https://doi.org/10.15407/intechsys.2025.01.050
UDC 004.8 + 004.032.26

M.I. KOROTIUK, Student,
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute",
37, Beresteyskyi ave., Kyiv, 03056, Ukraine
mariakorotiuk@gmail.com

N.A. RYBACHOK, PhD (Engineering), Assoc. Professor,
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute",
37, Beresteyskyi ave., Kyiv, 03056, Ukraine
https://orcid.org/0000-0002-8133-1148
rybachok@pzks.fpm.kpi.ua

AUTOMATED SENTENCE ALIGNMENT IN UKRAINIAN-GERMAN PARALLEL TEXTS

Introduction. *Sentence alignment in German parallel texts is a relevant task. It allows obtaining parallel data sets necessary for many computational linguistics tasks, such as parallel corpus construction and machine translation. The article describes the main tasks of sentence alignment, reviews existing methods and analyzes their ideas. Based on this analysis, a new method is proposed, it is based on the Bleualign approach, which uses machine translation systems and the BLEU metric to assess the similarity of sentences. However, it differs in the use of additional marker dictionaries for industry terms and conjunctions, including their synonyms. This article outlines the main tasks of sentence alignment, reviews existing methods, and discusses their ideas. Based on this analysis, a new method is proposed. This method is based on the Bleualign approach, which uses machine translation systems and BLEU metrics to evaluate sentence similarity, including the alignment of parts of complex sentences. However, it differs in the alignment process steps and introduces additional marker dictionaries for domain-specific words and conjunctions, including their synonyms.*

The purpose of the work *is to develop a method and software for automated sentence alignment in Ukrainian-German parallel texts.*

Methods. *The developed method is based on the Bleualign method and the BLEU metric. It is improved by the use of dictionaries of industry terms and conjunctions, and also provides a focus on one language pair — Ukrainian-German. The proposed method consists of*

Cite: Korotiuk M.I., Rybachok N.A. Automated Sentence Alignment in Ukrainian-German Parallel Texts. *Information Technologies and Systems,* Kyiv, 2025, Vol. 1 (1), 50–58. https://doi.org/10.15407/intechsys.2025.01.050

6 stages, allowing to align sentences in Ukrainian-German parallel texts. The proposed method is implemented in software using the Python programming language.

Results. *A new method for aligning sentences for Ukrainian-German parallel texts has been developed and its software implementation has been completed. The proposed method is based on statistical approaches and does not require significant computing resources. Unlike the Bleualign method, it uses dictionaries of industry terms and conjunctions for more accurate sentence alignment.*

Conclusions. *Further research will include experiments and comparison of the alignment results obtained using the proposed method with the results of the Bleualign method.*

Keywords: *sentence alignment, parallel texts, machine translation, BLEU metric, dictionaries.*

Terminology

Sentence alignment — the process of establishing correspondence between sentences with the same meaning in parallel texts.

Parallel texts — texts in different languages that convey the same meaning and are translations of each other.

BLEU metric — a statistical method for evaluating translation quality by comparing the source and target texts using n-grams.

Cosine similarity — a metric for calculating the similarity between vectors, measuring the cosine of the angle between them; it can be used to assess the similarity of vector representations of text features.

Machine translation — an automated process of translating text from one language to another using software, without direct human involvement.

Marker — a unique identifier that highlights specific characteristics of a word, phrase, or text segment, which hold particular significance in language processing, analysis, or translation.

Introduction and Problem Statement

In the field of computational linguistics, an important task is the development of new and improvement of existing methods for segment alignment of parallel texts, particularly sentences. Sentence alignment is a key step in obtaining data that can be used for building parallel corpora, training machine translation models, and analyzing the style and syntactic structure of sentences. The effectiveness of these tasks depends on the accuracy and volume of available parallel data.

There are many methods that can be applied to parallel texts belonging to different language pairs. However, their use often requires significant computational resources, such as through the use of multilingual models. Furthermore, the accuracy of such methods may be low if they are based solely on statistical approaches to data processing, especially if they do not take into account the lexical components of the sentences. Since the German language is widely used globally, and there are significantly fewer parallel data available for the Ukrainian language compared to English, it seems reasonable to develop a sentence alignment method for parallel texts in the context of these languages. Focusing on a specific language pair will allow for the creation of a sentence alignment method

optimized for the lexical characteristics of a specific field, with fewer computational costs compared to methods based on multilingual models, thanks to the use of statistical approaches. This will enable the method to be applied on standard hardware without specialized devices, making it more accessible to a broader audience.

In light of the above, the goal of this study is to improve the accuracy of automated sentence alignment in Ukrainian-German parallel texts by developing a method that ensures the acquisition of a parallel data set using statistical approaches, without requiring significant computational resources. To achieve this goal, it is necessary to review existing methods of automated sentence alignment, analyze their advantages and disadvantages, and, based on the results, develop a proprietary method and implement its software solution.

Literature Review

Among the existing methods of sentence alignment in parallel texts, the Gale-Church method, Hunalign, Bleualign, and Vecalign can be distinguished. Each of these methods uses different approaches to alignment, so their accuracy can vary significantly.

The Gale-Church method is based on the assumption that longer sentences in one language are typically translated into longer sentences in another language, and shorter sentences are translated into shorter ones [1]. The statistical approach, which compares the number of characters in sentences using dynamic programming methods, allows for fast results even for large texts. However, for obtaining accurate results, this might be insufficient, as it does not take into account the lexical components of sentences, which plays an important role in determining sentence similarity.

An alternative to the described method could be Hunalign — a hybrid method that combines the use of dictionaries and sentence length determination [2]. This allows for quite accurate results without the need for significant computational resources, provided that the dictionary is accurate and complete.

Since compiling and formatting a large dictionary can be a time-consuming process, especially when there is insufficient available data, the Bleualign method can be used. The main idea of this method is to use machine translation of the text and the BLEU metric as a measure of similarity to find reliable correspondences between sentences that serve as anchor points [3]. Due to the widespread use of machine translation systems, they can be applied to bring parallel texts to one language, avoiding the creation of a special dictionary with a large number of words, as is done in the Hunalign method. Importantly, these systems can account for semantic relationships in sentences as well. Furthermore, using already translated text allows the method to use only statistical approaches, making it relatively simple, yet fast and efficient. However, there may be difficulties when comparing sentences that contain synonyms or have a shifted word order, while maintaining the same meaning. This could lead to

a loss of some parallel data when evaluating their similarity using the BLEU metric, which is based on exact word matches.

This issue can be resolved by the Vecalign method, which implements a new approach for bilingual sentence alignment [4]. It is linear in terms of both time and memory usage and requires only bilingual vector representations of sentences. The method involves using multilingual vector models, such as LASER. This approach often provides very high accuracy in results due to its consideration of both semantic and lexical components of the sentence. However, it requires significant computational resources, which may be unavailable under certain conditions.

Also let's look at other sentence alignment methods and tools that use advanced approaches and improved text processing algorithms.

Web Align Toolkit — offers an intuitive interface for automatic alignment of parallel texts and supports various language pairs. The tool enables text preprocessing, which improves alignment accuracy. However, its algorithms may produce errors when processing texts containing numerous idioms or specialized terminology. Additionally, the tool does not support processing of very large texts [5].

InterText — a software with a graphical interface for semi-manual alignment, useful for correcting results from automated methods. Its modular architecture allows integration of additional dictionaries and rules for specific language pairs. Among the drawbacks of this project are limited support for modern data formats and decreased performance when processing texts exceeding 10,000 sentences [6].

Bertalign — a BERT-based approach ensuring semantic sentence alignment [7]. It shows strong results when processing context-dependent expressions and idiomatic phrases. The method is especially useful for aligning literary texts where preserving not only content but also stylistic features is crucial. However, it demands substantial computational resources, which may limit its application on standard hardware. Its effectiveness also sharply declines in the absence of high-quality pre-trained models for specific language pairs.

Lingtrain Aligner — an interactive tool combining automatic alignment with manual editing capabilities [8], suitable for creating high-quality parallel corpora where precision in each match is essential. Its modular architecture allows adaptation to specific tasks and language pairs. However, the system requires extensive training for effective use and may yield unstable results for small corpora.

All these modern tools demonstrate a shift from traditional statistical methods to neural and hybrid approaches. Nevertheless, their efficiency often depends on the availability of computational resources, the volume of parallel data, and text complexity, which may pose limitations for certain applications.

Method of Automated Sentence Alignment in Ukrainian-German Parallel Texts

The method proposed in this research is based on the idea of the Bleualign method. It is suggested to use machine translation and sentence similarity assessment in parallel texts, which are brought to one language using the BLEU translation quality evaluation metric. This method, like Bleualign, also supports multi-step sentence alignment, including "one-to-one" and "one-to-many" approaches. The authors' contribution lies in refining the algorithm by integrating an additional lexical component — domain-specific term and conjunction dictionaries, which are used as markers for sentence matching. Each dictionary is manually compiled and formatted so that synonyms for each term or conjunction are defined in both languages along with a unique marker. The accuracy of sentence alignment using the proposed method largely depends on the size of the dictionaries and the number of synonymous equivalents: the broader the lexical coverage and the more synonym pairs, the higher the alignment accuracy. At the same time, due to their thematic focus (conjunctions and terms within a specific domain), their volume remains moderate, allowing for the quick creation of such dictionaries for specific tasks.

It is expected that the proposed method will improve sentence alignment accuracy in Ukrainian-German parallel texts by 3 – 5% compared to the Bleualign method.

The proposed method consists of 6 stages, which are shown in the action diagram (Fig. 1).

Stage 1. Obtaining two parallel texts — the source (source) and target (target) texts, the source language of the text, as well as the domain term dictionary.

Alignment is possible for both Ukrainian as the source language and German as the target language, as well as in the reverse direction. The proposed method does not require additional formatting of text data since it can work with texts that contain paragraphs.

Stage 2. Choosing the sentence alignment method.

There are two alignment options: within paragraphs or for the entire text.

For the first option, paragraph alignment is performed first by creating a vector for each paragraph, containing the calculated number of dictionary term markers for each term in the given text segment. To achieve this, the text is preprocessed: tokenization is performed to determine the total number of words in the text. Tokenization is also necessary for further lemmatization or stemming of both the text and the dictionary terms. This is important for bringing terms to a uniform form; otherwise, they would be considered different.

Next, the obtained vectors are normalized by dividing by the number of words in the paragraph.

Then, paragraph similarity is determined by calculating cosine similarity. Paragraphs are considered parallel if their similarity exceeds a de-

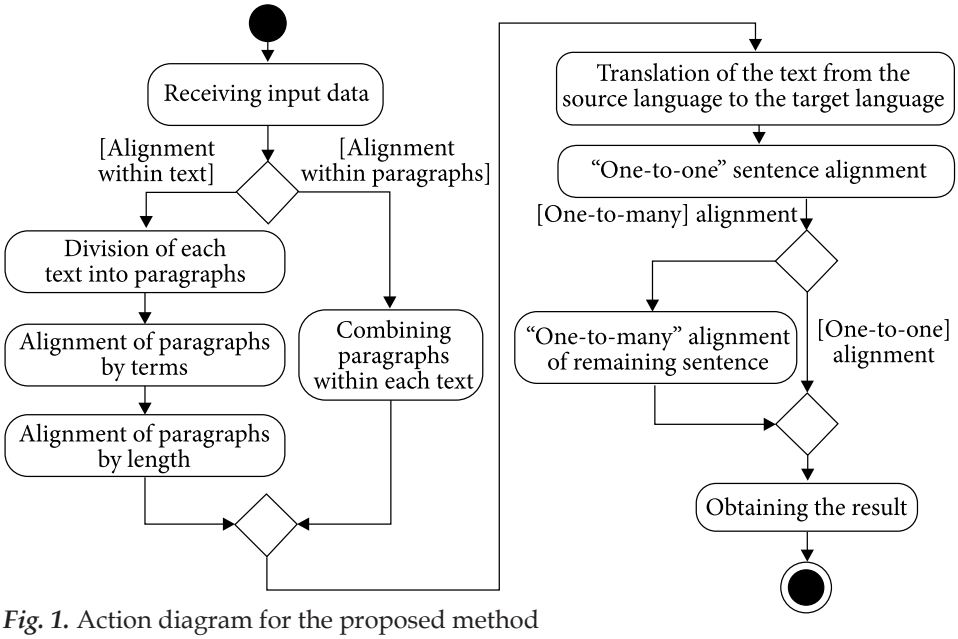

Fig. 1. Action diagram for the proposed method

fined threshold. Multiple parallel paragraphs may be found for each paragraph in the source text. However, there may be cases where some paragraphs do not have any parallel counterparts due to the absence of relevant terms in those text segments. In such cases, additional vectors for paragraphs are created based on the following indicators: the number of sentences in a paragraph and the average number of words per sentence.

Normalization for these additional vectors is performed in the same way as for term vectors. Cosine similarity calculations for these vectors help identify similar paragraphs, allowing an attempt to find parallel sentences among them at later stages.

In the second option, all paragraphs are combined, allowing sentence alignment to be performed regardless of their position within a paragraph or the text as a whole. This approach is suitable for small texts with different formatting due to the quadratic complexity of the method, as similarity is calculated for each possible sentence pair.

Stage 3. Translating the text from the source language to the target language.

Due to the availability of various libraries and resources for machine translation, the proposed method assumes the use of Ukrainian and German texts without requiring additional translations. Therefore, in implementation, the source text is automatically translated into the target language. This helps avoid errors that may arise when working with original and translated texts that do not fully match.

Stage 4. "One-to-one" sentence alignment within paragraphs or the entire text, depending on the results from the first stage.

Sentence alignment is performed in parallel texts that have been converted into the same language.

After tokenization and lemmatization or stemming, domain-specific terms are replaced with markers obtained from the respective dictionary. This reduces cases where synonyms are used in the translated source and target texts.

Sentence similarity is evaluated using the BLEU metric: sentences are considered parallel if their similarity exceeds a defined threshold. BLEU is usually used for translation quality assessment, but in this case, it is also useful because sentences in the same language are being compared. Thus, a high translation match indicates that they convey the same meaning.

A single sentence from the first text may have multiple parallel sentence options, just as a single sentence from the target text may correspond to multiple sentences from the source text.

It is important to correctly adjust the weights for n-grams used in the BLEU metric to account for word order in a sentence. Word order is important, but in translation, it may vary significantly while still conveying the same meaning.

Stage 5. Aligning sentences without identified parallels using the "one-to-many" approach (optional).

Complex sentences are split into simpler ones based on conjunctions identified in the dictionary, and the complex sentence is aligned with multiple simple ones. As in the previous stage, the BLEU metric is used to determine sentence similarity. This allows obtaining a greater number of parallel data.

Stage 6. Obtaining the result.

The alignment results are formatted as corresponding pairs of Ukrainian-German sentences.

Software Implementation of the Method

The software is implemented in Python as a command-line application with a modular architecture. The modules developed during the software implementation include:

- file handling module, which enables reading text data for alignment and saving results as a file;
- dictionary handling module, which retrieves necessary words and their corresponding markers;
- text processing module, which performs preprocessing and translation of text data;
- paragraph alignment module, which processes parallel texts and enables merging paragraphs or obtaining a set of parallel paragraphs from the target text for each paragraph in the source text;
- sentence alignment module, which implements sentence alignment using both "one-to-one" and "one-to-many" approaches and formats the alignment results;
- alignment process management module, which processes input data, calls all necessary modules, and provides the final result.

Python was chosen for software development due to its extensive set of tools for text processing, specifically utilizing the nltk and googletrans libraries.

All program modules include validation of input data, and functions incorporate error-handling mechanisms to address any issues that may arise during execution. This ensures stable and reliable program operation.

The required input data is provided via command-line arguments and includes parallel texts in the source and target languages, a dictionary of domain-specific terms, and the source language (Ukrainian or German).

Conclusions

Based on the conducted research on existing sentence alignment methods in parallel texts, an own method for Ukrainian and German has been proposed. This method builds upon the core idea of the Bleualign approach, allowing for lexical sentence alignment without relying on large dictionaries or significant computational resources. However, the proposed method introduces the use of a domain-specific dictionary containing terms relevant to the subject of the parallel texts, which are utilized as markers, along with conjunctions. This enhancement improves alignment accuracy by treating synonyms as equivalent terms.

A software implementation of the proposed method has been developed in Python.

Future research directions include:

- conducting experiments on sentence alignment in parallel texts across various domains, using specialized dictionaries;
- comparing alignment results obtained with the proposed method against those of the Bleualign method.

REFERENCES

1. Gale W., Church K. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics,* 1993, Vol. 19 (1), 75–102. URL: https://www.researchgate.net/publication/220355307_A_Program_for_Aligning_Sentences_in_Bilingual_Corpora [Accessed 27 Nov. 2024]
2. Halácsy P., Kornai A., Nagy V., Németh L., Trón V. Parallel corpora for medium density languages. *Recent Advances in Natural Language Processing IV,* 2007, Issue 1, 47–258. URL: https://www.researchgate.net/publication/282780901_Parallel_corpora_for_medium_density_languages [Accessed 27 Nov. 2024]
3. Sennrich R., Volk M. MT-based sentence alignment for OCR-generated parallel texts. *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers,* 2010, Issue 11. URL: https://www.researchgate.net/publication/281754851_MT-based_sentence_alignment_for_OCR-generated_parallel_texts [Accessed 27 Nov. 2024]
4. Thompson B., Koehn P. Vecalign: Improved Sentence Alignment in Linear Time and Space. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 2019, Issue 1, 1342–1348. URL: https://www.researchgate.net/publication/336999037_Vecalign_Improved_Sentence_Alignment_in_Linear_Time_and_Space [Accessed 27 Nov. 2024]. https://doi.org/10.18653/v1/D19-1136
5. Web Align Toolkit: Online parallel texts aligner and format converter. URL: http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit [Accessed 27 Nov. 2024]
6. InterText:parallel text alignment editor. URL: https://wanthalf.saga.cz/intertext [Accessed 27 Nov. 2024]

7. Liu L., Zhu M. Bertalign: Improved word embedding-based sentence alignment for Chinese-English parallel corpora of literary texts. *Digital Scholarship in the Humanities,* 2023, Vol. 38 (4), 621–634. URL: https://www.researchgate.net/publication/366682551_Bertalign_Improved_word_embeddingbased_sentence_alignment_for_Chinese-English_parallel_corpora_of_literary_texts [Accessed 27 Nov. 2024]. https://doi.org/10.1093/llc/fqac089
8. Lingtrain Aligner. URL: https://github.com/averkij/lingtrainaligner-editor/tree/t/master/docs2/docs/source [Accessed 27 Nov. 2024]

Received 06.02.2025

М.І. КОРОТЮК, студентка,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
просп. Беретейський, 37, м. Київ, 03056, Україна
mariakorotiuk@gmail.com

Н.А. РИБАЧОК, канд. техн. наук, доцент,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
просп. Беретейський, 37, м. Київ, 03056, Україна
https://orcid.org/0000-0002-8133-1148
rybachok@pzks.fpm.kpi.ua

АВТОМАТИЗОВАНЕ ВИРІВНЮВАННЯ РЕЧЕНЬ В УКРАЇНСЬКО-НІМЕЦЬКИХ ПАРАЛЕЛЬНИХ ТЕКСТАХ

Вступ. Вирівнювання речень в українсько-німецьких паралельних текстах є актуальним завданням, яке дає змогу отримувати набори паралельних даних, необхідних для багатьох завдань комп'ютерної лінгвістики, таких як побудова паралельних корпусів та машинний переклад. Стаття описує основні завдання вирівнювання речень, розглядає наявні методи та аналізує їхні ідеї. На основі цього аналізу пропонується новий метод, який ґрунтується на підході *Bleualign* і використовує системи машинного перекладу та метрику *BLEU* для оцінки схожості речень. Однак він відрізняється використанням додаткових словників маркерів для галузевих термінів та сполучників, включаючи їхні синоніми.

Мета. Розроблення методу та відповідного програмного забезпечення автоматизованого вирівнювання речень в українсько-німецьких паралельних текстах.

Методи. За основу розробленого методу використано метод *Bleualign* та метрику *BLEU*. Його удосконалено використанням словників галузевих термінів та сполучників, а також передбачено фокусування на одній мовній парі — українсько-німецькій. Запропонований метод складається із 6 етапів, які дозволяють виконати вирівнювання речень в українсько-німецьких паралельних текстах. Запропонований метод програмно реалізовано із використанням мови програмування *Python*.

Результати. Розроблено новий метод вирівнювання речень для українсько-німецьких паралельних текстів та виконано його програмну реалізацію. Запропонований метод базується на статистичних підходах і не вимагає значних обчислювальних ресурсів. На відміну від методу *Bleualign*, у ньому використано словники галузевих термінів і сполучників для більш точного вирівнювання речень.

Висновки. Подальші дослідження включатимуть проведення експериментів і порівняння результатів вирівнювання, отриманих при застосуванні запропонованого методу, із результатами методу *Bleualign*.

Ключові слова: *вирівнювання речень, паралельні тексти, машинний переклад, метрика BLEU, словники.*