Opinion diversity through hybrid intelligence

Meer, M.T. van der

2

An Empirical Analysis of Diversity in Argument Summarization

Presenting high-level arguments is a crucial task for fostering participation in online societal discussions. Current argument summarization approaches miss an important facet of this task—capturing diversity*—which is important for accommodating multiple perspectives. We introduce three aspects of diversity: those of opinions, annotators, and sources. We evaluate approaches to a popular argument summarization task called Key Point Analysis, which shows how these approaches are ill-equipped for (1) dealing with data from various sources, (2) representing arguments shared by few people, and (3) aligning with subjectivity in human-provided annotations. We find that both general-purpose LLMs and dedicated Key Point Analysis models vary along these three criteria, but have complementary strengths. Further, we observe that diversification of training data may ameliorate generalization. Addressing diversity in argument summarization requires a mix of strategies to deal with subjectivity.*

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2024. An Empirical Analysis of Diversity in Argument Summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2028–2045, St. Julian's, Malta. Association for Computational Linguistics.

2.1 Introduction

Getting an overview of the arguments concerning controversial issues is often difficult for those participating in ongoing discussions. In these discussions, many points are being communicated, there is no way to track which arguments were already encountered, and participants engage in haphazard miscommunication or conflicts. Automatic summarization is a way to provide a comprehensible overview of the opinions [15, 281]. However, generating summaries representative of the arguments involved in a discussion is difficult [32]. Argument summarization extends beyond text summarization because it separates argumentative and non-argumentative content, preserves the argumentative structure, and provides explicit stances on a central claim or hypothesis.

Summarizing arguments is challenging in many contexts, but the potential impact is high. For instance, after summarizing the arguments from societal discussions, the extracted arguments may shape new policies and may be used to justify decision-making [17, 153]. Similarly, businesses depend on review data to find customer feedback, which can steer product design [18].

Although arguments are often summarized by hand in practice [e.g., 264, 274, 279], recent developments in Argument Mining (AM) allow automatic analysis of argumentative text [224]. Obtaining summaries that faithfully represent open-ended opinions requires careful evaluation, especially in sensitive contexts, e.g., summarizing citizen feedback [109, 267].

One approach for generating comprehensive summaries of arguments is Key Point Analysis [KPA, 32]. In KPA, a corpus of opinions is analyzed for the *key points*, those arguments that are salient and repeated multiple times. However, some aspects of the KPA experimental design misalign with respect to real-world applications. We illustrate these blind spots, in particular, when applied to summarizing online societal discussions. We highlight three dimensions of **diversity** that are central to empowering citizens' opinions at scale [352]: (1) incorporating the long tail of opinions, (2) including diverse perspectives from annotators, and (3) being robust in handling data from multiple sources.

How current KPA approaches deal with the above dimensions of diversity is unexplored. We incorporate the standardized benchmark and two other datasets to experiment with different approaches. We develop specific analyses to uncover how KPA approaches fare on each dimension of diversity. In addition to the existing approaches, we use LLMs by prompting them to perform KPA, as they may be an attractive alternative to current models.

Applying KPA approaches across several datasets that vary in how they address diversity leads to mixed results. KPA approaches generalize poorly across data sources when used in transfer learning settings, though approaches reveal complementary merits across tasks. Further, their performance degrades when dealing with low-frequency opinions, i.e., opinions repeated by relatively few individuals. Finally, we observe that KPA approaches disregard subjective interpretations among individual annotators.

Contributions (1) We critically examine three dimensions of diversity—of opinions, annotators, and sources—in the KPA setup. (2) We analyze the behavior of existing metrics on one existing and two novel datasets. (3) We analyze multiple methods, including promp-t-based LLMs, broadening the scope of methods that can perform KPA.

2.2 Related Work

2.2.1 Key Point Analysis

KPA serves to separate argumentative from non-argumentative content, and condense argumentative content by matching arguments to key points [32]. Key points can be seen as high-level arguments that capture the gist of a set of arguments. While most work on KPA selects high-quality arguments as representatives, generating novel key points has been proposed as an alternative [376]. KPA has been applied across topics using data from discussion portals or online reviews [33, 34]. KPA is usually divided into Key Point Generation and Key Point Matching steps (see Section 2.3.1).

Multiple approaches exist for KPA [131]. Modeling choices consist of popular Transformer models such as BERT [301], enhanced representational quality using contrastive learning [10], and the incorporation of clustering techniques [231]. Our work aims to investigate some of the modeling choices employed in these works. For instance, in Li et al. [231], the authors discarded unmapped arguments, which may hurt the ability the represent minority opinions.

2.2.2 Opinion Summarization

Opinion summarization aims to generate summaries of an individual's subjective opinions [48, 180], often applied to product reviews [75]. Leveraging Transformer models is popular for opinion summarization [13, 16], though generic extractive summarization techniques are strong baselines [373]. Measuring bias in generated summaries has seen recent interest, specifically acknowledging that diverse opinions should be taken into account [176, 355] or postulating that diversity is a desirable trait when generating opinions [12, 420]. Our work applies these techniques to argumentation to obtain a high-level summary of opinions, and analyses differences in behavior for (in-)frequent viewpoints.

2.2.3 Diversity in Societal Decision Making

Sensitive decision-making contexts call for responses rooted in reason that serve social good rather than specific interests. One way of obtaining such responses is through evidence-based policymaking, which involves stakeholders and the broader public to strike decisions [64]. Citizen participation improves the support of the decisions when some requirements are met [260]. A key factor among those requirements is the involvement of a diverse group of citizens, independently voicing opinions [375]. Approaches to summarizing arguments in such citizen feedback face similar requirements.

In Argument Mining, we find recent work that aligns with these views, e.g., by a strong focus on the diverging perspectives among annotators in AM tasks [322]. Further, some preliminary work adjusts visualization for minority opinions [38]. However, in terms of data sources, most work is still centered on English-speaking content, with few multi-lingual or multi-cultural resources available [414].

2.3 Method

We formulate the KPA subtasks—*Key Point Generation* (KPG) and *Key Point Matching* (KPM). We then introduce the three dimensions of diversity and consider them when applying KPA.

Dataset	Data Source	Filter low freq.	Key Point source	Non-aggregated annotation	IRR
ArgKP	Human annotation	✓	Expert	✗	0.50-0.82 (κ)
PVE	Citizen consultation	✗	Crowd	✓	0.35 ($\kappa^\dagger$)
Perspectrum	Debate platforms	✗	Crowd	✗	0.61 (κ)

Table 2.1: Datasets and their diversity characteristics when considering the KPA task. The inter-rater reliability (IRR) is measured via Cohen's κ scores or prevalence and bias-adjusted Cohen's $\kappa^\dagger$ [PABAK, 357].

2.3.1 Task setup

We outline the two subtasks that constitute KPA, as originally introduced by [131].

Key Point Generation (KPG) focuses on generating *key points* $\mathcal{K}$ given a corpus of arguments $\mathcal{D}$ on a particular claim. Key points are high-level arguments that capture the gist of a collection of arguments. Key points oppose or support the claim.

Key Point Matching (KPM) *matches* arguments to key points. An argument matches a key point if the key point directly summarizes the argument, or if the key point represents the essence of the argument. We ensure that the stance of the key point (pro or con) matches the stance of the argument. Formally, given a set of key points $\mathcal{K}$ and a corpus $\mathcal{D}$, we score the match between an argument $d \in \mathcal{D}$ and a key point $k \in \mathcal{K}$ with a matching model $M(d,k)$. Assigning arguments to key points using match scores is flexible, and multiple strategies can be taken to reach a final decision (e.g. imposing a match score threshold) [33]. Since the assignment strategy is largely context-dependent, we evaluate the scoring mechanism itself, instead.

2.3.2 Modeling Diversity in Key Point Analysis

We focus on three main aspects of diversity.

Long tail opinions Several NLP models imitate biases that exist in datasets [51]. For argument summarization, focusing on majority arguments is one such form of bias, as it leads to possible misrepresentations. Failing to capture low-frequency arguments runs the danger of further estranging underrepresented viewpoints [204]. These methods need active correction from humans to account for this "long tail of opinions" [397]. For the KPA task, approaches have largely unknown behavior on capturing the long tail of opinions [278]. Additionally, LLMs struggle with learning long-tail knowledge [193], aggravating this issue. We experiment with subsampling the datasets to investigate the imbalanced data settings, which are representative of real-world use cases.

Annotators Datasets are labeled using a mix of crowd and expert annotators. Querying experts for key points may leave the impacted users (e.g., lay citizens) out of consideration [60]. Similarly, labels stemming from crowd annotation that are filtered for high agreement may disregard controversial or diverse opinions. Disagreement is a complex signal that includes subjective views, task understanding, and annotator behavior [21]. Having access to non-aggregated annotations would, e.g., allow for further modeling of patterns [89] or the reasons [241] underlying opinions. We investigate whether models trained on such annotations can identify disagreement.

Data sources Existing works investigate cross-domain generalization of KPA methods using data stemming from a single dataset, focusing on a cross-topic setting [33, 231, 330]. This dataset is gathered at a specific time. As discussions evolve, more nuanced positions may become relevant, and new real-world events impact the opinions. Further, these discussions usually take place on a single platform (e.g., Reddit threads, Twitter discussions), inheriting biases from the source [170]. Measuring the performance of KPA approaches should rely on diverse datasets, based on data gathered from different sources at different points in time. There have been some efforts in applying KPA across different contexts [34, 66, 145], but they apply approaches to a single dataset at a time, making direct comparison difficult. Our work examines the cross-dataset performance of these approaches to assess their relative strengths and weaknesses.

Table 2.1 shows the current datasets, and how they relate to the dimensions discussed above. In all three datasets, the arguments stem from user-submitted content. In one dataset (ArgKP), low-frequency arguments (i.e., opinions repeated by few individuals) are disregarded. Further, the ArgKP benchmark relies on expert-generated key points and does not include annotator-specific match labels. Perspectrum contains aggregated counts of match labels, but due to aggregation, we cannot identify annotator-specific patterns. Lastly, the inter-rater reliability differs for each dataset, with wide ranges, showing that the tasks are fundamentally subjective. We employ these three datasets for evaluating various KPA approaches and dive deeper into the three aspects of diversity.

2.4 Experimental Setup

We describe the data, KPA methods, and metrics involved in our experiments. The source code will be publicly available upon publication.

2.4.1 Data

Most work on KPA has used ArgKP, the dataset introduced by Friedman-Melamed et al. [131] in a shared task. We add two new datasets that match the KPA subtasks but have different characteristics.

ArgKP We adopt the shared task dataset, keeping the same split across claims as the original data. The ArgKP dataset contains claims taken from an online debate platform, together with crowd-generated arguments and expert-generated key points [32]. The arguments were produced by asking humans to argue for and against a claim, followed by filtering on high-quality and clear-polarity arguments. Key points were generated by an expert debater, who generated the key points without having access to the arguments. The final test set was collected after the initial dataset and has been curated to match some of the distributional properties of the training and validation sets.

PVE We use the crowd-annotated data stemming from a human-AI hybrid key argument analysis [397] based on a Participatory Value Evaluation (PVE), a type of citizen consultation. In this consultation process, citizens were asked to motivate their choices for new

Dataset	Train	Val	Test
ArgKP	24 (21K)	4 (3K)	3 (3K)
PVE	–	–	3 (200)
Perspectrum	525 (6K)	136 (2K)	218 (2K)

Table 2.2: Number of claims (and arguments) when splitting the dataset into training, validation, and test sets.

COVID-19 policy through text, which formed a set of comments for each proposed policy option. The performed key argument analysis resulted in crowd-generated key points, matching individual comments to key points per option. Since this is a small dataset, we only use it for evaluation.

Perspectrum Similar to ArgKP, Perspectrum contains content from online debate platforms. It extracts claims, key points, and arguments from the platform directly [71]. Part of the dataset is further enhanced by crowdsourcing paraphrased arguments and key points. The Perspectrum dataset is ordered into claims, which are argued for or against by perspectives, with evidence statements backing up each perspective. We use perspectives as key points, and evidence as arguments. We retain the same split over claims as the original data. The authors provide aggregated annotations on the match between arguments and key points. While this allows us to compute the agreement scores per sample, we cannot distill individual annotator patterns.

2.4.2 Approaches

We investigate different approaches with respect to their performance on the aspects of diversity. Appendix A.1 includes a detailed overview of the setup, parameters, and prompts. Similar to summarization techniques, most KPG methods are either *extractive*, taking samples as representative key points, or *abstractive*, formulating new key points as free-form text generation [113].

ChatGPT We use the OpenAI Python API [290] to run the KPA task by prompting ChatGPT. We differentiate between open-book and closed-book prompts. For the open-book prompts, we input the claim and a random sequence of arguments up to the maximum window (given a response size of 512 tokens) in the KPG task. For the closed-book model, we only input the claim, and the model synthesizes key points. In both approaches, KPG is abstractive. In KPM, ChatGPT predicts matches for a batch of arguments at a time, all related to the same claim.

Debater We use the Project Debater API [179], which supports multiple argument-related tasks, including KPA [35]. This approach uses a model trained on ArgKP and performs extractive KPG. We query the API for KPG and KPM separately.

SMatchToPR We adopt the approach from the winner of the shared task, which uses a state-of-the-art Transformer model and contrastive learning [10]. During training, the model learns to embed matching arguments closer than non-matching arguments. These representations are used to construct a graph with embeddings of individual argument sentences as nodes, and the matching scores between them as edge weights. Nodes with the maximum PageRank score are selected as key points. In our experiments, the model is trained using the training set of ArgKP and Perspectrum. This method performs extractive KPG. We experiment with RoBERTa-base and RoBERTa-large to estimate the effect of model size [248].

2.4.3 Evaluation Metrics

We evaluate models for KPG and KPM separately. For KPG, we adopt the set-level evaluation approach from Li et al. [231]. For KPM, we reuse the match labels provided by each dataset.

Key Point Generation (KPG)

KPG can be considered as a language generation problem [135] for evaluation. We rely on a mixture of reference-based and learned metrics, measuring both lexical overlap and semantic similarity. We use the following metrics:

ROUGE-(1/2/L) to measure overlap of unigrams, bigrams, and longest common subsequence, respectively. We average scores for all stance and claim combinations. Additional details on the ROUGE configuration are in Appendix A.1.3.

BLEURT [347] to measure the semantic similarity between a candidate and reference key point, which correlates with human preference scores. BLEURT introduces a regression layer over contextualized representations, trained on a set of human-generated labels.

BARTScore [445] to evaluate the summarization capabilities directly by examining key point generation. In contrast to BLEURT, BARTScore evaluates the likelihood of the generated sequence when conditioning on a source.

For each metric $\mathcal{S}$ that scores the overlap between two key points, we aggregate scores into Precision P and Recall R scores using Equations 2.1 and 2.2. For P, we take the maximum score between a generated key point a and the reference key points $\mathcal{B}$, averaging over all $n = |\mathcal{A}|$ generated key points. We perform the analogous for R. We report F_1 scores to balance precision and recall.

$$P = \frac{1}{n} \sum_{a \in \mathcal{A}} \max_{b \in \mathcal{B}} \mathcal{S}(a,b) \tag{2.1}$$

$$R = \frac{1}{m} \sum_{b \in \mathcal{B}} \max_{a \in \mathcal{A}} \mathcal{S}(a,b) \tag{2.2}$$

Key Point Matching (KPM)

We perform the KPM evaluation by obtaining match scores for key point-argument pairs. That is, for a key point k and an argument d, we check if a new model used in the KPA method would assign d to k. We reuse existing labels and do not use the results from KPG. Since we do not consider unlabeled examples between arguments and key points, we do not need to distinguish for undecided labels (as in Friedman-Melamed et al. [131]).

We evaluate each approach using mean average precision (mAP), taking the mean over average precision scores computed for claims C. Given a claim, we compute precision P_τ and recall R_τ for all match score thresholds τ, as in Equation 2.3. In case an approach outputs a binary match label instead of scores, we remap the scores to 0 and 1 for non-matching and matching pairs, respectively.

$$\text{mAP} = \sum_{C} \frac{\sum_{\tau}(R_{\tau} - R_{\tau-1})P_{\tau}}{|C|} \tag{2.3}$$

2.5 Results and Discussion

First, we report on the KPG and KPM evaluation. Then, we analyze how the aspects of diversity impact performance beyond a cross-dataset evaluation. We show results when conditioning on the long tail of opinions, look into the connection between annotator agreement and match score, and how performance changes for diverse data sources.

2.5.1 KPG Performance

Table 2.3 shows the results of KPG evaluation. Overall, no single approach performs best across all datasets. All models perform best on ArgKP except for closed-book ChatGPT, which performs the best on the PVE dataset. Thus, by adopting diverse datasets, we demonstrate that experimenting with a single dataset may inflate KPG performance.

ChatGPT consistently scores well on ROUGE and semantic similarity. This indicates that the abstractive generation of key points is beneficial. For PVE, we observe a strong tendency for open-book ChatGPT to adjust the generated key points to the linguistic style of the arguments. This clashes with the reference key points, which are paraphrased to make sense without the context of the original arguments. Hence, the closed-book model, which does not observe the source arguments, performs better, adopting more neutral language.

SMatchToPR performs best for Perspectrum. Although general-purpose LLMs are strong in zero-shot settings, a dedicated model for representing arguments achieves state-of-the-art results. The Debater approach is ranked lowest across all datasets, showing that training on a single dataset generalizes poorly to other datasets.

ROUGE and semantic similarity scores mostly agree, except for BLEURT on ArgKP. Here, we see that SMatchToPR slightly outperforms ChatGPT. We attribute this to the optimized representational qualities of SMatchToPR: it selects key points with high semantic similarity to many arguments, which is similar to how BLEURT provides scores based on contextualized representations.

Increasing model size (of SMatchToPR) improves performance for Perspectrum, but not for ArgKP and PVE. Because PVE is small, the pool of sentences to pick key point candidates from is limited, and possible improvements of the model are negligible when extracting the key points. For ArgKP, the ROUGE scores deteriorate, while the semantic similarity scores improve slightly. Intuitively, this matches expectations: the model can navigate the embedding space better, selecting key points that may be phrased differently but contain semantically similar content.

2.5.2 KPM Performance

Table 2.4 shows the results of KPM evaluation. ChatGPT, despite its strong performance on KPG, does not accurately match arguments to key points. Interestingly, the Debater outperforms the SMatchToPR model on the ArgKP dataset, but SMatchToPR is stronger on the PVE and Perspectrum datasets. SMatchToPR's strong performance on Perspectrum

Dataset	Approach	R-1	R-2	R-L	BLEURT	BART
ArgKP	ChatGPT	**34.3**	**12.5**	**30.3**	0.556	**0.540**
	ChatGPT (closed book)	29.5	7.1	25.6	0.314	0.256
	Debater	25.6	5.5	22.5	0.334	0.307
	SMatchToPR (base)	31.7	11.1	29.7	0.553	0.494
	SMatchToPR (large)	30.5	8.3	26.8	**0.563**	0.497
PVE	ChatGPT	18.5	3.9	15.3	0.329	0.369
	ChatGPT (closed book)	**27.1**	**8.6**	**21.4**	**0.376**	**0.378**
	Debater	13.3	0.0	13.3	0.294	0.188
	SMatchToPR (base)	21.3	3.7	16.6	0.351	0.344
	SMatchToPR (large)	21.3	3.7	16.6	0.351	0.344
Perspectrum	ChatGPT	21.3	5.7	18.2	0.355	0.322
	ChatGPT (closed book)	17.1	3.8	15.0	0.291	0.258
	Debater	9.4	0.4	8.5	0.197	0.210
	SMatchToPR (base)	22.5	6.5	19.3	0.257	0.232
	SMatchToPR (large)	**22.7**	**6.7**	**19.4**	**0.403**	**0.363**

Table 2.3: ROUGE scores and semantic similarity scores for the Key Point Generation task.

	mAP		
Name	ArgKP	PVE	Perspectrum
ChatGPT	0.17	0.27	0.46*
Debater	**0.82**	0.51	0.51
SMatchToPR (base)	0.76	0.53	0.80
SMatchToPR (large)	0.80	**0.61**	**0.82**

Table 2.4: Results for the Key Point Matching task. Closed-book ChatGPT scores are not available, since its KPA is made without observing arguments. The scores for ChatGPT on Perspectrum (*) were estimated on a subset of the test set to cut down costs.

and ArgKP is expected–they were included in its training. However, its good performance on PVE is interesting and it suggests that generalization is aided by more diverse data in training.

2.5.3 Analysis

Long tail diversity Most key points and claims are heavily skewed in the number of data points, except for PVE. Even for ArgKP, where key points with few matching arguments were removed, there is a strong imbalance across claims and key points in terms of associated arguments (see Figure 2.1).

Following this imbalance, we sort key points by the number of associated arguments such that the least frequent key points are considered first. Then, we introduce a cutoff parameter f to include arguments from a fraction of key points, starting with the least frequent. Using this parameter we perform matching only on low-frequency key point–arguments matches.

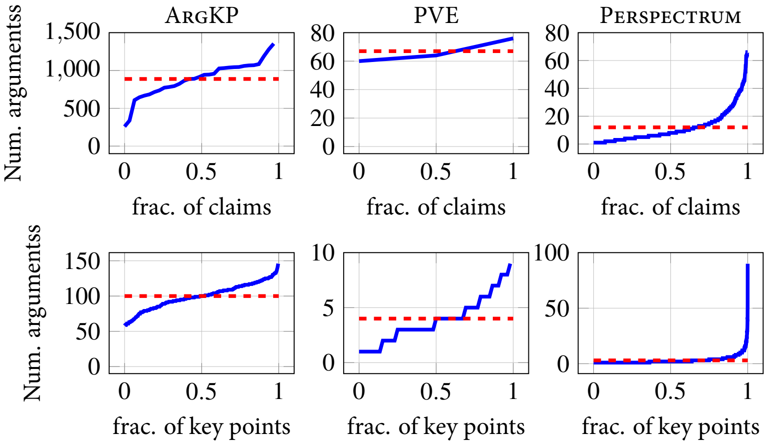

Figure 2.1: Number of arguments matched per claim (upper row) and key point (bottom row), sorted by frequency. The red dashed line shows the average number of arguments.

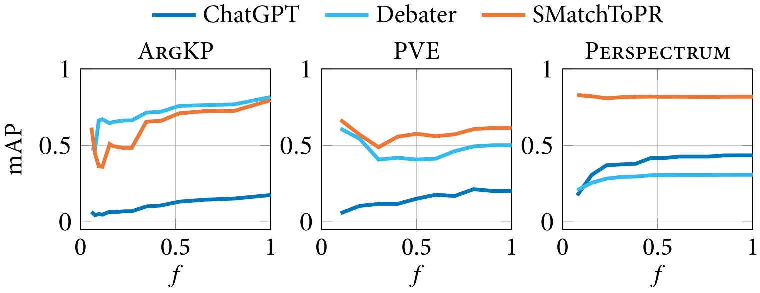

Figure 2.2: KPM performance when limiting data usage to a fraction f, starting with long tail first.

This allows us to investigate the approaches' performance in the long tail.

When we limit data usage by taking long tail arguments first, the performance of the KPA approaches, mainly on ArgKP and Perspectrum, decreases as shown in Figure 2.2. This shows that the ability to correctly match arguments is contingent on the frequency of the arguments. In some cases, the arguments associated with key points with the fewest matches can be matched, but there is a strong performance loss for low values of f. Across all datasets, ChatGPT suffers consistently in mAP when conditioning on low-frequency key points. For SMatchToPR on Perspectrum, there is almost no effect, showing that representation learning may positively impact the matching of key points to arguments even with low amounts of data. Performing the same experiment for KPG results in similar results: key points with a low number of matched arguments are harder to represent well.

Next, we investigate whether the arguments in the long tail are different from the majority. Here, the long tail consists of arguments for key points that see less than the median number of arguments per key point. We examine whether the sets of lexical items—noun phrase chunks (NPs) and entities—mentioned in the long tail arguments are included in the

		NP		Entity			
Left (long tail)	**Right (majority)**	left−right	right−left	left−right	right−left	NP-τ	Ent-τ
ArgKP	ArgKP	0.168	0.234	0.191	0.273	0.216*	0.373*
PVE	PVE	0.638	0.787	0.719	0.809	0.521*	0.389
Perspectrum	Perspectrum	0.397	0.807	0.401	0.797	0.361*	0.427*

Table 2.5: Fraction of NPs and Entities in **Left** that are not in **Right** & vice-versa. * indicates Kendall τ with $p < 0.05$.

	PVE		Perspectrum	
Approach	r	p	r	p
ChatGPT	0.030	0.687	0.039	0.469
Debater	0.163	0.029	-0.051	0.013
SMatch-base	0.097	0.195	0.093	0.215
SMatch-large	0.207	0.005	-0.03	0.123

Table 2.6: Pearson r correlation scores between predicted match scores and the annotator agreement per sample.

majority and vice versa. We also inspect the relative frequency of the shared lexical items via Kendall τ correlation on the NP and entity frequency rankings. Table 2.5 shows these results.

We see a large overlap of NPs and entities for ArgKP between the long tail and the frequent key points. We attribute this to the filtering of low-frequency data during dataset construction. For the other two datasets, we observe much less overlap—in most cases, more than half of the noun phrases and entities are unique to either part of the dataset. The only exception here is Perpectrum, where roughly 40% of the NPs and entities in the long tail are unique. When comparing the ranks of the intersecting lexical items, we observe moderate (but significant) rank correlation scores. Thus, the overlapping NPs and entities may not be in different frequencies in the two parts of the datasets. However, there is a strong indication of unique items in the long tail, in at least two of our datasets, showing that the long tail may contain novel insights.

Annotator agreement Due to subjectivity in the annotation procedures, we expect annotators to rate argument–key point matches differently. We investigate whether the performance of KPA models reflects this subjectivity. That is, we test if match scores x correlate with the agreement between annotators. Intuitively, when annotators agree, an argument and key point should be considered to match more objectively and thus may be easier to score for a model. From the two datasets that have a per-sample agreement score, we measure the Pearson r correlation between the annotator agreement percentage (as obtained from data) and each approach's match score $M(d,k)$. Results are shown in Table 2.6.

For all approaches, the correlations are negligible or weak at best [339]. This shows that the predictions made by the models fail to identify which matches are interpreted differently among annotators. Hence, these models are not able to represent the diversity stemming from annotation accurately [302].

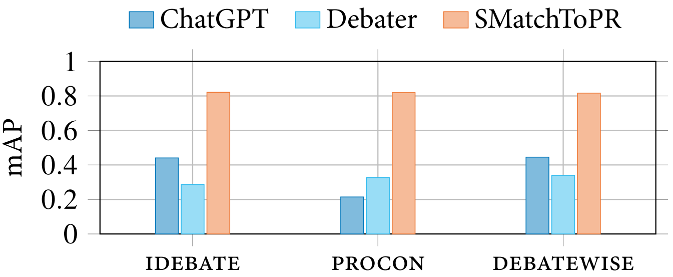

Figure 2.3: KPM performance for all approaches on the different data sources in PERSPECTRUM.

Data sources The KPG and KPM evaluations (Sections 2.5.1 and 2.5.2) indicate how the methods perform when applied to different datasets. The performance is dataset- and task-specific; no single approach performs both tasks best on any dataset. We further investigate the data sources in the PERSPECTRUM dataset, which was constructed using three distinct sources. Figure 2.3 shows the performance on each source separately. Although ARGKP and PERSPECTRUM share a data source, we find no overlapping claims and little repetition in content between the two (App. A.1.1). The SMatchToPR and Debater approaches are not sensitive to data source shift, but ChatGPT performance differs depending on the source data used, dropping considerably for the *procon* source. We find two factors that influence why these arguments are harder to match: (1) *procon* contains about 10 times fewer claims than the other two sources, and (2) *procon*'s arguments are copied verbatim from various cited sources, leading to large stylistic and argumentative differences.

2.6 Conclusion

We perform a novel diversity exploration of different KPA approaches on three distinct datasets. By splitting KPA into two subtasks (KPG and KPM), we investigate each subtask, independently.

First, we find that an LLM-based approach works well for generating key points, but fails to match arguments to key points reliably. Conversely, smaller fine-tuned models are better at matching arguments to key points but struggle to find good key points consistently. Second, using a single training set yields poor generalization across datasets, showing that data source impacts a KPA approach's ability to generalize. Diversification of training data leads to promising results. Third, across all datasets, we see that existing methods for KPA are insensitive to long tail diversity, decreasing performance for key points supported by few arguments. Finally, all models are insensitive to differences between individual annotators, disregarding subjective interpretations of arguments and key points.

We showed how multiple aspects of diversity, a core principle when interpreting opinions, are not evaluated using the standard set of metrics. Our analysis revealed interesting complementary strengths of the KPA approaches. Future efforts could focus on addressing diversity, either by mining for minority opinions directly [425], or by identifying possibly subjective instances using socio-demographic information [43]. Further, models can be enhanced with subjective understanding [322], or work together with humans to jointly address some of the diversity issues [19, 397].

Limitations

We identify five limitations of our work.

Diversity definition Our definition of diversity is specific to three dimensions, but there may be additional dimensions. For example, our unit of analysis is at the *argument* level. Diversity may also be analyzed for the opinion holders or those affected by decisions in policy-making contexts.

Novel key points Our evaluation of KPG and KPM employs existing key points. However, KPA methods may generate novel or unseen key points. Evaluating such novel key points is nontrivial and it may require experiments involving human subjects.

Resource limitations KPA approaches are resource intensive. We limited some approaches where (1) it would become too expensive to run KPA because of the complexity of the number of comparisons (e.g., Debater approach), or (2) the models do not support a big enough window to fit all arguments (e.g., ChatGPT context window is limited). While there are alternatives (e.g., GPT-4), they drastically increase the cost.

Dataset diversity The arguments in our data are in English, and limited to data gathered from online sources. Further, the users involved in collecting the datasets we employ may not be demographically representative of the global population. We conjecture that increasing the diversity of the data sources would make our conclusions stronger. However, publicly available datasets, especially non-English sources, for this task are scarce. We make our code and experimental data public to incentivize further research in this direction.

Data exposure We cannot verify whether the data from the test sets have been used when training the LLMs. This would make the model familiar with the vocabulary and have a more reliable estimation of the arguments' semantics. That likelihood is the smallest for PVE since it is the most recent dataset, gathered with new crowd workers.

Ethical Considerations

There are growing ethical concerns about NLP (broadly, AI) technology, especially, when the technology is used in sensitive applications. Argument summarization can be used in sensitive applications, e.g., to assist in public policy making. An ethical scrutiny of such methods is necessary before their societal application. Our work contributes toward such scrutiny. The outcome of our analysis shows how KPA methods fail to handle diversity. Potential technological improvements may lead to better results, but due diligence is required before applying such methods to real-world use cases.

We do not collect new data or involve human subjects in this work. Thus, we do not introduce any ethical considerations regarding data collection beyond those that affect the original datasets. A potential concern is that reproducing our results may involve using (possibly paid) services for running KPA. However, we aimed to make the analyses feasible with limited budget and resources.