Segmentation of Argumentative Texts by Key Statements for Argument Mining from the Web

Ines Zelch[1,2] **Matthias Hagen**[1] **Benno Stein**[3] **Johannes Kiesel**[4]

[1]Friedrich-Schiller-Universität Jena [2]Leipzig University [3]Bauhaus-Universität Weimar
[4]GESIS – Leibniz Institute for the Social Sciences

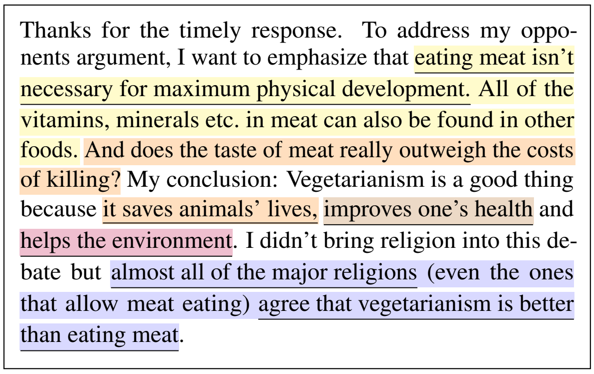
Abstract

Argument mining is the task of identifying the argument structure of a text: claims, premises, support/attack relations, etc. However, determining the complete argument structure can be quite involved, especially for unpolished texts from online forums, while for many applications the identification of argumentative key statements would suffice (e.g., for argument search). To this end, we introduce and investigate the new task of segmenting an argumentative text by its key statements. We formalize the task, create a first dataset from online communities, propose an evaluation scheme, and conduct a pilot study with several approaches. Interestingly, our experimental results indicate that none of the tested approaches (even LLM-based ones) can actually satisfactorily solve key statement segmentation yet.

1 Introduction

The field of argument mining deals with the identification and extraction of arguments from a text. A fundamental step in argument mining is text segmentation, which deals with the separation of different statements (argumentative discourse units) from each other (Stede and Schneider, 2019). When placed in relation to each other, the statements form a tree structure in which the root node represents the topic statement of the text and the nodes of the first level correspond to the main statements on this topic (Lawrence et al., 2014). Other nodes in the tree are, for example, explanations and examples that support their parent node.

But identifying the segments within a text is challenging, mainly because the exact segment boundaries are often up for debate (Pevzner and Hearst, 2002; Ghosh et al., 2014). Natural language texts, especially in debate forums and other argumentation-rich media, are rarely comprised of clear-cut and well arranged statements. Ambiguities, interjections, digressions and other factors prevent the establishment of general rules for the delimitation of segments. However, many argument analysis tasks do not require to identify the fine-grained argument structure of a text. To know what an argumentative text is about, it is usually sufficient to know its main points, which are thus typically the elements of interest in applications such as key point generation (Bar-Haim et al., 2020a,b), argumentation summarization (Syed et al., 2020) or argument search (Wachsmuth et al., 2017).



Thanks for the timely response. To address my opponents argument, I want to emphasize that eating meat isn't necessary for maximum physical development. All of the vitamins, minerals etc. in meat can also be found in other foods. And does the taste of meat really outweigh the costs of killing? My conclusion: Vegetarianism is a good thing because it saves animals' lives, improves one's health and helps the environment. I didn't bring religion into this debate but almost all of the major religions (even the ones that allow meat eating) agree that vegetarianism is better than eating meat.

Figure 1: Argumentative text on *vegetarianism*, highlighting key statements (underlined) and optional supplements of segments (colored).

So far, main points have mostly been gathered by crowdsourcing (e.g., Misra et al. (2015); Friedman et al. (2021)). For an automated extraction of the main points of a text, we propose an alternative and coarser-grained segmentation task that aims to separate the *key statements*—the level-one nodes in the argument tree—from each other. To model the ambiguity of boundaries, we only require a segment to (1) completely cover a single key statement and (2) not overlap with other key statements. Other contents can be included in the segment, but do not have to be. To illustrate the segmentation goal, Figure 1 shows examples of key statements (underlined) and text passages that could be included in a segment (colored text without underlining). To address this new task on realistic web data, we

Proceedings of the 12th Argument Mining Workshop, pages 228–242
July 31, 2025 ©2025 Association for Computational Linguistics

use the args.me corpus, which provides texts on different controversial topics crawled from four different debate portals (Ajjour et al., 2019b). We apply a range of segmentation approaches, including simple sentence and paragraph segmenters, two previous models for argument unit segmentation, and two different LLMs.[1]

Section 2 provides an overview of the challenges of the segmentation task and presents previous segment approaches. Section 3 defines key statements against the background of different terms of argumentative text units and formalizes the coarse-grained segmentation task. Section 4 outlines the evaluation procedure and Section 5 presents the corresponding results. Amongst others, we find that a segmentation by paragraphs provides a good baseline and LLMs achieve the best results. The predictions of the previous argument unit segmentation approaches are often too short to be useful.

2 Background

This section outlines various challenges of argument segmentation and introduces a selection of approaches that tackle this task.

Challenges Some challenges arise due to the overall distribution of arguments in texts. A segmentation at sentence boundaries is usually not sufficient, as multiple propositions might be contained in a single sentence, or a proposition may stretch over more than one sentence (Stede and Schneider, 2019). Multiple arguments can enforce each other in so called compound argumentations (Palau and Moens, 2009; Stab and Gurevych, 2017) and have to be recognized as being part of the same segment. Further, segments can be embedded into another (Lawrence and Reed, 2020) so that they cannot be separated appropriately. Another challenge are implicit statements which are difficult to capture on text-level, for example enthymemes (implicit premises that are considered obvious), rhetorical questions or sarcasm (Lawrence and Reed, 2020; Trabelsi and Zaïane, 2019; Hasan and Ng, 2014). Sometimes, propositions require knowledge beyond the text span under consideration, such as back-references to (parts of) previous statements (Lawrence and Reed, 2020), or missing co-references that have to be resolved. Finally, there is the problem of segment evaluation since humans often disagree on the exact boundaries, and the importance of different types of errors might depend on the application of the resulting segments (Pevzner and Hearst, 2002; Ghosh et al., 2014).

Related Work Approaches that tackle the argument segmentation task, usually process a text on either sentence or token level. On sentence level, argument segmentation is typically approached as classification task, labeling a sentence as argumentative (probably even more fine-granular, for example as claim or premise, pro or con) or as not argumentative (Reimers et al., 2019; Lippi and Torroni, 2015; Moens et al., 2007). On token level, several approaches are based on BiLSTM architectures or BERT, sometimes in combination with Conditional Random fields or other additional components (Fu et al., 2023; Alhindi and Ghosh, 2021; Trautmann et al., 2020; Chernodub et al., 2019). For example, Ajjour et al. (2017) use a BiLSTM model with different textual features including POS-tags, information about clauses, phrases, and sentences as well as a list of discourse markers. Others propose rule-based approaches (Fujii and Ishikawa, 2006) or use the parse tree representations of the sentences (Guilluy et al., 2023; Dumani et al., 2020; Persing and Ng, 2016). Recent approaches also use LLMs for the segmentation task (D'Agostino et al., 2024). All approaches have in common that they extract argument units with specific boundaries.

3 Conceptualizing Key Statement Segmentation

The extraction of argumentative text units has been addressed under varying terms and definitions and with different scope. Against this background, we formalize the segmentation task for our use case.

3.1 Defining Key Statements

In order to describe, illustrate, and categorize the concept of key statements of a text, we relate and contrast it with existing concepts.

Key statements are argument discourse units Stede and Schneider (2019) define an argumentative discourse unit (ADU) as a text segment "that plays a single role for the argument being analyzed, and is demarcated by neighboring text spans that play a different role, or none at all [for the argumentation.]" The nature of "role" can vary between analyses and ADUs can thus span multiple sentences, or be shorter than a sentence. ADUs are in this sense the argumentative counterpart to the

[1] Our code is available at https://github.com/webis-de/argmining25-argument-segmentation.

elementary discourse units (EDUs) in rhetorical structure theory (see Taboada and Mann (2006) for an introduction). Typically, ADUs are a text's statements (Lawrence and Reed, 2020) and can be seen as nodes in a tree in which the edges indicate support and attack relationships and with the topic statement as the tree's root node. In this view, key statements are a subset of all ADUs in a text, namely the ADUs at depth (or level) one, i.e., the children of the root node.

Key statements are linked to key points In argumentation, one distinguishes between the written or spoken words (statement) and their abstract meaning (proposition). Key statements are the salient statements in an argumentative text. As statements, they are linked to the abstract propositions a reader forms in their mind while comprehending the text. In this sense, the concept of "central propositions" of a text, introduced by Misra et al. (2015) and extracted by means of abstractive summarization, coincides with the propositions of key statements. Furthermore, the concept of "key points" of a topic, introduced by Bar-Haim et al. (2020a,b); Friedman et al. (2021), follows the same idea, but defines salience with respect to a topic—described by a collection of texts—and not single texts. Key statements can thus be used to infer central propositions and potential key points.

Key statements are not aspects Though seemingly related, key statements are not the salient parts of single statements or propositions. For example, typical aspect terms in statements on minimum wage increases are "job" and "economy," which indicate that the statements concern effects on the respective aspect (Trautmann, 2020). The same concept for segments longer than single words has been coined "point at issue" by Fujii and Ishikawa (2006). The corresponding concept for propositions has been coined "argument facet" by Misra et al. (2015). In contrast to these, key statements are complete statements.

3.2 Formalizing Key Statement Segmentation

Having defined key statements, we define the task of segmenting by key statements as follows:

> Given an argumentative text and the controversial topic it discusses, segment the text such that each segment contains exactly one key statement (as per the topic).

The most important difference compared to previous argument segmentation approaches is that this task definition allows segments to encompass more text than just the key statement. Instead of defining specific segment boundaries, we permit some variability in the segments in order to account for the ambiguity of the segmentation task, as explained in Section 1. Key statements only define a minimum set of ADUs that the segments must cover in terms of content, and they must be correctly separated from each other.

3.3 Segmentation Approaches

We apply a range of different approaches to the segmentation task. Based on the structural text features, we apply a segmentation at sentence boundaries using NLTK's sentence tokenizer, and at each new paragraph based on HTML tags (`<br>` and `<p>`). Additionally, we apply a re-implementation of the unit segmentation by Ajjour et al. (2017), and TARGER (Chernodub et al., 2019) which is usable via an API. Finally, we prompt PaLM and GPT-4 as representatives of LLMs.[2] The prompt used for the segmentation with PaLM (provided in Figure 4 in Appendix A) is derived from a prompt by Chen et al. (2024) for the segmentation of Wikipedia pages into propositions. For GPT-4, the prompt is subtly varied for better results. We also tested prompt optimization with DSPy (Khattab et al., 2022, 2023), but did not achieve further improvements by this. For all created segments, we automatically filter those with less than three whitespaces in order to reduce noise without potential argumentative content.[3] We further test the effect of filtering segments classified as non-argumentative by the model provided by Reimers et al. (2019),[4] since the applied approaches do not necessarily distinguish between argumentative and non-argumentative propositions of a text.

4 Developing an Evaluation Framework for Key Statement Segmentation

To evaluate the coverage of the key statements by predicted segments, we provide a test set with manually extracted key statements, and propose an automatic matching approach that assigns segments to semantically equivalent key statements (indepen-

[2] PaLM 2 and GPT-4o mini

[3] For example, for the sentence approach, this removes segments like "It's that simple.", "1.2 Contention 1" or links.

[4] github.com/UKPLab/acl2019-BERT-argument-classification-and-clustering

The given texts contain arguments on different controversial claims (“topics”). Your task is to annotate the text’s key statements.
A key statement is a minimal text passage that expresses exactly one proposition that directly supports or attacks the text’s topic. It can range from a single phrase to multiple sentences.
If a key statement is repeated in a similar way, annotate only the occurrence that expresses the statement best.

Figure 2: Instructions for the expert annotators.

Forced marriages are not supported theologically by any of the major religions.[ann2, ann3] Whilst different religions may disagree on the nature of marriage and its formation, all are agreed that some level of consent is necessary.[ann1] Forced marriage is no more than a barbaric tribal custom which has no place in a modern society.[ann2]

Figure 3: Example annotations for a text on the topic “We should abandon marriage.” The brackets show which annotators annotated each sentence as key statement. The first two of the three sentences are semantically very similar and which one to annotate as key statement is somewhat arbitrary.

dent of the specific segment boundaries). Furthermore, we introduce suitable evaluation categories and measures.

4.1 Compiling a Dataset for Key Statement Segmentation

In order to evaluate key statement segmentation approaches on relevant web data and analyze the relationship between the key statements and key points, we create a dataset by sampling texts from the args.me corpus of online discussion forums (Ajjour et al., 2019a). In this sampling process, we focus on texts discussing topics that are related to the topics in IBM’s Key Point Analysis Shared Task (Friedman et al., 2021). The dataset consists of 50 texts that comprise 1,263 sentences and 25,201 words in total,[5] and cover 14 different controversial IBM topics. We manually annotated the key statements of these texts, resulting in 147 ground truth segments, covering 204 sentences and 4,019 words (16%). Figure 2 shows the annotation instruction for our three expert annotators.[6] The dataset is available online.[7]

To analyze the ambiguity of the annotation task, ten argumentative texts were annotated independently by all three annotators.[8] A discussion between the annotators revealed a general agreement; a major ambiguity resides in semantically similar sentences that could all be selected as key statements. Due to this ambiguity, traditional measures for inter annotator agreement are unsuitable for the task. For illustration purposes, consider the situation in Figure 3, where it is somewhat arbitrary which of the first two sentences to choose as key statement. Still, Cohen’s Kappa would produce a negative score for annotators ann1 and ann3, who both agree that the last sentence is not a key statement. We thus performed a manual matching of annotated key segments between annotators and use the pairwise Jaccard index[9] to assess the agreement, resulting in medium to high scores between 0.47 and 0.80 macro-averaged across the ten texts. Given the ambiguity of the task, moderate agreement in argument segmentation studies is a common result (Habernal and Gurevych, 2017; Ghosh et al., 2014; Palau and Moens, 2008). The amount of words that the annotators marked as key statements is similar (between 41% and 51%). To complete our agreement assessment, Section 4.3 shows that evaluation results vary only slightly when switching between the annotations of the different annotators as ground truth. We thus conclude that a reliable annotation of key statements is possible, except for ambiguities induced by repetitions. For future datasets, one could consider to change the annotation instructions for repetitions (last sentence of Figure 2) to suggest the first occurrence instead of the commonly used but more ambiguous “best” occurrence to potentially reduce these ambiguities.

4.2 Matching Segments to Key Statements

To match predicted segments to ground truth key statements, we require an approach that goes beyond simple string matching for multiple reasons. Firstly, a text may contain paraphrases of the same statement. In such cases, all different formulations should be matched to the corresponding ground truth key statement (see, for example, the highlighted text snippets in Figure 5). Secondly, the ground truth key statements are not necessarily continuous text snippets. However, they should

[5] As per NLTK’s word and sentence tokenizer.

[6] Members of our research group

[7] Data: https://doi.org/10.5281/zenodo.14865977

[8] Using doccano (Nakayama et al., 2018)

[9] Jaccard index: $\frac{|\text{segments both annotated}|}{|\text{segments at least one annotated}|}$

still be matched to segments from extractive approaches, such as sentence or paragraph segmentation. Thirdly, using LLMs often results in segments that are not strictly extractive, but these should still be matched to the key statement that is semantically most similar. For example, a ground truth key statement "Prostitution and recreational drugs are totally different: with prostitution you are not really harming anyone and recreational drugs can have a negative effect on people" should be matched with the LLM-generated segment "Prostitution is different from recreational drugs; it doesn't inherently harm others, unlike addictive drugs."

To match segments, we tested different similarity measures at various thresholds (skipped for brevity) against a human matching. We found that a combination of 3-gram overlap (threshold: 0.12), difflib's SequenceMatcher,[10] (threshold: 0.5), and an SBERT sentence transformer model[11] (threshold: 0.9) yields the best performance for PaLM segments and key statements, and outperforms each single measure: Counting it as a match if the similarity is above the threshold for at least one of the three measures, we reach a precision of 0.90, a recall of 0.79, and a very good F1 of 0.84.

4.3 Distinguishing Segment Match Categories

In order to distinguish between different kinds of mismatches between predicted segments and ground truth key statements, we derive different matching categories (Table 1). Key statements are *missed* if they are not covered at all, predicted segments without a corresponding key statement are *spurious*. A *match* between two segments can be *correct* or either *incomplete*, *impure* or *incomplete&impure*. The categories are illustrated with an example in Table 7 (Appendix).

To assess the correct matching category automatically, we build upon the segment matching and corresponding similarity scores. For each predicted segment (precision perspective), we count the number of ground truth segments to which it was matched. If this count is one, we consider the key statement to be correctly covered (*match*), if it is zero, the predicted segment falls in category *spurious*, and if it is greater than one, we assume that the prediction condenses multiple key statements into a single segment (*impure*). Similarly, we count the number of matches for each ground truth segment (recall perspective). Again, it is a *match* if this count is one, if it is zero, the ground truth segment is *missed* by the prediction; if it is greater than one, we assume that the ground truth segment is erroneously divided into multiple predicted segments and therefore only *incompletely* covered. An exception are predictions with a similarity > 0.9 to a key statement for either the SequenceMatcher or SBERT. They are considered as *match* rather than *incomplete*, to take into account that a statement can be repeated with different wording throughout a text, so that multiple matches would be possible. Segment pairs where both ground truth and prediction are matched multiple times are assigned an *incomplete&impure* label. In a strict evaluation, we only consider correct matches, whereas a relaxed evaluation comprises all matched segments (including incomplete and impure ones).

To further extend our assessment of inter annotator agreement, this time with respect to implications of disagreement on segmentation results, we evaluate each approach against the annotations of each annotator separately (cf. Section 4.1). Table 2 shows the mean and standard variation over annotators for relaxed precision, recall, and F1. As the low standard deviation indicates, evaluation results vary only slightly for different annotators, which shows a general agreement among annotators.

4.4 Measuring the Key Point Coverage

In order to assess how critical incomplete, impure and even missed segments are, we can estimate their effect on the end application (as described by Pevzner and Hearst (2002)), which will be the creation of key points in future work. It can be analyzed whether missed predictions lead to a complete loss of key points, or if they are still covered by other texts in the corpus. We therefore map the manually extracted key statements and predicted segments to the key points of the Key Point Analysis Shared Task 2021 (Friedman et al., 2021), which summarize the most important premises for a controversial topic (five pro, five con).[12] We apply the best matching approach (Alshomary et al., 2021) that participated in the shared task to map the segments to their most similar key point, but only if the calculated similarity is > 0.9. Key points covered by key statements should also be covered by the predicted segments.

[10] docs.python.org/3/library/difflib.html

[11] all-mpnet-base-v2, https://www.sbert.net/

[12] github.com/ibm/KPA_2021_shared_task

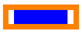
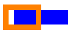
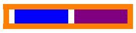
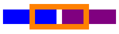
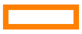
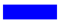
Category	Explanation	
Matched correct	A key statement is covered correctly by a prediction (additional text may be included)	
Incomplete	A key statement is covered partially or split into multiple segments	
Impure	Different key statements are merged into a single predicted segment	
Incomplete & impure	Different (incomplete) key statements are merged	
Spurious	A predicted segment matches no key statement (e.g., non-argumentative text, examples)	
Missed	A key statement is not covered by any predicted segment	

Table 1: Explanation of segment match categories. In the pictogram, blue and and purple rectangles illustrate key statements (ground truth), whereas orange boxes represent predicted segments. For calculating strict precision and recall, only matched correct segments are counted as true positives, whereas the relaxed measures also count incomplete and impure (and both combined) as such.

Measure	Approach					
	PaLM	GPT-4	Paragr.	Sent.	Ajjour	Targer
Precision	0.49±0.02	0.27±0.04	0.50±0.02	0.23±0.00	0.21±0.02	0.17±0.02
Recall	0.66±0.08	0.65±0.03	1.00±0.00	1.00±0.00	0.90±0.02	0.98±0.03
F1 strict	0.47±0.02	0.29±0.05	0.29±0.05	0.26±0.05	0.22±0.04	0.18±0.01
F1 relaxed	0.56±0.03	0.37±0.04	0.67±0.01	0.39±0.00	0.35±0.02	0.28±0.03

Table 2: (Micro-) average and standard deviation for (relaxed) precision, recall, and (strict and relaxed) F1 score calculated by evaluating each approach for each of the three ground truths (one per annotator).

	All		Filtered	
	Man.	Auto.	Man.	Auto.
matched | Precision	0.54	0.56	0.63	0.57
– correct	0.33	0.31	0.44	0.43
– incorrect	0.21	0.25	0.19	0.14
spurious	0.46	0.54	0.37	0.43
matched | Recall	0.84	0.74	0.64	0.59
– correct	0.64	0.58	0.52	0.50
– incorrect	0.20	0.16	0.12	0.09
missed	0.16	0.27	0.36	0.41
F1 micro strict	0.44	0.40	0.48	0.46
F1 micro relaxed	0.66	0.56	0.63	0.58

Table 3: Comparison of automatic (auto.) and manual (man.) matching results using PaLM segments, reporting micro average scores from a precision and recall perspective. Row 'incorrect' summarizes incomplete and impure matches, 'matched' covers all correct and incorrect segments.

5 Results

The evaluation results for the automatic matching approach as well as the effectiveness of the segmentation approaches are reported in the following.

5.1 Matching Categories

The automatic assignment of matching categories is evaluated using the PaLM segments. Table 3 compares the segmentation effectiveness based on the automatic matching procedure with the effectiveness based on a manual matching. The differences in the scores mainly result from a shift of correct or incomplete segments (manually labeled) to spurious or missed (automatically labeled). The overall precision and recall are very similar, the slightly better scores resulting from the manual assignments indicate that we do not erroneously improve the overall results by the automatic estimation of the categories.

5.2 Segmentation

Table 4 shows the effectiveness of the presented segmentation strategies, whose segments are automatically matched to the key statements. Afterwards, the matching categories are assigned automatically. The scores are reported on micro-level, averaging over all predicted and ground truth segments of all argumentative texts together. The upper half of the table presents a precision-oriented evaluation. The matched/precision row shows the relaxed proportion of predicted segments that are matched to a key statement, comprising correct, incomplete and impure segments; the remaining predictions are spurious. The recall-oriented results are similarly arranged in the second half of the table. The precision of the approaches can be improved at the cost of recall by filtering segments classified as non-argumentative by Reimers et al.'s

Measure	Approach								
	PaLM f.	**PaLM**	**GPT-4**	**Paragr.**	**Sent. f.**	**Sent.**	**Ajjour**	**Targer**	
# Segments	173	285	470	347	408	1125	1174	1759	
matched | Precision	0.57	0.46	0.28	0.42	0.38	0.22	0.17	0.14	
– correct	0.43	0.31	0.18	0.22	0.21	0.08	0.07	0.05	→
– incomplete	0.11	0.13	0.10	0.13	0.16	0.13	0.10	0.09	→
– impure	0.02	0.01	0.00	0.04	0.00	0.00	0.00	0.00	→
– incomplete & impure	0.01	0.01	0.00	0.03	0.01	0.01	0.00	0.00	→
spurious	0.43	0.54	0.71	0.58	0.62	0.78	0.82	0.86	→
matched | Recall	0.59	0.74	0.69	0.93	0.79	1.00	0.90	1.00	→
– correct	0.50	0.58	0.52	0.49	0.56	0.59	0.55	0.53	
– incomplete	0.05	0.10	0.14	0.12	0.21	0.38	0.34	0.45	→
– impure	0.03	0.03	0.02	0.25	0.01	0.01	0.00	0.01	→
– incomplete & impure	0.01	0.03	0.01	0.07	0.01	0.02	0.01	0.01	→
missed	0.41	0.27	0.30	0.07	0.20	0.01	0.10	0.00	→
F1 micro strict	0.46	0.40	0.27	0.30	0.31	0.14	0.12	0.09	
F1 micro relaxed	0.58	0.56	0.41	0.58	0.52	0.36	0.30	0.25	

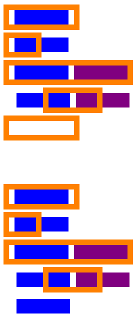


Table 4: Evaluation of different argument segmentation approaches from a precision- and recall-oriented perspective, reporting the micro average scores. The rows 'matched' cover all correct and incorrect segments. The total number of manual reference segments is 147. For PaLM and the sentence approach, the results after filtering (f.) are shown for comparison.

(2019) argument classifier. In Table 4, this is exemplarily shown for the segments created by PaLM and the sentence approach (columns 'PaLM f.' and 'Sent. f.'), the filtering results for all segmentation approaches can be found in Table 8 (Appendix).

Both sentence and paragraph segmentation approaches produce segments that cover almost the complete text (except the filtered short segments) and therefore have a high recall. Accordingly, a high number of spurious segments causes a low precision. The recall of the paragraph segmentation is not a hundred percent because the semantic matching approach may miss a key statement, for example if the segment contains much additional content. Also, the paragraph segments vary considerably in length (depending on the text formatting) and can thereby result in short spurious segments (see example in Table 10 in the Appendix). The two argument unit segmentation approaches, TARGER and Ajjour, create a high number of segments that are usually shorter than a sentence. The created segments are in some cases useful and succinct,[13] but in most of the cases not self-contained and not argumentative (e.g., "agree that would be absurd", "As for lives saved"), resulting in the lowest precision scores of all approaches. PaLM provides a better balance between precision and recall, produces considerably less segments than the sentence approach, TARGER and Ajjour, and does not return the complete text, like paragraph and sentence approach. GPT-4 has a lower precision and recall than for example the paragraph segments, however, it has a higher proportion of correct matches and therefore a comparable strict F1. Still, the segmentation with LLMs has room for improvement.

Measure	Approach			
	PaLM f.	**PaLM**	**GPT-4 f.**	**GPT-4**
matched | Pre.	0.67	0.58	0.53	0.47
– correct	0.32	0.24	0.19	0.14
– incomplete	0.32	0.31	0.34	0.33
– impure	0.03	0.02	0.00	0.00
– inc. & imp.	0.00	0.01	0.00	0.00
spurious	0.32	0.42	0.48	0.53
matched | Rec.	0.62	0.84	0.65	0.90
– correct	0.34	0.45	0.31	0.42
– incomplete	0.20	0.27	0.32	0.45
– impure	0.08	0.10	0.01	0.02
– inc. & imp.	0.00	0.02	0.01	0.01
missed	0.37	0.16	0.36	0.11
F1 micro strict	0.33	0.31	0.24	0.21
F1 micro relaxed	0.65	0.69	0.57	0.62

Table 5: Evaluation with fix boundaries for PaLM and GPT-4 segments (with filtering in columns 'f.').

To compare our evaluation setup with the traditional evaluation based on explicit segment boundaries, we map the key statements and LLM-created segments to the original text, to verify whether the key statements' boundaries are within the bound-

[13] for example, "DP sometimes kills innocents", "Violates the right to life" for the topic "abolish capital punishment"

aries of the predictions. Our evaluation shows that the results with explicit boundaries are less accurate. Key statements or segments by LLMs can consist of disconnected text passages (e.g., leaving out lengthy explanations). Mapping such segments to contiguous text passages to get the exact boundaries, can result in segments much longer than the original one. On the one hand, this produces longer key statements, which are more difficult to cover, on the other hand, it can cause predictions to cover additional content, so that key statements might be matched by mistake. The results in Table 5 show that the number of correctly matched segments is lower than with our proposed evaluation approach. At the same time, the number of incomplete matches is much higher, which results from the changed segment sizes. This leads to an improvement of the relaxed F1, while the strict F1 is clearly lower for the LLM-generated segments.

5.3 Key Point Coverage

Table 6 shows the key point coverage of the different segmentation approaches with the key points covered by the manual segments as reference. Since segmentation approaches with a high output number have a higher probability to cover all key points, we only report the numbers of the more reasonable approaches. Although the paragraph approach covers almost the complete text, it does not cover all key points. As before, this can most probably be explained by the potentially greater length of these segments which might prevent the matching model from a correct mapping, since the matching approach for key points (Alshomary et al., 2021) was trained on segments of sentence length. This might also explain why the filtering approach removes more relevant segments for paragraphs. The highest coverage of key points is achieved for segments from GPT-4, although it has lower F1 scores than PaLM and the paragraph approach. This indicates that an approach can provide a good overall coverage of argumentative propositions in a pool of texts, even if not every single key statement is covered. Also, the filtering approach works very well on the GPT-4 segments, as the key points coverage does not drop here. All in all, using the key points of the IBM shared task gives only an estimation of the covered key points. Ideally, we would generate new key points for the underlying data, however, this is beyond the scope of this work.

PaLM	**PaLM f.**	**GPT-4**	**GPT-4 f.**	**Paragr.**	**Paragr. f.**
0.74	0.70	0.87	0.87	0.70	0.57

Table 6: Coverage of key points for segments created by three approaches (with filtering in columns ’f.’), with key points covered by the key statements as reference.

5.4 Qualitative Analysis

Different challenges of the segmentation task that were described in Section 2 are also present in the texts from the args.me corpus. For example, rhetorical question express opinions only implicitly, like “should we ban Kentucky fried chicken because it too can be used as an instrument for terrorists ?” (topic “We should prohibit flag burning”). Another problem is the citation of counterarguments that can result in segments with the opposite stance. In the passage “*hate is the motivational force behind the burning*. Untrue”, the contrary proposition is indicated by a single negating word. A sentence segmenter can never capture this, but PaLM creates the following segment: “In the first round, my opponent claimed that hate is the motivational force behind flag burning. This is untrue. [. . .]”. Beyond that, LLMs allow to address most of the other segmentation challenges as well, for example, they can segment texts independently of sentence boundaries, make them self-contained, and filter non-argumentative content (illustrated in Figure 6 for the example text in Section 1). Other text attributes that require reformulations to obtain a meaningful segment are careless mistakes in writing that can distort the meaning and even turn a statement into the opposite, as in the following sentence: “Study shows that there there is not enough evidence to support the fact that the death penalty does not act as a deterrence.”, where the writer is actually arguing against the death penalty, but reverses the statement by adding a “not” too many. Implicit references to previous posts, cannot be resolved without further knowledge: “With the Cain and Abel story [. . .] the bible never said that they were the only people on the earth” (referring to a passage where his opponent argues with illogical parts of the bible). A downside of the use of LLMs are reformulations that change the original content. An example is the generated segment “The Dutch euthanasia’s have doubled since 1998.”, whereas the text originally states “The euthanasia’s in Belgium have doubled since 1998”. In our case, this only happens in exceptional cases and is therefore not further considered, but it is important to keep this possibility in mind. Moreover, LLMs have problems with argumentative chains (such as ”1: Every state of the universe is caused by another state. 2: If every state of the universe if caused by another state, then an initial state is logically impossible. 3: From 1 and 2, an initial state is logically impossible. 4: From 3, there can be no cause of the initial state. 5: According to the definition of god, god cannot exist.”) that are consistently separated rather than kept together.

6 Conclusion

In this work, we formalize the new task of text segmentation by key statements, the most salient argumentative statements in a text that form the basis for an abstract overview of the contained arguments. We provide detailed insights into the theoretical background of the task, and into its evaluation that takes the ambiguity of segmentation into consideration. Moreover, we demonstrate the suitability of the proposed coarse-grained segmentation approach for less structured web documents, such as discussion forums, and apply a range of segmenters of varying complexity. For the evaluation, we provide a first test set with human annotations of key statements in 50 texts. First experiments on this test set show that a segmentation by paragraphs represents a strong baseline for the task. While previous unit segmentation approaches result in a high number of very short segments, LLMs provide the most promising results so far. They additionally have the advantage that they allow subtle adaptations to the text which can be useful in order to tackle segmentation challenges, such as resolving missing co-references or formulating implicit statements more explicitly. All approaches benefit in terms of precision when applying an additional filtering step to remove non-argumentative segments. Our results suggest that a combination of different approaches, such as paragraphs and LLMs, could lead to even better results. Also, chain-of-thought prompting could further improve the effectiveness of LLMs on this task. Apart from that, we plan to investigate the usefulness of key statements for the generation of key points in future work.

7 Limitations

The presented setup for the automatic segmentation of an argumentative text by key statements entails different limitations which are important to consider. First of all, we propose a “stacked” evaluation approach where multiple steps are performed until the final results are available. Although all steps are evaluated, each step is a potential source of error. For example, the quality of the intermediate matching step (and optionally of the classification approach for filtering non-argumentative segments) influences the final effectiveness of the different segmentation approaches. The estimation of the key point coverage additionally relies on the approach for mapping segments and key points. The key point coverage as calculated in this paper, is additionally limited by the number of key points provided in the ArgKP 2021 dataset. Table 9 shows example segments where none of the provided key points is suitable. It is therefore desirable, to extend the existing key points in future work. Regarding the data base, it should be emphasized that working with unstructured documents from the web is always more challenging than with curated data, and that automatic analysis methods can only be applied to a limited extend. Finally, we are aware of the limited size of our test dataset. However, it covers a considerable range of texts with different levels of quality and structuredness, and is thus sufficient to demonstrate the concept of our proposed evaluation setup. Moreover, it can be extended for further evaluations.

8 Ethical Considerations

All annotators gave their consent to the use of their key statement annotations. They all have an academic background, but we collected no further demographic information as they are not relevant in our context, and could not be sufficiently anonymized for three people. Since no personal data were collected, an approval by an ethics review board was not necessary. The texts collected from the debate portals might contain harmful content, but we do not take responsibility for offensive content of any kind. All argumentative texts were processed by annotators and authors independent of their personal opinion on the expressed statements. As already noted, the use of LLMs for the segmentation task has the potential to distort the content, so their output should always be verified. All scientific artifacts used in this work are free to use for research purposes. This mainly concerns the args.me corpus (CC BY 4.0), data from the IBM shared task (Apache License 2.0), the argument classification model by Reimers et al.

(2019) (Apache License 2.0) and the TARGER API (MIT License). All artifacts are used in the context of argumentation mining and analysis which is consistent with their original use.

Acknowledgments

This publication has been partially supported by the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU), and by the German Federal Ministry of Education and Research (BMBF) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019a. Modeling Frames in Argumentation. In *Proceedings of EMNLP-IJCNLP 2019*, pages 2922–2932.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit Segmentation of Argumentative Texts. In *Proceedings of ArgMining@EMNLP 2017*, pages 118–128.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019b. Data Acquisition for Argument Search: The args.me Corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*, pages 48–59.

Tariq Alhindi and Debanjan Ghosh. 2021. "Sharks are not the threat humans are": Argument Component Segmentation in School Student Essays. In *Proceedings of BEA@EACL, 2021*, pages 210–222.

Milad Alshomary, Timon Gurcke, Shahbaz Syed, Philipp Heinisch, Maximilian Spliethöver, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. 2021. Key Point Analysis via Contrastive Learning and Extractive Argument Summarization. In *Proceedings of ArgMining@EMNLP 2021*, pages 184–189.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From Arguments to Key Points: Towards Automatic Argument Summarization. In *Proceedings of ACL 2020*.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative Argument Summarization and Beyond: Cross-Domain Key Point Analysis. In *Proceedings of EMNLP 2020*.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X Retrieval: What Retrieval Granularity Should We Use? In *Proceedings of EMNLP 2024*, pages 15159–15177.

Artem Chernodub, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann, and Alexander Panchenko. 2019. TARGER: Neural Argument Mining at Your Fingertips. In *Proceedings of ACL 2019*, pages 195–200.

Giulia D'Agostino, Chris A. Reed, and Daniele Puccinelli. 2024. Segmentation of Complex Question Turns for Argument Mining: A Corpus-based Study in the Financial Domain. In *Proceedings of LREC/COLING 2024*, pages 14524–14530.

Lorik Dumani, Christin Katharina Kreutz, Manuel Biertz, Alex Witry, and Ralf Schenkel. 2020. Segmenting and Clustering Noisy Arguments. In *Proceedings of LWDA 2020*, pages 23–34.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 Key Point Analysis Shared Task. In *Proceedings of ArgMining@EMNLP 2021*, pages 154–164.

Yujie Fu, Yang Li, Suge Wang, Xiaoli Li, Deyu Li, Jian Liao, and Jianxing Zheng. 2023. Hierarchical Enhancement Framework for Aspect-based Argument Mining. In *Proceedings of EMNLP 2023*, pages 1423–1433.

Atsushi Fujii and Tetsuya Ishikawa. 2006. A System for Summarizing and Visualizing Arguments in Subjective Documents: Toward Supporting Decision Making. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text 2006*, pages 15–22.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing Argumentative Discourse Units in Online Interactions. In *Proceedings of ArgMining@ACL 2014*, pages 39–48.

Samuel Guilluy, Florian Méhats, and Billal Chouli. 2023. Constituency Tree Representation for Argument Unit Recognition. In *Proceedings of ArgMining 2023*, pages 35–44.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Comput. Linguistics*, 43(1):125–179.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are You Taking this Stance? Identifying and Classifying Reasons in Ideological Debates. In *Proceedings of EMNLP 2014*, pages 751–762. ACL.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing Retrieval and Language Models for Knowledge-Intensive NLP. arXiv/2212.14024.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. arXiv/2310.03714.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining Arguments From 19th Century Philosophical Texts Using Topic Based Modelling. In *Proceedings of ArgMining@ACL 2014*, pages 79–87.

Marco Lippi and Paolo Torroni. 2015. Context-Independent Claim Detection for Argument Mining. In *Proceedings of IJCAI 2015*, pages 185–191.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn A. Walker. 2015. Using Summarization to Discover Argument Facets in Online Idealogical Dialog. In *Proceedings of NAACL HLT 2015*, pages 430–440.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic Detection of Arguments in Legal Texts. In *Proceedings of IAAIL 2007*, pages 225–230.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Raquel Mochales Palau and Marie-Francine Moens. 2008. Study on the Structure of Argumentation in Case Law. In *JURIX 2008*, pages 11–20.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation Mining: The Detection, Classification and Structure of Arguments in Text. In *Proceedings of ICAIL 2009*, pages 98–107.

Isaac Persing and Vincent Ng. 2016. End-to-End Argumentation Mining in Student Essays. In *Proceedings of NAACL HLT 2016*, pages 1384–1394.

Lev Pevzner and Marti A. Hearst. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Comput. Linguistics*, 28(1):19–36.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of ACL 2019*, pages 567–578.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Comput. Linguistics*, 43(3):619–659.

Manfred Stede and Jodi Schneider. 2019. Finding Claims. In *Argumentation Mining*, pages 57–76. Springer.

Shahbaz Syed, Roxanne El Baff, Johannes Kiesel, Khalid Al Khatib, Benno Stein, and Martin Potthast. 2020. News Editorials: Towards Summarizing Long Argumentative Texts. In *Proceedings of COLING 2020*, pages 5384–5396.

Maite Taboada and William C. Mann. 2006. Rhetorical structure theory: looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

Amine Trabelsi and Osmar R. Zaïane. 2019. PhAITV: A Phrase Author Interaction Topic Viewpoint Model for the Summarization of Reasons Expressed by Polarized Stances. In *Proceedings of ICWSM 2019*, pages 482–492.

Dietrich Trautmann. 2020. Aspect-Based Argument Mining. arXiv/2011.00633.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-grained argument unit recognition and classification. In *Proceedings of AAAI 2020*, pages 9048–9056. AAAI Press.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an Argument Search Engine for the Web. In *Proceedings of ArgMining 2017*, pages 49–59.

A Appendix

Decompose the ’content’ into clear and simple arguments, ensuring they are interpretable out of context.

1. Maintain the original phrasing from the input whenever possible.
2. Decontextualize each argument by adding necessary modifier to nouns or entire sentences and replacing pronouns (e.g. ’it’, ’he’, ’she’, ’they’, ’this’, ’that’) with the full name of the entities they refer to.
3. Each argument can consist of multiple sentences.
4. Present the results as a list of strings.

Input: Good morning my fine opponent and thank you for this wonderful debate. I will start off with a small overview of my points, and then leave it to you for the next round. [. . .] Teens should be able to develop self-expression and their personal identity. Instead, they might resort to unconventional piercings and tattoos... School uniforms encourage followers not leaders. The practice discourages independent thinkers. This follower mentality could extend into adulthood. [. . .]

Output:

- Teens should be able to develop self-expression and their personal identity. With school uniforms, they might resort to unconventional piercings and tattoos...
- School uniforms encourage followers not leaders. The practice discourages independent thinkers. This follower mentality could extend into adulthood.
- . . .

Input: <*new text*>
Output:

Figure 4: PaLM prompt for segmentation.

[. . .] So if there is no evidence left behind and there are scientific explanations for things usually described to god. Than the evidence points more towards there being no god. Just because we can’t know absolutely doesn’t mean that based on the evidence we can’t make an educated guess about what is most likely true. "In short, you have faith God doesn’t exist therefore live your life as if he did not. " I don’t have faith that he does not exist I have no reason to believe he does [. . .] We can never know anything absolutely but we do have evidence that that can help us know what is most likely true.

Figure 5: Text on topic ’We should adopt atheism’, highlighting two different formulations of the same statement.

- Eating meat isn’t necessary for maximum physical development.
- All of the vitamins and minerals in meat can be found in other foods.
- The taste of meat does not outweigh the costs of killing.
- Vegetarianism saves animals’ lives.
- Vegetarianism helps one’s health.
- Vegetarianism helps the environment.
- Almost all major religions agree that vegetarianism is better than eating meat.

Figure 6: Segments generated by GPT-4o mini for the argumentative text in Section 1, Figure 1.

Match Category	Manual	Model
matched	man	Eating meat isn’t necessary for maximum physical development.
matched	man	Eating meat isn’t necessary for maximum physical development. All of the vitamins, minerals etc. in meat can be found in other foods.
incomplete (a)	man	Eating meat isn’t necessary.
incomplete (b)	man	part 1: Eating meat isn’t necessary part 2: All of the vitamins, minerals etc. in meat can be found in other foods.
impure	man	Eating meat isn’t necessary for maximum physical development. All of the vitamins, minerals etc. in meat can be found in other foods. And does the taste of meat really outweigh the costs of killing?
incomplete & impure	man	Eating meat isn’t necessary. And does the taste of meat really outweigh the costs of killing?
missed	man	—
spurious	—	Thanks for the timely response.

Table 7: Examples for different error categories in segment matching. Manually extracted main proposition (man): “Eating meat isn’t necessary for maximum physical development.”

Measure	Approach					
	PaLM filt.	GPT-4 filt.	Paragr. filt.	Sent. filt.	Ajjour. filt.	Targer. filt.
# Segments	173	272	154	408	413	465
matched | Precision	0.57	0.35	0.63	0.38	0.29	0.30
– correct	0.43	0.24	0.36	0.21	0.18	0.18
– incomplete	0.11	0.10	0.16	0.16	0.11	0.11
– impure	0.02	0.01	0.08	0.00	0.00	0.00
– incomplete & impure	0.01	0.00	0.03	0.01	0.00	0.01
spurious	0.43	0.65	0.37	0.62	0.70	0.69
matched | Recall	0.59	0.52	0.66	0.79	0.64	0.73
– correct	0.50	0.41	0.38	0.56	0.49	0.56
– incomplete	0.05	0.08	0.07	0.21	0.14	0.15
– impure	0.03	0.02	0.18	0.01	0.0	0.01
– incomplete & impure	0.01	0.01	0.03	0.01	0.01	0.01
missed	0.41	0.47	0.33	0.20	0.36	0.27
F1 micro strict	0.46	0.30	0.37	0.31	0.26	0.27
F1 micro relaxed	0.58	0.42	0.65	0.52	0.41	0.44

Table 8: Effectiveness of different argument segmentation approaches after filtering segments labeled as non-argumentative (classification approach by (Reimers et al., 2019)). The evaluation is done from a precision- and recall-oriented perspective, reporting the micro average scores. The rows ’matched’ cover all correct and incorrect segments, the total number of manual reference segments is 147.

Segment PaLM	**Key Point**
Every time the DP is used the right to life is violated.	State-sanctioned killing is principally wrong
Study shows that there there is not enough evidence to support the fact that the death penalty does not act as a deterrence.	The death penalty is ineffective in deterring crimes
88% of expert criminologists concur that the death penalty doesn't deter violent crime, despite what these "flimsy" studies might suggest.	The death penalty is ineffective in deterring crimes
People generally support capital punishment because they believe criminals do not deserve to live.	State-sanctioned killing is principally wrong
The purpose of the justice system is, ideally, to make an impartial decision not to satisfy the lust for vengeance possessed by the victim's loved ones. The "bonus" of satisfying the family is hardly adequate reason to support the death penalty.	The death penalty helps the victim/their family
Pro argues that the death penalty is justified because it is saving money that would otherwise be used for life imprisonment; if anyone is trying to put a dollar value on human life, it would be Pro.	The death penalty saves costs to the state
Even law enforcement admits that the death penalty is "the least efficient use of taxpayers' money".	The death penalty saves costs to the state
The use of the death penalty is actually far more expensive than the maintenance of a LIP inmate.	The death penalty saves costs to the state
What makes it right for the guilty person to be deserved of the same thing he's being executed for? [. . .] life imprisonment is a better means of punishing the guilty.	
It is exceedingly rare for those confined in prison to escape.	
Justice is not killing people to make ourselves feel at ease. Justice is not an eye for an eye, a tooth for a tooth until we are all blind and toothless.	
. . . our "justice" system would condemn these men to death without providing a chance for contrition, repentance, or redemption. [...]	
Ethical justifications are not based upon economic gains ([. . .] human life cannot be compared to material goods).	

Table 9: Examples of segment–key point matches for the topic 'We should abolish capital punishment'. The segments in the lower half have no suitable key point in the ArgKP 2021 dataset (Friedman et al., 2021).

#	Segments
0	Viewers keep in mind I will first be finishing addressing my opponents round 2 rebuttal.
1	Defense: C.4, Jury less likely to condemn.
2	I would like some evidence that the jury are able to choose the sentence of the criminal, because I've been searching for it but can't find it.
3	The DP saves only the lives of criminals being murdered, besides cases when murderers are let go with should not happen. The DP sometimes kills innocents.
4	Defense: C.5, Innocence.
5	Here's a case in which a man was framed by the police. [1]
6	This source says that well over eighty people in the past quarter century have been condemned but then released before execution [2]
7	This source shows accounts of 11 innocents being executed. [3]
8	Here is a quote from one study taken.
9	" In my current research into probable innocents that have been executed, I have uncovered at least 74 cases in which wrongful executions have most likely taken place." [2]
10	Let me also add what I said in the previous round. There have undoubtedly been cases in which the innocent have been executed but have not been proved innocent afterwards. After being executed there is usually not much need for someone to try too prove the innocence of someone who is already dead. So there are undoubtedly instances in the past where we have executed an innocent man but did not know so, and still do not know.
11	With life imprisonment there is zero chance of killing an innocent man.
12	Defense: C.6, Life imprisonment just as effective removing those who cause harm.
13	The only people murderers can harm is their fellow inmates, assuming they were not sentenced to solitary confinement. This is far outweighed by the fact that executing innocents is a much bigger a problem than murderers and rapists killing each other.
14	Defense: C.8, Violates the right to life.
15	1. Not outweighed: The lives saved by the DP are the lives of rapists and murderers. The lives saved by life imprisonment are the lives on innocent people wrongly condemned.
16	2. "Not comparable morally": The murderer of course has no right to take another mans life. So what makes it right for us to take his life?
17	3. Murderer is guilty, but not deserved of death: What makes it right for the guilty person to be deserved of the same thing he's being executed for? Of course he's guilty, but life imprisonment is a better means of punishing the guilty.
18	4. DP Vengeful: My opponents analogy's are faulty. His first analogy doesn't even make sense because what he's saying is it would be absurd to kidnap someone to show that kidnapping is wrong. I agree that would be absurd, which is why killing people to show killing people is wrong is also absurd. His second analogy is completely wrong because cops don't punish those who speed by speeding.
19	5. Violating anothers rights does not deprive you of your own: John Stuart Mill is essentially saying the "eye for eye tooth for tooth" concept is right. Proving that the DP is vengeful. This concept is widely accepted as wrong.
20	6. I'm not sure exactly what my opponent means by personal liberty, but putting a man in prison for murder is easily justified while the DP is not.
21	7. The fact of whether war is justified is completely another matter.
22	8. "Protecting the right to life": Every time the DP is used the right to life is violated. As for lives saved, see my first point.
23	It is fallacious reasoning to assume that, because murder rates were dropping at the time the DP was used that means it was because of the DP.
...	...
33	"Prisoners prefer life"
34	I think that it all depends for different people.
35	Being locked in a single small room in solitary confinement for years on end is certainly not very pleasant.
...	...
49	http://beta.nodeathpenalty.org... [1]
50	http://www.the-slammer.org... [2]
51	http://www.justicedenied.org... [3]

Table 10: Examplary paragraph segments. Gray segments are removed since they contain less than 3 whitespaces—the segment at index 33 is the only relevant passage that is lost by this filtering approach.