the same;

– based on bilingual lexical information.

Algorithms that include reference symbols align sentences based on information in the tagged case or spelling similarity.

Length-based methods are susceptible to spaces, as they can lead to incorrect alignment from one point of space to the end of the corpus.

To calculate the similarity of two structural units of texts, a particular criterion of similarity is introduced, for example, the number of translation equivalents in the dictionary. Then, the obtained weight is normalized to the length of the text so that the values of different text units are comparable.

The dynamic programming method is used to solve the smoothing problem optimally. However, it is impossible to calculate the entire matrix for large enough texts due to the considerable time required.

The Kazakh language belongs to the agglutinative group, and it isn't easy to find large and high-quality parallel corpora for such a group of languages. The difficulty lies in that when aligning the Kazakh sentence of the source language with English as the target language, one Kazakh word can be translated into three words. For example, the word "tuystarymizdan" has the following translation in

English – "from our relatives", which complicates parallel alignment at the word level.

The problem described in the Kazakh language also occurs in other language pairs, consisting of Kazakh as a source or target language. The task becomes more complicated when parallel corpora for languages belonging to different language groups and languages from other group families are developed. For example, the Kazakh language belongs to the Turkic language family, and the English language belongs to the Indo-European family.

Aligned parallel corpora are used in several different linguistic and computational linguistics areas.

2. Literature review and problem statement

Although parallel corpora are necessary language resources for many natural language processing tasks, they are rare or even unavailable for many language pairs. Instead, comparable corpora containing parallel pieces of information are widely available that can be used in applications such as statistical and neural machine translation systems.

In [1], sentences and dictionaries use the corpus, which must be aligned as the only source of information. This paper identifies problems such as sentence-level alignment, determining which sentence in the original matches each sentence in translation, and vocabulary alignment, to determine which word in one language is equivalent to each word in the translation. Their job was to find related words in a pair of possible pairs of sentences; the more associated words in a pair, the more likely it would contain equivalent sentences. However, the presented approach can be used to align individual sentences and cannot analyze the text as a whole. And the definitions of related words depend on finding related words in a sentence, which may not always be within the same sentence.

Alignment for the English-Hindi language pair is considered in [2]. The score of the source and target sentences is determined by comparing fragments of both sentences using English and Hindi vocabulary. The authors divided all source and target sentences into smaller blocks based on language blocks. In the authors' proposed method, the source sentence is first compared with a set of possible sentences that can be a translation of the original sentence. Each such comparison is assigned a comparison score. Based on the comparison score, the parallel corpus was aligned. This approach is very interesting, but may not always be applicable to different types of languages. For example, for Turkic languages with an agglutinative syntactic structure, it is not possible to compare the methods of separation by blocks, because many syntactic structures (as prepositions, conjunctions, declensions, which are separately formed in the text as in English, German or Spanish) are formed in word formation with the help of sets of suffixes, the definition of which requires additional analysis.

[3] describes a statistical method for aligning parallel corpora with their translations in two parallel corpora. The only information about the offers used to calculate the alignments was the number of tokens they contained. The work did not use the linguistic details of the sentence. This approach did not use the language details of the sentence and the properties of the language. This approach may not give good alignment results for different types of languages or with different structures.

The algorithm developed by the authors in [4] uses features derived from the distributive properties of words and does not use language-specific knowledge. Instead, the work used the context of sentences and the concept of Zipf's word vectors, which effectively model the distributive properties of words in a given sentence for English, Bulgarian, Czech, Estonian, Lithuanian, Serbo-Croatian, and Slovenian, which are analytic languages. The paper considers only West Germanic languages and East Slavic languages, and it is impossible to apply these features of the algorithm to other groups of languages.

Extracting parallel corpora from websites is also a direction to build parallel corpora. For example, many methods have been proposed to extract parallel texts in sentences or phrases from Wikipedia, such as similarity in [5]. In addition, the topic categories from Wikipedia were used to align multilingual corpora in [6]. Such resource as Wikipedia is very convenient since it is multilateral and multilingual. However, search of data according to the portal depends on the set tags and since developers for different languages are different people, materials can be not advising and not cover sufficient completeness.

The study [7] proposes a generative latent Dirichlet distribution model for extracting parallel fragments from similar documents. The experimental results are significantly improved if the extracted fragments generated by the proposed method are used to expand the existing parallel corpus in a statistical machine translation system. According to human estimates, the accuracy of the proposed method for the English-Persian assignment is about 59.7 %.

Expanding the range of application of corpus methods and structures, various works have been published in the journal [8], which brings together research and reports for an audience of researchers and practitioners interested in the range of applications of corpus linguistics. The role of applied corpus linguistics is to provide a forum for further theorizing corpus data analysis methods, exchanging examples and new methods, and promoting the development and consolidation of applied corpus linguistics as a significant force in information technology-based social research.

The authors [9] conducted a preliminary survey using PRISMA guidelines, searched the most widely used Information Technology (IT) databases, and identified free and accessible Arabic corpora. As a result, they identified a total of 48 available sources of corpora available free of charge in Arabic. The results were classified by corpus type into five categories depending on their primary purpose. [9, 10] presented applications of methods and structures of a monolingual corpus, but there is no work related to the alignment or development of a parallel corpus.

The authors have presented a study of sentence alignment using a small corpus of reference alignments and two large corpora containing aligned novels for English-Spanish and English-French language pairs in [11]. The proposed study also had several obvious weaknesses in implementation. The proposed model needs to compute scores for null links, which is a nearly impossible task because "real" deletions are difficult to predict based only on the text.

The algorithms and approaches presented above have been actively used and developed for well-known many resource languages. That allowed them to get quite good results. Below are some works on the topic of the study of the Kazakh language.

The cross-lingual sentence embeddings are proposed for low-resource languages in [12]. The method uses cross-lingual sentence embeddings trained from parallel sentence pairs. The paper [13] tackles the inconsistencies by investigating the neighborhoods of a given sentence pair. In these works, the alignment is directly related to the dictionary, and the accuracy of the alignment depends on the volume of the dictionary used.

In [14], the parallel corpus for the Kazakh language was assembled using Bitextor, and alignment was done using the LibTagAligner library. The aligned parallel corpus with LibTagAligner amounted to about 30,000 sentences. The authors managed to get not bad practical results. The inconvenience of this approach is that it takes a lot of time to collect texts using Bitextor, and to collect a parallel corpus from sites, you need to have corresponding web pages, which is not always the case for different languages.

Practical problems with agglutinative languages are also discussed in [15]. The work used the Helsinki FiniteState Toolkit (HFST) to process rule-based analysis, researched its benefits for morpheme-based alignments, and used the GIZA++ tool, which traverses two words alignments and merges the alignments. This approach is limited and only allows you to extract the correct phrase table from the word alignment.

In [16], texts from four Kazakh bilingual news sites were examined. The authors created a parallel corpus of texts based on criminal topics. To obtain results, the authors needed to carry out a lot of work on data preparation and preprocessing – lexical correspondences of both languages were developed in advance and the meanings of the tags of parts of speech were used to coordinate the corpus. As a result, 60 % of the proposals of the assembled corpora were automatically aligned correctly to the given (only criminal) topic.

Various works and approaches for processing, collecting and aligning parallel corpora were considered [17]. Of course, each work contributes to the development of different languages and their applications. Unfortunately, many approaches are limited to alignment at the level of words, phrases, or just one sentence. Also, many approaches to collecting corpora depend on the electronic resources available. These are websites, dictionaries, monolingual corpora, and so on. This is not always convenient for low-resource languages, such as Kazakh, Kyrgyz, Uzbek, etc. And of course, it is necessary to take into account the linguistic features of each language. When analyzing and working with different (not related groups) languages, corresponding difficulties arise. For example, English and Kazakh are from different language groups and have many differences and inconsistencies.

In [18], pairs of HTML documents were parsed site-by-site using the BeautifulSoup Python library to produce document-level aligned raw texts. In addition, sentence alignment was performed on tokenized and lowercase texts using Hunalign [19].

Work on the alignment of the parallel corpus of the English-Kazakh pair is very small and limited. The considered approaches cannot be applied because the Kazakh language has complex syntactic and semantic structures. This must be taken into account when analyzing and comparing with another language. The study proposes an alignment approach based on a bilingual dictionary of synonyms and expansion based on the grammar of the Kazakh language. In addition, it demonstrates the suitability of the proposed automated system for processing whole texts to create parallel corpora for the English-Kazakh language pair with high accuracy.

3. The aim and objectives of the study

The study aims to develop an information system that automatically aligns the parallel corpora for the English-Kazakh language pair.

The following objectives were set to achieve the aim:

– to develop the two-stage parallel corpus alignment for the English-Kazakh language pair and show the results obtained by the developed aligning method for English-Kazakh;

– to develop a general concept of extending technology for parallel corpora with the Kazakh language.

4. Materials and methods

4. 1. A two-stage alignment method for the Kazakh language

A parallel corpora or parallel translations corpora is a corpus consisting of texts in one language and their translation into another. Creating a parallel corpus may include several of the following steps, such as text alignment, text markup, etc.

This section described a two-stage alignment algorithm for an aligning synthetic parallel corpus of the English-Kazakh (and vice versa) language pair.

The first part of the alignment uses the Hunalign tool. The Hunalign program relies on a series of statistical weights (sentence length, character match, punctuation structure, etc.) and assigns a particular alignment probability factor to each aligned pair of sentences. It aligns bilingual text at the sentence level. If it is below zero, then the alignment of these text segments is unlikely. Sentence pairs with a negative coefficient and all glued sentences are considered doubtful segments that need to be checked manually first. Questionable segments are highlighted in red. The editor reviews them, and errors are corrected manually. On the Hunalign input, the text must be tokenized and segmented into bilingual sentences. Its output is a sequence of bilingual pairs of sentences with weights.

Fig. 1 shows a synonym-based alignment algorithm, where SL is the source language, and TL means target language.

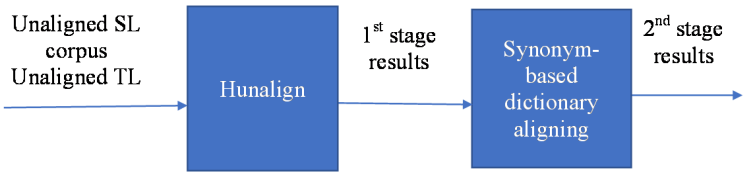

Fig. 1. The proposed method workflow

The command to run for alignment with Hunalign is the following:

HUNALIGN_PATH -text -bisent -utf HUNALIGN_DICTIONARY_FILE_PATH kaz_file eng_file >> RESULT_FILE_PATH.

Here:

– HUNALIGN_DICTIONARY_FILE_PATH is the path to the dictionary file;

– RESULT_FILE_PATH specifies the path where the file with the results is saved;

– text is the result that must be in text format, but not in the default (numeric) format;

– bisent is only bi-sentences (one-to-one alignment segments) that are printed.

The most common evaluation processes for corpora alignment quality typically rely on manually aligned and annotated corpora, which are used as a gold standard. Unfortunately, no such qualitative gold standard existed for the mentioned languages. Therefore, the method of "TF-IDF" (term frequency-inverse document frequency) was used to establish important words [20–23] (1)–(3). Important words were used to measure word overlap between sentences. It can be concluded whether pairs of sentences aligned through Hunalign are correctly aligned.

$$tf - idf(t,d,D) = tf(t,d) \cdot tdf(t,D), \quad (1)$$

$$tf(t,d) = \log\left(1 + freq(t,d)\right), \quad (2)$$

$$tdf(t,d) = \left(\frac{N}{count\left(d \in D : t \in d\right)}\right), \quad (3)$$

where t – unique term (or word), d – document (corpus), D – set of all documents (parallel corpora).

The weight in *tf-idf* will be weights between 1 and 0 of all unique words in parallel corpora.

Look at all sentences in the source language for each word in the TF-IDF wordlist. If a word appears in a sentence, store all the words in the corresponding target language sentence as a possible alignment.

Before running the command, it needs to prepare the following data:

– kaz_file – text in the Kazakh language;

– eng_file – text in the English language;

HUNALIGN_DICTIONARY_FILE (en_kaz.dic) is the bilingual dictionary. The content of a bilingual dictionary file has the following format:

– about @ туралы[turaly];

– ambassador @ елші[yelshi];

– asia @ азия[aziya].

The algorithm for the two-stage alignment of the parallel corpus of the English-Kazakh pair is based on the following:

– a dictionary with synonyms for the English-Kazakh pair is created, which has the following format:

English word *w*: synonym 1 in Kazakh, synonym 2 in Kazakh, ..., synonym *n* in Kazakh.

The volume of the dictionary of synonyms for the English-Kazakh pair is 31,097:

– the resulting corpus is cleaned with Hunalign and divided into two files with the corresponding language (for the Kazakh language and separately for the English language), in which each sentence begins on a new line;

– the length of each monolingual corpus is compared;

– each sentence of the monolingual English language is checked line by line with an English word from the English-Kazakh dictionary of synonyms.

At the same time, a search is done for the words by synonyms of the monolingual corpus of the Kazakh language (Fig. 2).

The sentences are saved to a new file when finding words from the dictionary of synonyms in each monolingual corpus. If there is no match, then the sentences are deleted.

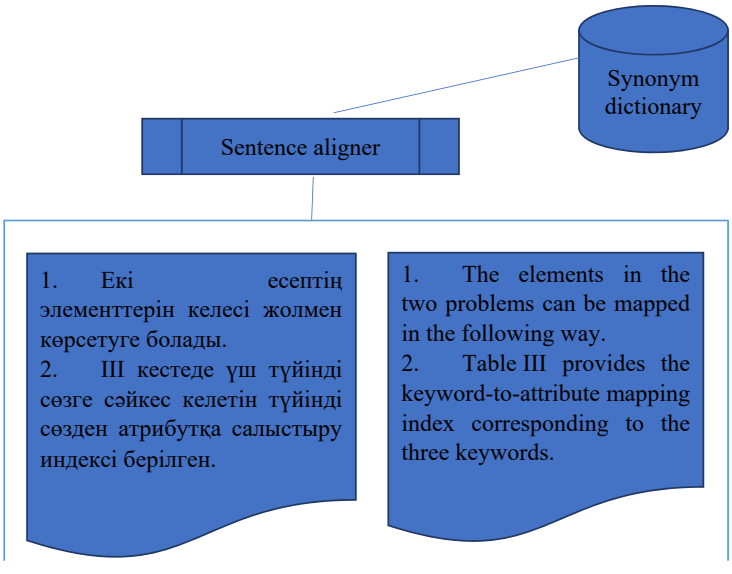

Fig. 2. Proposed synonym-based alignment method on the example of English-Kazakh pair

The first part, as well as the English-Kazakh alignment, uses the Hunalign tool. On the Hunalign input, the text must be tokenized and segmented into sentences. The second part of the alignment is based on a dictionary with synonyms for the English-Kazakh language pair. Therefore, it is best to use the dictionary alignment method for the English-Kazakh language pair belonging to different language groups, namely the dictionary of synonyms.

4. 2. Parallel corpora extending method for the Kazakh language

In this part of the paper, a method for generating simple sentences for the Kazakh language will be considered to increase the volume of the parallel corpora. The developed method takes into account the morphological and syntactic rules of the Kazakh language. The sentences are formed depending on the part of speech. The Kazakh language has nine parts of speech. In composing a sentence, the part of speech is selected not at once but according to the meaning of the sentence and the idea being expressed.

In order to generate a simple sentence, nouns, numbers, verbs, and adverbs were considered from the part of speech.

Another important rule to consider is tenses. There are three tenses in the Kazakh language: present, past, and future. They indicate when the event occurred. The present tense refers to an action that takes place while speaking; the past tense describes an action that took place before the present tense; the future tense refers to an action that did not yet occur while speaking.

The following formula is used to calculate possible numbers for the generation of simple sentences consisting of different parts of speech in the Kazakh language (4)–(13):

$$C_n^k = \frac{n!}{k!(n-k)!}. \quad (4)$$

The calculation for a combination of 1 part of speech from 9, where n=9, k=1:

$$C_9^1 = \frac{9!}{1!(9-1)!} = 9. \quad (5)$$

The calculation for a combination of 2 parts of speech from 9 (n=9, k=2):

$$C_9^2 = \frac{9!}{2!(9-2)!} = 36. \quad (6)$$

The calculation for a combination of 3 parts of speech from 9, where n=9, k=3:

$$C_9^3 = \frac{9!}{3!(9-3)!} = 84. \quad (7)$$

The calculation for a combination of 4 parts of speech from 9, where n=9, k=4:

$$C_9^4 = \frac{9!}{4!(9-4)!} = 126. \quad (8)$$

The calculation for a combination of 5 parts of speech from 9, where n=9, k=5:

$$C_9^5 = \frac{9!}{5!(9-5)!} = 126. \quad (9)$$

The calculation for a combination of 6 parts of speech from 9, where n=9, k=6:

$$C_9^6 = \frac{9!}{6!(9-6)!} = 84. \quad (10)$$

The calculation for a combination of 7 parts of speech from 9, where n=9, k=7:

$$C_9^7 = \frac{9!}{7!(9-7)!} = 36. \quad (11)$$

The calculation for a combination of 8 parts of speech from 9, where n=9, k=8:

$$C_9^8 = \frac{9!}{8!(9-8)!} = 9. \quad (12)$$

The calculation for a combination of 9 parts of speech from 9, where n=9, k=9:

$$C_9^9 = \frac{9!}{9!(9-9)!} = 1. \quad (13)$$

Below are possible combinations that take into account the syntax and morphological features in the construction of a sentence of the Kazakh language with 2 parts of speech with examples:

– noun+verb;
– noun+adjective;
– pronoun+adjective;
– pronoun+verb;
– adverb+verb;
– numeral+verb;
– adverb+adjective;
– шылау+verb;
– imitation word+verb;
– adjective+imitative words;
– noun+numeral;
– interjection+adjective;
– interjection+noun;
– interjection+verb;
– interjection+numeral;
– pronoun+adjective;
– pronoun+noun;
– verb+pronoun.

Eighteen combinations are possible from the 36 combinations (6). Below, Table 1 presents some parts of speech combinations with examples to it.

Table 1

The combinations with examples

Combination of parts of speech	Examples in Kazakh
Noun	Күз
Noun+Verb	Күз келді
Noun+Adjective+Verb	Қоңыр күз келді
Noun+Adverb+Verb	Күз кеш келді
Noun+Adjective+Adverb+Verb	Қоңыр күз кеш келді

By creating a sentence-generating corpus for the Kazakh and English language pairs, we have contributed to the development of modern parallel corpus and increased the number of the parallel corpus.

Fig. 3 shows the simple sentence generation for the Kazakh-English language pair.

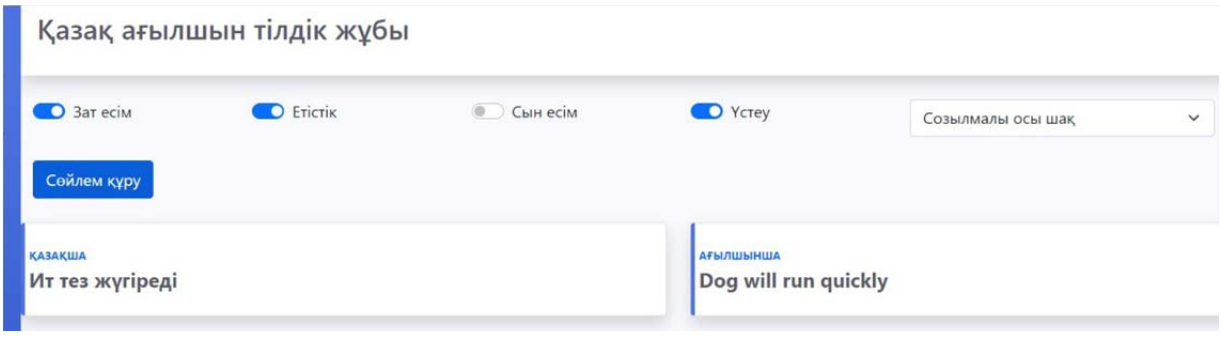

Fig. 3. The generation result for a combination of three parts of speech

5. Research results of parallel corpora extending and aligning for the Kazakh language

5. 1. Results for parallel corpora extending technology

In this paper, a maximum combination of 4 parts of speech was used to generate a sentence, and it also took into account all the endings of the person, cases, possessives, and tenses of verbs.

During the work, each combination was studied, and the number of sentences in Table 2 was generated for each.

Table 2

Results for simple sentence generation for Kazakh-English pair

Part of speech combination	Number of generated sentences
Noun	325
Noun+Verb	275
Noun+Adjective+Verb	175
Noun+Adverb+Verb	150
Noun+Adjective+Adverb+Verb	120
Total	1,045

The number of generation of simple sentences directly depends on the completeness of the dictionary.

5. 2. Results for parallel corpora aligning technology

All text resources and corpora are stored in the database; primary text data is recorded in the database resources folder. This data is processed – the text must be tokenized and divided into sentences, and then only submitted for alignment of this approach and system. As a result, the resulting parallel corpora are written to a separate folder – the results of our database. The resulting folder of results differs from the original one, because irrelevant data is deleted after processing. The request to the database goes with the help of the read/request function, and in the future, the results obtained are written to a new *.txt file in the database.

In the experimental part, the first step was using the Hunalign tool. The second step used the synonym-based alignment method because Hunalign returned incorrectly aligned sentences. The alignment results based on the synonym method are shown in Table 3.

Table 3

Experimental alignment data for English-Kazakh language pair

Number of original sentences	Number of aligned sentences	Aligned correctness with the proposed method in percentage, %	Aligning correctness with Hunalign in percentage, %
5,287	4,652	88	84
21,000	19,110	91	87

The parallel English-Kazakh corpus contains texts from some Kazakh news sites, such as nu.edu.kz and akorda.kz [24, 25].

Therefore, alignment results depend on the completeness of the words in the synonym dictionary.

Good alignment results directly depend on the volume and coverage of the dictionary of synonyms of the entire corpus. As with many knowledge search, natural language processing, or pattern recognition systems, the performance of parallel text alignment algorithms is usually measured in terms of precision (14), recall (15), and *F*-score (16) [26].

$$\text{precision} = \frac{TP}{TP + FP}, \tag{14}$$

$$\text{recall} = \frac{TP}{TP + FN}, \tag{15}$$

$$F\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{16}$$

where TP – true positive, FP – false positive, and FN – false negative. This paper used the method for calculating precision, *F*-score, and recall for all mentioned language pairs.

Table 4 presents the evaluation results for the alignment method based on the synonym dictionary for the Kazakh-English language pair.

Table 4

Evaluation results for the alignment method for Kazakh and English

Language	Recall	Precision	Accuracy	F-score	Language
Kazakh	0.883	0.883	0.883	0.871	Kazakh
English	0.937	0.937	0.937	0.906	English

"Recall" indicates the classifier's ability to find all positive patterns, that is, correctly aligned sentences in the corpus. "Accuracy" defines correctly classified occurrences as the total number; that is, it is the correct number of matches to the total number in the corpus. Finally, "Precision" indicates the ability of the classifier not to label a negative pattern as positive; in the corpus, the "negative pattern" is a misaligned sentence, and "positive" is a correctly aligned sentence.

Table 4 shows the results of different assessments for aligning the Kazakh-English language pair. According to the estimates, it can be determined that the alignment method based on the synonym dictionary is effective for the mentioned pair.

6. Discussion of the obtained results of parallel corpora aligning and extending

The presence of rare words or phrases in the text will cause errors during the text-based automatic alignment method. For this reason, it is necessary to create new algorithms and techniques that preserve alignment quality to reduce costs. During the study, linguistic features of the Kazakh language were analyzed, and mathematical methods considering the features of the mentioned language for their solution were carried out. Most research works consider the high-resource languages in aligning and extending parallel corpora; in addition, there are very few works for low-resource languages, such as the Kazakh language. In our study, an attempt is to improve the results of the Hunalign tool by using a synonyms dictionary and extending parallel corpora volume by generating simple sentences based on a combinatorial approach, which considers the grammar of the

Kazakh language. A two-stage aligning method is proposed. This makes it possible to align a corpus with relatively better quality to resolve the lack of a good quality corpus. A simple sentence generation method is presented too. The peculiarity of the methods lies in integrating computational and linguistic models of text analysis at the preprocessing stage.

A two-stage alignment method was proposed because of an improvement in the alignment quality compared to using Hunalign alone. Comparative results are presented in Table 3. From Table 3, it is seen that the proposed method gives an improvement of 3–4 % over Hunalign.

The second mentioned method is for extending parallel corpora to increase the volume of parallel corpora of the Kazakh language. At the beginning of the research, the grammatical rules of the two languages under consideration were studied. This is because grammatical rules are one of the most important elements in sentence formation. This method checked the correctness of sentence generation with the compounding of the proper endings and changes in time.

The biggest limitations of this study lie in using a limited size of the dictionary. Both mentioned methods' results depend on the dictionary's completeness. Future work plans to add more words to the dictionary to prove these algorithms' effectiveness.

7. Conclusions

1. A two-stage alignment method is proposed for parallel corpora alignment for the Kazakh language. The developed method consists of two stages, the first stage uses Hunalign, the second uses a synonym-based alignment method, which improves the results of Hunalign by over 3–4 %. The developed method of two-stage alignment gives the best alignment quality, but the quality depends on the completeness of the synonym dictionary. For example, the alignment technology based on a dictionary of synonyms gave an average of 89 % correct alignment for the English-Kazakh language pair.

2. The extending technology is performed on low-resource languages, namely for English-Kazakh and Kazakh-English language pairs. The parallel corpora generation method is based on the combinatorial approach, where we take into account the syntactic and morphological features of simple sentences in the Kazakh language side. The results of the generated sentence show that the method works well for generating sentences with up to 4 parts of speech. It is planned for the future to consider sentences with more than four parts of speech in sentences.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Acknowledgments

This research was performed and financed by the grant Project IRN AP08052421 of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

1. Nazar, R. (2011). Parallel corpus alignment at the document, sentence and vocabulary levels. Procesamiento del Lenguaje Natural, 47, 129–136. Available at: https://core.ac.uk/download/pdf/16370668.pdf
2. Bharati, A., Sriram, V., Krishna, A. V., Sangal, R., Bendre, S. (2002). An Algorithm for Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings of ICON-2002: International Conference on Natural Language Processing. Mumbai. Available at: https://arxiv.org/ftp/cs/papers/0302/0302014.pdf
3. Brown, P., Lai, J., Mercer, R. (1991). Aligning Sentences in Parallel Corpora. ACL '91: Proceedings of the 29th annual meeting on Association for Computational Linguistics, 169–171. doi: https://doi.org/10.3115/981344.981366
4. Bicici, E. (2008). Context-Based Sentence Alignment in Parallel Corpora. Computational Linguistics and Intelligent Text Processing, 434–444. doi: https://doi.org/10.1007/978-3-540-78135-6_37
5. Adafre, S. F., de Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources. Available at: https://aclanthology.org/W06-2810.pdf
6. Smith, J. R., Quirk, Ch., Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, 403–411. Available at: https://aclanthology.org/N10-1063.pdf
7. Bakhshaei, S., Safabakhsh, R., Khadivi, S. (2019). Extracting parallel fragments from comparable documents using a generative model. Computer Speech & Language, 53, 25–42. doi: https://doi.org/10.1016/j.csl.2018.07.002
8. Flinn, A. (2021). Review of Doval and Sánchez Nieto (2019). Parallel Corpora for Contrastive and Translation Studies: New resources and applications. Applied Corpus Linguistics, 1 (2), 100007. doi: https://doi.org/10.1016/j.acorp.2021.100007
9. Ahmed, A., Ali, N., Alzubaidi, M., Zaghouani, W., Abd-alrazaq Alaa A, Househ, M. (2022). Freely Available Arabic Corpora: A Scoping Review. Computer Methods and Programs in Biomedicine Update, 2, 100049. doi: https://doi.org/10.1016/j.cmpbup.2022.100049
10. Sennrich, R., Volk, M. (2011). Iterative, MT-based Sentence Alignment of Parallel Texts. NODALIDA 2011 Conference Proceedings, 175–182. Available at: https://aclanthology.org/W11-4624.pdf
11. Xu, Y., Max, A., Yvon, F. (2015). Sentence alignment for literary texts. Linguistic Issues in Language Technology, 12. doi: https://doi.org/10.33011/lilt.v12i.1383

12. Chaudhary, V., Tang, Y., Guzmán, F., Schwenk, H., Koehn, P. (2019). Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings. Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2). doi: https://doi.org/10.18653/v1/w19-5435
13. Artetxe, M., Schwenk, H. (2019). Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/p19-1309
14. Zhumanov, Z., Madiyeva, A., Rakhimova, D. (2017). New Kazakh Parallel Text Corpora with On-line Access. Lecture Notes in Computer Science, 501–508. doi: https://doi.org/10.1007/978-3-319-67077-5_48
15. Kartbayev, A. (2015). Refining Kazakh Word Alignment Using Simulation Modeling Methods for Statistical Machine Translation. Lecture Notes in Computer Science, 421–427. doi: https://doi.org/10.1007/978-3-319-25207-0_38
16. Rakhimova, D. R., Turganbaeva, A. O. (2020). Normalization of Kazakh language words. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 20 (4), 545–551. doi: https://doi.org/10.17586/2226-1494-2020-20-4-545-551
17. Khairova, N., Mamyrbayev, O., Mukhsina, K. (2019). The Aligned Kazakh-Russian Parallel Corpus Focused on the Criminal Theme. Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems (COLINS-2019). Volume I: Main Conference. Kharkiv, 116–125. Available at: http://ceur-ws.org/Vol-2362/paper11.pdf
18. Assylbekov, Zh., Myrzakhmetov, B., Makazhanov, A. (2016). Experiments with Russian to Kazakh sentence alignment. The 4-th International Conference on Computer Processing of Turkic Languages "TurkLang 2016". Available at: https://nur.nu.edu.kz/handle/123456789/1694
19. Hunalign. Available at: https://github.com/danielvarga/hunalign
20. Singhal, A., Buckley, C., Mitra, M. (1996). Pivoted document length normalization. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '96. doi: https://doi.org/10.1145/243199.243206
21. Wu, H. C., Luk, R. W. P., Wong, K. F., Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems, 26 (3), 1–37. doi: https://doi.org/10.1145/1361684.1361686
22. Lavin, M. J. (2019). Analyzing Documents with TF-IDF. Programming Historian, 8. doi: https://doi.org/10.46430/phen0082
23. Arroyo-Fernández, I., Méndez-Cruz, C.-F., Sierra, G., Torres-Moreno, J.-M., Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF–IDF. Computer Speech & Language, 56, 107–129. doi: https://doi.org/10.1016/j.csl.2019.01.005
24. Nazarbayev University. Available at: https://nu.edu.kz/
25. Akorda. Available at: https://www.akorda.kz/en
26. Sueno, H. T., Gerardo, B. D., Medina, R. P. (2020). Converting text to numerical representation using modified Bayesian vectorization technique for multi-class classification. International Journal of Advanced Trends in Computer Science and Engineering, 9 (4), 5618–5623. doi: https://doi.org/10.30534/ijatcse/2020/211942020