Sentence-Alignment in Semi-parallel Datasets

Steffen Frenzel and Manfred Stede
Applied Computational Linguistics
University of Potsdam
steffen.frenzel|manfred.stede@uni-potsdam.de

Abstract

In this paper, we are testing sentence alignment on complex, semi-parallel corpora, i.e., different versions of the same text that have been altered to some extent. We evaluate two hypotheses: To make alignment algorithms more efficient, we test the hypothesis that matching pairs can be found in the immediate vicinity of the source sentence and that it is sufficient to search for paraphrases in a 'context window'. To improve the alignment quality on complex, semi-parallel texts, we test the implementation of a segmentation into Elementary Discourse Units (EDUs) in order to make more precise alignments at this level. Since EDUs are the smallest possible unit for communicating a full proposition, we assume that aligning at this level can improve the overall quality. Both hypotheses are tested and validated with several embedding models on varying degrees of parallel German datasets. The advantages and disadvantages of the different approaches are presented, and our next steps are outlined.

1 Introduction

The task of sentence alignment originated in the context of machine translation, as the preparatory step for word or phrase alignment, which eventually informed bilingual translation models. In this paper, we address the somewhat different scenario of monolingual semi-parallel text, i.e., different versions of the same text. A well-known case is that of simplified language, where a text in standard language has been mapped to a text that is easier to process for audiences with limited knowledge of the language or people with cognitive or intellectual disabilities. In addition to this, we study two other settings that to our knowledge have not received attention yet. The first are sets of different biographic encyclopedia articles on the same person (authors from the former German Democratic Republic). The second is a specific use case from the Social Sciences, viz. the writings of the philosopher Hannah Arendt, who frequently published second (edited) versions of her works. All our data is in German, but our methods are in principle language-neutral and can be adapted to other target languages, and also to multilingual alignment tasks.

These datasets are well-suited for our experiments for several reasons: First, they represent different levels of difficulty in terms of segmentation and alignment. While the plain-language data contains relatively short and concise sentences and the data is relatively parallel, Hannah Arendt's essays offer significantly greater challenges; they are more heavily altered and the syntactic complexity is greater. The encyclopedia entries represent a special case, as some of the texts are written in terse style, often avoiding full clauses. However, in terms of content they are less parallel than the plain-language texts and, therefore, form an interesting complement.

In this paper we test two hypotheses:

- Matching pairs of text units should be found in similar positions in the two text versions, and it should therefore be sufficient to search for paraphrases in a predefined 'context window'. This approach should make the alignment models more efficient and could even improve alignment quality.
- Complex, heavily-altered sentences can be difficult to align, because only parts of the sentences are matching. Therefore, alignment quality should be improved by aligning on the (often sub-sentential) level of Elementary Discourse Units instead of sentence level. We expect this effect to be greater on complex data like the Arendt essays than on simple data like the plain-language texts.

The paper is structured as follows: In Section 2, we first describe related work for the most impor-

Proceedings of LaTeCH-CLfL 2025, pages 87–96

tant concepts of this paper - the notion of semi-parallel texts, EDU segmentation and sentence alignment. In Section 3, we present our datasets in detail. We provide content descriptions in Section 3.1 and corpus statistics in Section 3.2. In Section 3.3 we describe the process and the results of our manual annotation study. In Section 4, we explain methods and results of our experiments - separately for the topics of segmentation, embedding and alignment. Section 5 provides a qualitative error analysis, and Section 6 summarizes our conclusions and describes next steps.

2 Background & Related Work

2.1 Semi-parallel texts

The term 'parallel corpora' originates from research on statistical machine translation (SMT), where parallel texts were generally understood as direct translations into another language (Wołk and Marasek, 2017). However, parallel and non-parallel texts are difficult to clearly distinguish from each other; instead, it is often seen as a scale of 'comparable' corpora (Cheung and Fung, 2004). Such comparable texts have long been the subject of research, with most work focusing on the extraction of parallel sentences from these corpora (e.g., Tillmann (2009); Rauf and Schwenk (2011); Smith et al. (2010); Chu et al. (2013)). These papers use the term *quasi-comparable* texts for loosely related texts that can be written on the same topic or on different topics (Cheung and Fung, 2004).

In addition, research on paraphrase detection and paraphrase generation is also relevant for our work on semi-parallel text versions. Paraphrases map possibilities to change sentences on a lexical, morphological or syntactic level without affecting the meaning (Wahle et al., 2023). Many works have already been published on both paraphrase detection (e.g., Gold et al. (2019); Liu and Soh (2022)) and paraphrase generation (e.g., Bandel et al. (2022); Yang et al. (2022)). Paraphrases are also analyzed as a phenomenon of intertextuality in the context of digital humanities (e.g., Sier and Wöckener-Gade (2019)).

Our definition of semi-parallel texts is based on this research, but for the purposes of this paper we refer only to monolingual text variants. These are texts that are more or less closely related to each other and deal with the same topics. They may be texts that have been reformulated by the author for different audiences, written by different authors on the same topic, or simplified in order to be accessible to more people. In any case, due to their high similarity of content it should be possible to compute a meaningful alignment.

2.2 EDU Segmentation

The notion of 'Elementary Discourse Unit' (EDU) originated in the field of discourse parsing, especially in the tradition of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), where a text is first divided into EDUs, which are then recursively connected to each other via coherence relations (Cause, Contrast, Elaboration, etc.). Intuitively, an EDU is an independent clause or an adjunct clause that makes a complete contribution to the discourse; specific annotation guidelines then typically describe language-specific syntactic criteria. To illustrate, the sentence in Example 1 consists of two EDUs, while the matrix and complement clauses in Example 2 do not constitute two independent contributions:

(1) [This novel reads well,] [though it is a bit too long.]

(2) [In the bookshop I was told that this novel is a bit too long.]

RST parsers thus contain a segmentation component, but the notion of EDU is relevant also for other tasks. An early stand-alone segmenter for English, built on top of a syntactic parser, was SLSeg (Tofiloski et al., 2009). A more recent approach using a BiLSTM-CRF approach is NeuralEDUSeg (Wang et al., 2018). For German, a syntax-oriented approach was implemented by Sidarenka et al. (2015), who utilized a constituent and a dependency parser for two variants of a segmentation module. Recently, a few multilingual models have been built as part of a shared task (Braud et al., 2023).

The training data situation for German has very recently improved with the introduction of a new RST-annotated corpus (Shahmohammadi and Stede, 2024). For our work, we thus use their RST parser and extract from its output the sequence of EDUs computed for an input text.

2.3 Sentence Alignment and Evaluation

Sentence alignment is the task of matching sentences of two text versions that have the greatest semantic similarity. Early sentence aligners initially used scoring functions that only compared

the number of words or characters, because they assumed strong parallelism (Brown et al., 1991; Gale and Church, 1993). In later work (e.g. Moore (2002)) also lexical features and heuristics were used to improve speed and alignment quality. For example, LERA (Pöckelmann et al., 2022) models the alignment problem in a graph theoretic way and makes the alignment decision with a distance function based on the Jaccard index (Jaccard, 1901).

Sentence alignment algorithms are usually applied to bilingual, parallel texts. The use of machine translation (MT) methods to convert both texts into a common language was therefore widespread. For example, Sennrich and Volk (2010) use the BLEU score to carry out alignments in machine-translated texts.

Since the introduction of BERT by Devlin et al. (2019), the use of sentence embeddings has become increasingly established in this field of research. Reimers and Gurevych (2019) improved the computation of sentence embeddings with their Sentence-BERT (SBERT) model, reducing the enormous computational effort of the classical BERT model.

Embedding vectors can then be compared using classical similarity calculations such as the cosine similarity or the Euclidean distance. One of the first papers to implement this approach to sentence alignment was VecAlign (Thompson and Koehn, 2019). Both VecAlign and SentAlign (Steingrimsson et al., 2023) are based on bilingual sentence representations such as LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2022).

Recently, Molfese et al. (2024) introduced CroCoAlign - an algorithm that, in contrast to the models mentioned so far, incorporates more contextual information for disambiguating possible sentence mappings.

3 Data and Manual Annotation of Alignment

In this section, we first describe the three sources of data that we are using and how we constructed the corpora; this includes segmenting texts into EDUs. Then we report on our inter-annotator agreement study on the alignment task.

3.1 The Datasets

Hannah Arendt essays: In our experiment, we aligned two different versions of the essay on Franz Kafka: firstly, the original version 'Franz Kafka', which appeared in 1948 in the publication 'Sechs Essays' (*Six Essays*), and secondly, a radio broadcast entitled 'Franz Kafka - von Neuem gewürdigt' (*Franz Kafka – newly appreciated*), which was also published in 1948. The essays are part of the Hannah Arendt Edition, a digital, open-access edition that is hosted by Freie Universität Berlin.[1]

GDR literature encyclopedias: This dataset consists of encyclopedia entries on authors from the GDR. Two entries on the same person were manually selected from a larger dataset, but from different encyclopedias. For selecting the articles, particular attention was paid to finding entries that were as detailed as possible and ideally written in complete sentences, even though this was not possible for all entries. In all cases, one Wikipedia article was used as the reference text, with the second entry coming from different encyclopedias.

Plain language dataset: The third dataset consists of news reports, each of which is available in an original and a simplified version. The dataset was originally created to train models for text simplification tasks. Although the texts are closely related, it is possible that information has been lost during the simplification process or that the grammatical structure has been changed. This data is part of the APA-RST dataset (Hewett, 2023).

3.2 Corpus Statistics

As the three datasets come from different genres, they are structured differently and each present their own challenges. In two of the three cases, the data is available both in full sentences and in EDUs; however, many of the GDR encyclopedias were not initially written in sentence form and therefore EDU segmentation was not possible in this case. Detailed corpus statistics are listed in Table 1.

Hannah Arendt essays: This dataset includes two variants of the essay 'Franz Kafka'. The original version is slightly longer (36 sentences) than the radio broadcast and also features longer sentences - this is likely due to the change in target audience. In comparison to the other datasets at hand, the essays from Hannah Arendt provide the longest and most complex sentences with an average of around 30 words per sentence.

GDR literature encyclopedias: This dataset consists of encyclopedia articles about 61 authors

[1] https://hannah-arendt-edition.net/home?lang=en

from the GDR. The alignment is performed between the Wikipedia article and one other encyclopedia entry about this person, so the dataset consists of 122 documents in total. Since some of the encyclopedia entries were written with heavily-abbreviated sentences, this dataset is well-suited to test the performance of the alignment models at sub-sentence level, but it cannot be used to compare it for EDUs and whole sentences.

Plain language dataset: This dataset consists of 449 different news reports, each of which is available in the original and simplified version. In addition, we segmented both versions of the 449 reports into EDUs. In contrast to the essays by Hannah Arendt, the sentences are shorter and less complex; in many cases they cannot be segmented into more than one EDU.

3.3 Manual Annotation

Samples of all three datasets were selected for manual annotation of sentence alignment. Since context is an important factor for alignment decisions, documents were randomly selected for manual annotation rather than sentences.

Two annotators worked on the study. Both are students of Computational Linguistics and therefore trained in the linguistic characteristics of texts and their computational processing. The annotators were given guidelines for manual annotation. These guidelines specified that the basis for alignment must always be semantic similarity rather than surface form. It was specified that multiple alignments of the same element should only be made in justified exceptions and that, in contrast, there is no obligation to align all elements. Following these guidelines, the following alignment patters are allowed: `[1:0, 0:1, 1:n, n:1]`. However, `[n:m]` alignments are not possible.

To create a gold standard, the main annotator labeled encyclopedia entries on 11 different authors, 11 different newspaper reports from the plain language dataset, and the essays on Franz Kafka. To measure the inter-annotator agreement (IAA), the second annotator also processed almost half of this data. IAA for all datasets and additional statistics of the manual annotation can be found in Table 1.

Datasets:	Arendt	GDR-Data	Plain-language
Corpus Statistics			
Documents	2	122	898
Sentences	402	-	43,255
Segments	1,036	1,745	48,282
Words	12,323	17,571	440,000
Avg. Segments / Sentence	2.575	-	1.12
Avg. Words / Sentence	30.45	-	10.2
Avg. Words / Segment	11.825	10.06	9.1
Results of Manual Annotation			
Total: Aligned Sentences	402 (201)	-	568 (207)
Total: Aligned Segments	1,036 (512)	194 (88)	648 (237)
Cohen's Kappa: Sentences	0.772	-	0.917
Cohen's Kappa: Segments	0.843	0.785	0.909
Non-aligned Sentences	24.6% (28.97%)	-	25.6% (22.7%)
Non-aligned Segments	40.1% (47.8%)	52% (47%)	24.9% (24.9%)

Table 1: Statistics for all three corpora and results of the manual annotation.

4 Experiments and Results

4.1 Methods

Next, we describe our methods separately for segmentation, embedding and alignment.

4.1.1 Segmentation

Our alignment procedure should make it possible to carry out alignments both at sentence level and at EDU level. The first step in our pipeline is therefore the EDU segmentation of sentences. This step requires language-specific models, which are rare, especially for German. For the work described here, we used a modified version of the DPLP parser (Ji and Eisenstein, 2014), which was trained by Shahmohammadi and Stede (2024) on a corpus covering three different genres (blog posts, news, commentary). The parser produces complete RST trees from which the EDUs are then extracted.

4.1.2 Embedding

In order to process large texts efficiently, we use sentence embeddings for the numerical represen-

tation of language data. As we work exclusively with German data, we require embedding models that can process German texts. Several monolingual and multilingual models are suitable for this purpose. Furthermore, there are major differences in our data in terms of sentence length and grammatical complexity. We need embeddings that can process long, convoluted sentences from Hannah Arendt's essays as well as short EDUs and keyword-like entries from the lexicon articles.

Since, to our knowledge, there are no models that have been explicitly trained on EDUs, we tried out various embedding models on our test data and selected the following two models for the final experiments:

- `T-Systems-onsite/cross-en-de-roberta-sentence-transformer`: This is an `xlm-roberta-base` model (Conneau et al., 2019) that was fine-tuned by Philip May on the STSbenchmark dataset for processing English and German texts.
- `paraphrase-multilingual-mpnet-base-v2` This is a multilingual sentence-BERT model for STS tasks, trained on parallel data for more than 50 languages (Reimers and Gurevych, 2019).

We tested both embedding models in all runs, but since the RoBERTa model led consistently to better results, we decided to omit the second model for this task.

4.1.3 Alignment

We are also testing two different approaches for the automatic alignment of embeddings; one considers all possible unit pairs, the other reduces the candidate set. We cannot use existing alignment algorithms such as VecAlign (Thompson and Koehn, 2019) or SentAlign (Steingrimsson et al., 2023), since these approaches are designed to align parallel texts and cannot produce mappings that violate the parallel sentence ordering (for illustration, see the crossing lines in Figures 1 and 2).

The first approach uses the `paraphrase mining` function from the `Sentence Transformers` module (Reimers and Gurevych, 2019). It takes a list of strings as input and calculates sentence embeddings from them. The embedding model required for this can be defined manually. The function then uses cosine similarity to calculate the semantic similarity of all possible pairs of elements of the input. Finally, it outputs one or more possible matches for each element, sorted in descending order of cosine similarity. The function also offers the option to use other measurement units instead of cosine similarity to determine the similarity.

We generate several possible matches for each element and use a customized function to calculate the final alignments from there. This function is designed in a way that we have several adjustment options for fine-tuning. For example, we can specify that two elements should only be aligned if a certain cosine similarity is exceeded. We can also use a binary parameter to determine whether the same element may be aligned multiple times or not.

Our second approach is based on the assumption that the best matches of a sentence are to be found in an adjacent part of the second text version, i.e. that the index positions of the matched sentences are close. We have therefore developed a customized function that iterates over the first text version and searches for possible alignments in a neighboring section of the second text. The advantage of this approach is the reduced requirements in terms of computing power and time, as only the similarity to a few possible matches has to be calculated for each sentence.

The function is designed in a way that a threshold for alignments can be defined here as well. It can also be determined whether multiple alignments should be permitted and the size of the context window can be varied. Finally, the model for calculating the embeddings and the distance measure for determining the semantic similarity can also be specified using optional parameters. These setting options are intended to ensure that the algorithm can be flexibly adapted to the requirements of the different datasets at hand.

In Table 2, we list the fine-tuning settings of both approaches in the 'settings'-column.

4.2 Results

4.2.1 Alignment Algorithms

Comparing the results of the two alignment functions shown in Table 2, the context window generally performs better. For all datasets except the EDU-segmented plain language data, the context window leads to better alignments - on average approx. 0.4 higher Cohen's Kappa. In the case of EDU-segmented plain language data, however, the paraphrase function achieves a Cohen's Kappa that is approx. 0.4 higher.

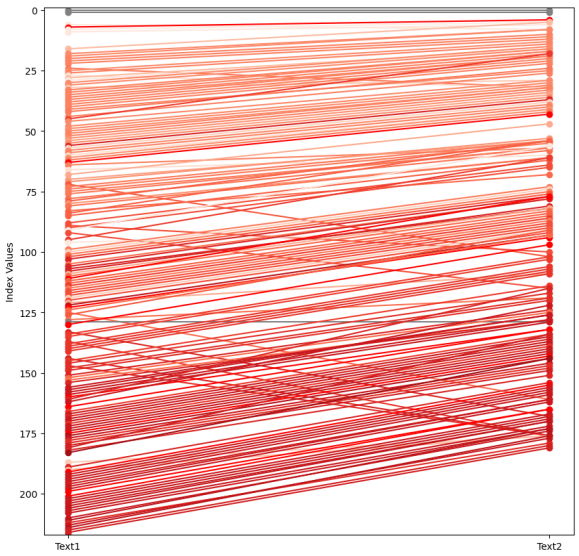

Figure 1: Alignments of Arendt essays on sentence level

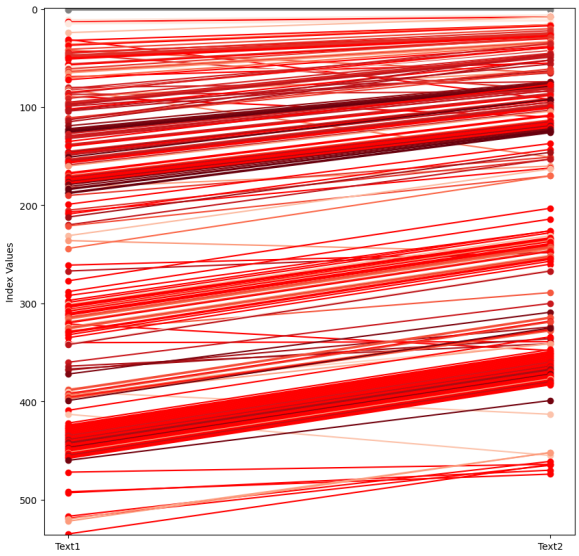

Figure 2: Alignments of Arendt essays on EDU level

In addition, the context window is also faster in all cases, on average approx. 14%. The biggest difference in terms of calculation time is for the encyclopedia entries, where the context window takes 36% less time than the paraphrase mining function. The smallest difference, on the other hand, is for the Kafka essays at sentence level - here the context window is only around 3% faster. These results correlate with the size of the context window, which in turn result from the properties of the datasets. If the texts are short or highly parallel, a small context window is sufficient to find the correct matches and the function can save a lot of time. With the long and heavily modified Kafka essays, on the other hand, much larger context windows are required to achieve good alignments and the efficiency advantage of the function shrinks accordingly. This can also be seen from the visualizations of the alignment throughout the Kafka essays in Figures 1 and 2. Since the text versions are of different length and some parts are heavily altered, the gap between aligned sentences is bigger (indicated by darker colors).

In most cases, the EDU-segmented texts can also be aligned faster than whole sentences. On average, however, the difference is smaller than the difference between the two alignment functions. It is particularly surprising that the EDUs also have an advantage with the paraphrase function, as significantly more elements have to be compared with each other at this level. However, it seems to be possible to calculate the embeddings of EDUs much faster, which results in an overall runtime advantage.

4.2.2 Alignment Level

A direct comparison between the alignment of EDUs and whole sentences (see Table 2) shows that the models achieve slightly better results on the sentence level than on EDU level, both for the Arendt data and for the plain language data.

Plain language dataset: The more sensitive RoBERTa model achieves better scores than the `sbert` model in all runs. The best run is achieved on the sentence level: With the embeddings of the RoBERTa model and a relatively high threshold of 0.55 cosine, a Cohen's Kappa of 0.76 is achieved between the manual alignment and the automatic alignment. The F1 score is also 0.76 in this case. If EDUs are used for alignment instead, the values across all runs are approx. 0.1 points below the runs with whole sentences. In the best run (RoBERTa, again 0.55 cosine), 0.65 Cohen's kappa is achieved.

Hannah Arendt essays: A similar picture emerges for this text pair: Full sentences again lead to better scores than EDUs. The differences between the various settings are therefore more evident here. However, the overall best performances - both for EDUs and for whole sentences - are again achieved with the RoBERTa embeddings and the threshold of 0.55 cosine. The Cohen's kappa here is 0.65 for whole sentences and 0.49 for EDUs.

GDR literature encyclopedias: Although this data is only available in a keyword-like form, good

Dataset	Alignment Algorithm	Settings	Prec	Recall	F1	Kappa	Computing time
Plain Lang - Sents	Context Window	model 1, Threshold: 0.55, multi-align: True, Window-size: 10	0.902	0.762	0.764	**0.761**	102.3 Sec
	Paraphrase Function	model 1, Threshold: 0.4, multi-align: True	0.898	0.743	0.740	0.731	117.2 Sec
Plain Lang - EDUs	Context Window	model 1, Threshold: 0.55, multi-align: True, Window-size: 20	0.839	0.673	0.671	0.650	**93.3** Sec
	Paraphrase Function	model 1, Threshold: 0.6, multi-align: True	0.817	0.699	0.684	0.693	97.4 Sec
Kafka - Sents	Context Window	model 1, Threshold: 0.55, multi-align: True, Window-size: 50	0.816	0.684	0.648	**0.656**	104.2 Sec
	Paraphrase Function	model 1, Threshold: 0.55, multi-align: True	0.780	0.657	0.623	0.630	107.3 Sec
Kafka - EDUs	Context Window	model 1, Threshold: 0.55, multi-align: True, Window-size: 75	0.805	0.695	0.655	0.589	**100.7** Sec
	Paraphrase Function	model 1, Threshold 0.6, multi-align: True	0.802	0.572	0.609	0.578	116.4 Sec
Encyclopedias	Context Window	model 1, Threshold 0.55, multi-align: True, Window-size: 10	0.888	0.786	0.740	**0.697**	**61.0** Sec
	Paraphrase Function	model 1, Threshold 0.6, multi-align: True	0.860	0.672	0.660	0.600	83.7 Sec

Table 2: Best overall results for different datasets and alignment algorithms.

alignment results are generated here in various runs with a Cohen's Kappa of up to 0.7.

The results also show that in general it seems to work better to allow aligning the same elements multiple times and combining this setting with a cosine threshold. In all runs this led to better results than restricting multiple alignments and removing the threshold.

5 Error Analysis

The most severe difficulties arise for the Arendt essays. When the aligned sentences are examined more closely, it becomes clear that an incorrect assignment may have been made, even though the matched sentences generally fit together well thematically. 'Meaning-heavy' terms like names and nouns, which have a great influence on the sentence embeddings, occur repeatedly throughout the dataset and thus make correct assignment more difficult. Therefore, part of the problem is that the embeddings of such complex sentences are not fine-grained enough to select the actual correct sentence from several potentially-matching sentences. This phenomenon can be observed in the following examples (English translations created by us, not by Hannah Arendt):

(3) [Original] **Das gemeinsame Erlebnis der Leser Kafkas ist eine allgemeine, unbestimmbare Bezauberung [...], eine klare Erinnerung an merkwürdige und scheinbar unsinnige Bilder und Beschreibungen – bis sich ihnen eines Tages der verborgene Sinn mit der plötzlichen Deutlichkeit einer einfachen und unangreifbaren Wahrheit enthüllt.**

The common experience of Kafka's readers is a general, indefinable enchantment [...], a clear memory of strange and seemingly nonsensical images and descriptions – until one day the hidden meaning is revealed to them with the sudden clarity of a simple and unassailable truth.

(4) [Found match] **Das einzige, was den Leser in Kafkas Werk lockt und verlockt, ist die Wahrheit selbst, und diese Verlockung ist Kafka in seiner stillosen Vollkommenheit geglückt, daß seine Geschichten auch dann in Bann schlagen, wenn der Leser ihren eigentlichen Wahrheitsgehalt erst einmal**

nicht begreift.

The only thing that lures and entices the reader in Kafka's work is the truth itself, and Kafka succeeded in this enticement with such quiet perfection that his stories cast a spell even if the reader does not at first grasp their actual truthfulness.

(5) [Correct match] **Kafkas eigentliche Kunst besteht darin, daß der Leser eine unbestimmte, vage Faszination, die sich mit der unausweichlich klaren Erinnerung an bestimmte, erst scheinbar sinnlose Bilder und Begebenheiten paart, [...] aushält, bis sich die wahre Bedeutung der Geschichte sich enthüllt.**

Kafka's real art lies in the fact that the reader endures an indeterminate, vague fascination, which is coupled with the inescapably clear memory of certain images and events that at first appear to make no sense [...] until the true meaning of the story is revealed.

As already mentioned in the last section, the alignment of the Arendt texts is made more difficult by the fact that the texts were also heavily altered at paragraph level. Parts were added or omitted and the sentence order was changed considerably. These characteristics make it very difficult (in particular for the context window) to find the correct correspondences, as the window size would have to be increased significantly and the efficiency advantages of this approach would be lost as a result.

6 Conclusion

The context window shows a superior performance compared to the paraphrase mining function both for alignment quality and alignment speed. However, there are still problems to be solved: If text versions are altered heavily, the window size has to be extended to find the best match. To mitigate this, a previous paragraph alignment could be implemented and the sentence alignment could be performed in a second step.

The role of EDU segmentation is difficult to assess. The use of EDUs in the alignment process can only make sense if the sentences are so long on average that several EDUs are created. However, even the experiments on the Hannah Arendt data showed that the models achieve slightly worse values on average with EDUs than with whole sentences. Several aspects should be considered here: Of all the data we worked with, Hannah Arendt's essays are by far the least parallel and therefore the most difficult to align. This can already be seen from the proportion of unaligned items in the test data: While less than 25% of the data in the plain language dataset was not aligned, this proportion is more than 35% in the EDUs of the Arendt essays. In addition, alignment is made very difficult by the length of the texts. Both the news reports of the plain language dataset (46 segments on average) and the encyclopedia entries (14 segments on average) are short, and correspondences are to be expected in the immediate vicinity. Arendt's essays, on the other hand, consist of more than 500 segments. They were restructured on paragraph level and are of different lengths.

In contrast, both models achieve very good values on the encyclopedia data, with Kappa scores ranging from 0.65 to 0.75. This also shows that the generally poorer scores of the Arendt essays are due more to the difficulty of the dataset than to the problems caused by the use of EDUs in the alignment process.

In order to solve the problems described here, several tasks must be tackled in the next steps: To conclusively evaluate the usability of EDUs for the alignment of complex, semi-parallel texts, further data should be included, which to some extent form a compromise of the datasets available here: They should be longer and more complex than the plain language and encyclopedia data, but more similarly structured on the textual level than the Arendt essays. In addition, further models for EDU segmentation should be considered, which may also be fine-tuned on the data available.

Also, it is necessary to thoroughly check the quality of the sentence embeddings. It has been discovered as part of the problem that the embeddings cannot clearly distinguish similar but non-identical phrases. A study that specifically measures the similarity of exchanged words, sentence structures and paraphrases could help to develop more precise embeddings for this use case.

Finally, a previous paragraph alignment should be tested to mitigate the fact that increased window sizes are necessary to combat alterations on paragraph level. With these additions, it should be possible to further improve sentence alignment on semi-parallel datasets.

Acknowledgments

We thank our student assistants Dietmar Benndorf and Maximilian Krupop for annotating training data, and we are grateful to the anonymous re-

viewers for their helpful feedback. Data from GDR literature encyclopedias was provided by the research group "Forschungsplattform Literarisches Feld DDR".[2] Our work is supported by the Deutsche Forschungsgemeinschaft (DFG), project (524057241) "Semi-automatische Kollationierung verschiedensprachiger Fassungen eines Textes“.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, ACL '91, page 169–176, USA. Association for Computational Linguistics.

Percy Cheung and Pascale Fung. 2004. Sentence alignment in paralllel, comparable and quasi-comparable corpora.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2013. Chinese–Japanese parallel sentence extraction from quasi–comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 34–42, Sofia, Bulgaria. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Darina Gold, Venelin Kovatchev, and Torsten Zesch. 2019. Annotating and analyzing the interactions between meaning relations. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36, Florence, Italy. Association for Computational Linguistics.

Freya Hewett. 2023. APA-RST: A text simplification corpus with RST annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 173–179, Toronto, Canada. Association for Computational Linguistics.

Paul Jaccard. 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, page 547–579.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Annual Meeting of the Association for Computational Linguistics*.

Timothy Liu and De Wen Soh. 2022. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601, Dublin, Ireland. Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281.

Francesco Molfese, Andrei Bejgu, Simone Tedeschi, Simone Conia, and Roberto Navigli. 2024. Crocoalign: A cross-lingual, context-aware and fully-neural sentence alignment system for long texts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2209–2220, St. Julian's, Malta. Association for Computational Linguistics.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA. Springer.

[2] www.ddr-literatur.de

Marcus Pöckelmann, André Medek, Jörg Ritter, and Paul Molitor. 2022. LERA—an interactive platform for synoptical representations of multiple text witnesses. *Digital Scholarship in the Humanities*, 38(1):330–346.

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*, 25(4):341–375.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Sara Shahmohammadi and Manfred Stede. 2024. Discourse parsing for German with new RST corpora. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 65–74, Vienna, Austria. Association for Computational Linguistics.

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. 2015. Discourse Segmentation of German Texts. *JLCL*, 30(1):71–98.

Kurt Sier and Eva Wöckener-Gade. 2019. Paraphrase als Ähnlichkeitsbeziehung. ein digitaler zugang zu einem intertextuellen phänomen. In *Platon Digital. Tradition und Rezeption*. Propylaeum.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Los Angeles, California. Association for Computational Linguistics.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Christoph Tillmann. 2009. A beam-search extraction algorithm for comparable data. In *Annual Meeting of the Association for Computational Linguistics*.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (Short Papers)*, pages 77–80, Suntec, Singapore. Association for Computational Linguistics.

Jan Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase types for generation and detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 12148–12164. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Krzysztof Wołk and Krzysztof Marasek. 2017. Unsupervised construction of quasi-comparable corpora and probing for parallel textual data. In *Multimedia and Network Information Systems*, pages 307–320, Cham. Springer International Publishing.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Haibo Zhang, Xue Zhao, Wenqing Yao, and Boxing Chen. 2022. GCPG: A general framework for controllable paraphrase generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4035–4047, Dublin, Ireland. Association for Computational Linguistics.