News Editorials: Towards Summarizing Long Argumentative Texts

Shahbaz Syed [*] **Roxanne El Baff** [†] **Khalid Al-Khatib** [‡] **Johannes Kiesel** [‡]
Benno Stein [‡] **Martin Potthast** [*]

[*]Leipzig University [†]German Aerospace Center (DLR)
[‡]Bauhaus-Universität Weimar
<shahbaz.syed@uni-leipzig.de>

Abstract

The automatic summarization of argumentative texts has hardly been explored. This paper takes a further step in this direction, targeting news editorials, i.e., opinionated articles with a well-defined argumentation structure. With Webis-EditorialSum-2020, we present a corpus of 1330 carefully curated summaries for 266 news editorials. We evaluate these summaries based on a tailored annotation scheme, where a high-quality summary is expected to be *thesis-indicative*, *persuasive*, *reasonable*, *concise*, and *self-contained*. Our corpus contains at least three high-quality summaries for about 90% of the editorials, rendering it a valuable resource for the development and evaluation of summarization technology for long argumentative texts. We further report details of both, an in-depth corpus analysis, and the evaluation of two extractive summarization models.

1 Introduction

News summarization has been, and still is subject of active research to this day (Yao et al., 2017; Shi et al., 2018; Lin and Ng, 2019). However, the inverted pyramid structure of news reports, where the lead paragraphs often have a summarizing quality in and of themselves (PurdueOWL, 2019), induces a bias in many recent news summarization approaches to just copy the opening sentences (Kedzie et al., 2018; Jung et al., 2019). Hence, these approaches fail at argumentative news articles, such as editorials, whose structure differs from that of news reports.

Editorials represent the views of an organization (newspaper) on long-standing societal issues and aim to shape public opinion (Firmstone, 2019; Gajevic, 2016). Compared to news reports, which aim to inform objectively about current events, editorials subjectively assess a controversial topic in order to persuade its audience of a specific stance toward it (Van Dijk, 1995; Hynds and Archibald, 1996). This difference in their goals leads to a difference in linguistic choices. An editorial is usually composed of three discourse parts, namely lead, body, and conclusion (Van Dijk, 1992; Rich, 2015). The lead introduces the issue at hand by starting with an anecdote or a question. The body elaborates on arguments and background information, while the conclusion provides an evaluation as well as (possibly) implicit suggestions and calls to action (Bolívar, 2002). The summary of an editorial must hence be constructed with care in order to preserve its argumentative structure and its persuasive means. Research on automatic (news) summarization has so far neglected argumentative texts in general, and the genre of editorials in particular. With this paper we contribute to closing this gap, taking effective steps toward editorial summarization:

1. We define an annotation scheme tailored to editorial summaries, requiring a high-quality summary to be thesis-indicative, persuasive, reasonable, concise, and self-contained.
2. We create a corpus of 1330 summaries for 266 news editorials (five summaries each), manually acquired and evaluated by operationalizing the proposed annotation scheme.
3. We analyze each summary of the corpus with respect to content overlap, distribution of evidence types, adherence to the editorial structure, and annotator indications regarding summary quality.

4. We evaluate two unsupervised extractive summarization models (four variants total) in comparison to the acquired references, and their potential to identify an editorial's core message (the thesis).

The evaluation indicates a high suitability of the corpus for research and development: For 90% of the editorials there are at least three high-quality summaries, and for 52% all five are. The analyses also reveal that multiple summaries can be collected for an editorial with low content overlap, that good summaries include more *third party evidence* to justify an editorial's thesis, and that editorials' summaries have a distinct structure compared to those of news reports, with a specific contribution from each editorial discourse unit (lead, body, and conclusion). The corpus and other resources are publicly available.[1]

2 Related Work

The summarization of argumentative text has hardly been studied: Egan et al. (2016) automatically summarize political online debates by extracting key content from their arguments as "points" (verbs and their syntactic arguments). Similarly, Bar-Haim et al. (2020) propose to map crowdsourced arguments to "key points." They created the ArgKP corpus, containing 24,000 ⟨argument, key point⟩ pairs, extracted from the IBM-Rank-30k dataset (Gretz et al., 2020). To the best of our knowledge, no argument corpus comprises summaries of long-form monological argumentative text.

Most of the commonly used corpora for automatic news summarization, such as the NYT corpus (Sandhaus, 2008), Gigaword (Napoles et al., 2012), CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016), XSum (Narayan et al., 2018), and NEWSROOM (Grusky et al., 2018), primarily consist of (non-argumentative) news reports and only one ground truth summary per report. Although the DUC shared task datasets (Over et al., 2007) provide multiple summaries per document (500 news reports), they are very short (up to 14 words or 75 bytes), similar to Gigaword and XSum (up to two sentences). These corpora, stemming from the news domain, may contain some editorials; the ones in the NYT corpus were studied by Li et al. (2016), Al-Khatib et al. (2017), El Baff et al. (2018), and El Baff et al. (2020) for tasks such as summarization, analysis of rhetorical strategies, and argumentation quality assessment. But Li et al. (2016) observes that the accompanying summaries in this corpus are teasers rather than actual summaries. In our work, we focus on composing summaries exclusively for news editorials by aiming to capture their core argumentation, providing multiple and comparably longer ground truth summaries (20% of an editorial's segments) for each editorial.

A scheme for annotating argumentative roles of sentences for summarizing research articles was presented by Teufel et al. (1999). However, they only analyzed the effectiveness of this scheme and did not collect or evaluate any summaries. The key difference between other news summarization corpora and ours is the use of a (genre-specific) annotation scheme that unifies the summary acquisition and evaluation. Other summarization corpora lack such unification and only adopt the notion of "salience" (importance) of sentences in a text (Peyrard, 2019) to automatically extract summaries or crowdsource their acquisition (El-Haj et al., 2010). In the absence of a human-written ground truth, parts of a text, such as the title, highlighted sentences, or lead sentences, are used as proxies for summaries of the source documents. While such heuristics help create large corpora, the infeasibility of evaluating all the ground truth summaries leads to increased noise in the datasets, severely limiting the task of summarization and its evaluation (Kryscinski et al., 2019). We evaluated each summary in our corpus for its quality and provide labels for high (low) quality per quality dimension defined in our annotation scheme.

3 Annotation Scheme for Editorial Summaries

As per Hidi and Anderson (1986), humans produce two types of summaries: writer-based ones and reader-based ones. A writer-based summary is produced to facilitate one's own comprehension of a text. A reader-based summary intends to inform others about a text's core message, possibly to evoke further interest in the reader. News, in particular, may also be accompanied by a teaser, namely an incomplete summary that aims to attract people to read the entire news article (report or editorial) (Li et al., 2016). In extreme cases, teasers can become clickbait (Potthast et al., 2016), constructed to manipulate their

[1] https://webis.de/publications.html\#?q=COLING+2020

readers to visit an online news article (e.g., by invoking strong curiosity). For our corpus, we strive for reader-based summaries.

The intention of a reader-based summary is often to substitute the original text. For informational texts, such as news reports, this is roughly performed by the omission of irrelevant sentences (deletion), the subsumption of details into higher-level categories (generalization), and the integration of details into topic sentences (construction) (Kintsch and Van Dijk, 1978). Reorganization and rewording are possible (Johnson, 1983), but new ideas must not be introduced (Brown and Day, 1983; Kintsch and Van Dijk, 1978). In this regard, an editorial aims to persuade its readers of one central claim (thesis) through its monological argumentation (Al-Khatib et al., 2016; Wachsmuth et al., 2018). It is composed of argumentative discourse units (ADUs, typically statements) that form arguments to support the thesis (Peldszus and Stede, 2013; Stab and Gurevych, 2014). These arguments implement the author's strategy, incorporating not only logical, but also emotional and credible means of persuasion (Aristotle, translated 2007). The core message of an editorial corresponds to its thesis and its most persuasive segments; thus, an editorial's summary—and that of long argumentative texts in general—should aim to preserve both. We propose an annotation scheme tailored for editorial summaries, defining five quality dimensions that emphasize argumentation as well as summarization quality:

1. *Thesis-indicativeness.* The thesis of an editorial can be stated as a call for action or as an opinion (Van Dijk, 1992). The summary should thus explicitly contain the thesis or indicate it.
2. *Persuasiveness.* As the goal of an editorial is to persuade, the same applies to its summary. As per Wachsmuth et al. (2017), the summary should aim to be effective (i.e., aim to persuade the target audience of its thesis).
3. *Reasonableness.* The summary should help its audience to reach the thesis and rebut plausible counter-arguments to it.
4. *Conciseness.* A summary should be significantly shorter than the editorial and lack any superfluous phrasing or information.
5. *Self-containedness.* A summary should be comprehensible with general knowledge, without referring to additional resources.

Altogether, we strive to compile a corpus of editorial summaries that come close to the following definition:

> A high-quality summary of an editorial *indicates its thesis*, argues for this thesis in a *persuasive* and *reasonable* manner, and is *concise* yet *self-contained*.

Defining such an annotation scheme as a prerequisite allowed us to collect high-quality summaries and evaluate editorial summarization approaches in a unified manner. Nevertheless, just as for other kinds of summarization, it is subjective to determine the "core" parts of a text (Winograd, 1984). This circumstance is prevalent in editorials, where the argumentative structure and even the thesis might not be explicitly stated, leaving room for interpretation. Therefore, both the data collection and evaluation must not rely on a single ground truth summary. Below, the operationalization of our annotation scheme is described.

4 Summary Acquisition

Based on a corpus of news editorials that have previously been annotated with regard to argumentative discourse units, we crowdsource the generation of multiple reader-based summaries using Amazon's Mechanical Turk. The summarization is framed as an annotation / extraction task, where segments from an editorial are selected to compose a summary.

Data Source The news editorials corpus of Al-Khatib et al. (2016) forms the data source of our study. It comprises 300 editorials from three different news portals: Al Jazeera, Fox News, and The Guardian. After reviewing them, we omitted 34 ones falsely labeled as editorial, and very short ones. Each editorial has been segmented and annotated via crowdsourcing with the following argumentative discourse unit (ADU)

ADU type	Example
Assumption	Many have simply lost faith in global climate negotiation summits such as COP 20 starting in Lima, Peru, today.
Anecdote	We were in-between lessons during our first class, when we suddenly heard the sound of shooting.
Common-Ground	Politicians are meant to act in the interests of their people.
Statistic	In the early 1900s, Argentina ranked among the world's top 10 in per capita income.
Testimony	"I saw my brother drown in front of my eyes," said Hamid.

Table 1: Examples of ADUs (evidence types) selected from different editorials.

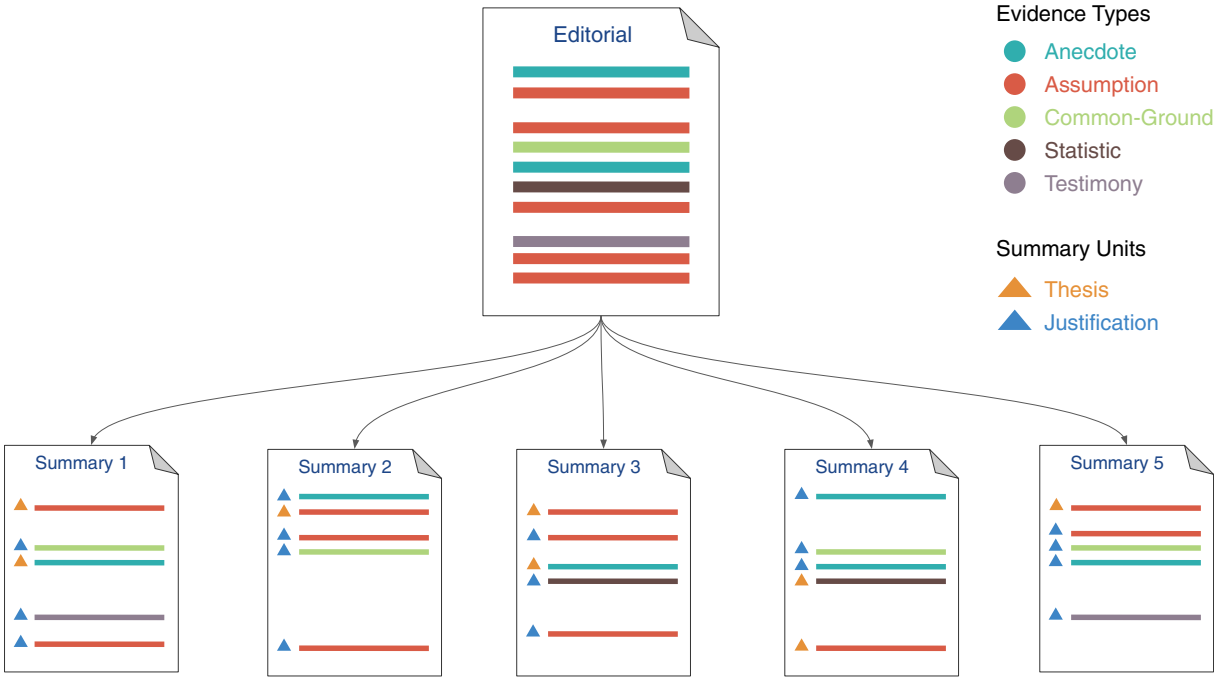


Figure 1: An editorial is comprised of multiple ADUs (evidence types), a subset of which are extracted to compose summaries. Each extracted ADU serves either as thesis or justification in the summary (summary units). Annotators can select up to two ADUs as the thesis.

types: anecdote, assumption, common ground, statistics, testimony, and other (collectively "evidence types", see Table 1). We adopt these ADU segments as our selection units for creating summaries (see Figure 1), since they are (mostly) well-formed texts by definition (Al-Khatib et al., 2016). We choose this corpus because its annotations enable our detailed argumentation analysis of the acquired summaries as well as our evaluation of automatic summarization models. Although each editorial in the corpus is accompanied by a very short summary (one sentence), extracted automatically from its web page, these summaries are insufficient to study their argumentative nature.

Annotation Task The manual selection of summary segments often relies on the concept of importance or salience, i.e., the importance of each segment decides whether it is to be included in the summary (Hardy et al., 2019; El-Haj et al., 2010). However, alongside capturing important content, the summary should also adhere to our annotation scheme. To operationalize this in the summary acquisition process, we specifically asked the workers to label each editorial segment as one of:

1. *Thesis*: segments that represent what the author wants to persuade the reader of.
2. *Justification*: segments that support the thesis.
3. *Background*: segments that provide background information to the reader.

4. *Not-in-summary*: segments that should not be in the summary.

Summary length was limited to be 20% of an editorial's segments. In many editorials, the thesis may not be explicitly stated but rather implied by the author. For this reason, we allowed up to two segments to be labeled as thesis, which allows for inspecting the worker agreement on the editorial's core message.

We also asked each worker to self-assess (1) their prior knowledge of the editorial's topic (background), (2) if they agreed with the author's opinion (stance), (3) their general interest in the topic (interest), and (4) the persuasiveness of the editorial (persuasiveness). These questions constitute a profile of the worker, which allows for a profile-dependent analysis of the summaries. A similar profile was considered in evaluating spoken argumentation by Jovičić (2004), where the effectiveness of the conveyed arguments depended on the audience. In our case, it turns out that the quality of our summaries is indirectly influenced by the workers (more details below).

Pilot Study We carried out a pilot study with 25 editorials, one editorial per human intelligence task (HIT), to check if our initial guideline required any revisions. Each editorial was annotated by five workers, resulting in 125 summaries. We did not show the segments' evidence types (from our data source) to avoid any selection bias. However, we excluded the segments of an editorial annotated as "Other" as these segments are not argumentative. Although this somewhat affects the readability, most argumentative segments are well-formed, and our emphasis on the composition of useful and self-contained summaries in the guideline mitigates this problem to some extent.

From the 125 summaries, we obtained a total of 1180 summary segment labels: 19.66% thesis, 54.15% justification and 26.19% background. We found that 28% of unique summary segments were annotated interchangeably as justification or background. Thus, to simplify the final annotation, alongside the thesis label, we only used the justification label to annotate the segments that either support the thesis or provide background information.

Final Annotation Using the three labels thesis, justification, and not-in-summary (to undo previous selections), we acquired summaries for the remaining 241 editorials. Similar to the pilot study, each editorial was annotated by five workers, and to ensure quality, we chose workers with an approval rating of at least 98% and 1000 accepted HITs from three native English speaking countries (US, UK, and Canada). We chose these countries, in particular, to render the task more relevant to workers, since most editorials discuss topics related to these regions. This is further reflected in the self-assessment questionnaire, where 76.11% of the workers stated that they have sufficient background knowledge about the editorial topics they annotated.

Thesis Agreement In general, the agreement among summaries is expected to be low, not necessarily due to poor annotations, but due to the subjectivity of the importance notion (Hardy et al., 2019; Mani, 2001) and argumentative text perception. Besides, agreement tends to further decrease as the length of the summary increases (Jing et al., 1998). However, the agreement on the thesis segment(s) indicates that the workers agree on the core message of the editorial. Hence, we consider worker agreement only on their selected thesis segments. As workers can label up to two segments as thesis, we consider full (two common segments) and partial (one common segment) agreement. As shown in Table 3a, the 61% majority agreement is promising, considering the challenging nature of the task, especially that a thesis can be indirectly implied when not explicitly stated in the editorial.

Our corpus consists of 1330 summaries having 12,806 labeled segments, with 14.7% labeled as thesis and 85.3% as justification. Table 3b shows the summary lengths in terms of segment and word counts.

5 Evaluation of Summaries

Adherence to the DUC Guideline Manual qualitative evaluation of summaries is often carried out according to the DUC guideline (Dang, 2005): summaries must be grammatical, non-redundant, exert referential clarity, and have focus, as well as structure and coherence. Gillick and Liu (2010) found this to be an expensive and a rather difficult task for non-experts. Thus, many summarization studies either avoid manual evaluation completely, or carry out only partial studies, rendering comparisons across papers

Quality Dimension	Explanation	Majority Agreement	Summaries
Thesis-relevance	The thesis is relevant to the title, i.e., it could be the main point(s) of the editorial with the given title.	76%	81.7%
Persuasiveness	A persuasive summary aims to convince its readers to take a stand on a particular topic. To this end, it uses persuasion techniques such as: providing logical arguments to support its stand, invoking certain emotions on the readers, and/or using effective phrases.	76%	86.5%
Reasonableness	A reasonable summary adequately supports its thesis, i.e., the thesis is supported by a sufficient number of arguments.	79%	89.8%
Self-containedness	A self-contained summary is understandable by most of the readers, i.e., no need for additional information to get its thesis and follow its argumentation. Also, a self-contained summary refers to entities (people, locations, events, etc.) without any confusion in the usage of pronouns.	74%	84.1%
Overall score	–	74%	82.4%

Table 2: Summary quality dimensions and the guideline given to workers, derived from our annotation scheme to render it comprehensible to non-experts. Thesis-relevance is an indirect assessment of *thesis-indicativeness*. The majority agreement column shows percentages of at least 2/3 agreement on each dimension, and for the overall score. The last column shows the percentage of summaries per editorial (averaged over all 266), which satisfy the corresponding quality dimension. On average, 82.4% of the summaries are high-quality per editorial, i.e., they satisfy at least three quality dimensions.

difficult (Hardy et al., 2019). Even ground truth summaries themselves are rarely evaluated. To ensure the quality of our corpus, we therefore thoroughly evaluate each acquired summary for quality.

We argue that, by the construction of the summaries, their grammaticality and non-redundancy are sufficiently fulfilled, and that, by definition of our annotation scheme, the remaining DUC criteria are covered. Our summaries inherit the grammaticality of the editorials they were derived from. They are sequences of ADUs extracted from the editorials, and although ADUs may be part of longer sentences, they do form complete sentences in and off themselves (Al-Khatib et al., 2016). Similarly, non-redundancy is inherited from the editorials; given their high writing quality, we can expect less redundant text, whereas if a certain point is repeated in an editorial to emphasize it, it stands to reason its summary may do so. Local redundancies, such as repeated names where an anaphora would suffice, cannot be avoided, since we did not ask the crowd workers to revise the summaries. Referential clarity, focus, structure, and coherence form part of our annotation scheme: Assessing the reasonableness of a summary includes checking for justifications to support its thesis, which is an indirect judgment of a summary's focus, i.e., the summary contains only related segments that together support its thesis. Likewise, assessing if a summary is self-contained considers referential clarity (i.e., no confusing usage of pronouns) as well as structure and coherence (i.e., the summary is well-organized).

Evaluation Task Similar to the acquisition of the summaries, we crowdsourced their qualitative evaluation. Table 2 shows how we explained the different quality dimensions of our annotation scheme to the crowd workers. The judgments for each dimension were made on a four-point scale (strongly disagree, weakly disagree, weakly agree, strongly agree). For persuasiveness, owing to the infeasibility of measuring this for some (often unknown) target audience (Wachsmuth et al., 2017), we restrict this dimension to being persuasive in general. Similarly, reasonableness of argumentations in theory also includes their acceptability by the target audience (Wachsmuth et al., 2017). Again, due to the infeasibility of measuring this for our summaries, we restrict this dimension to having adequate justifications for their thesis. All (five) summaries of an editorial were evaluated in one HIT, each performed by three workers with the same selection criteria as in the summarization task. We only showed an editorial's title alongside each summary,[2] with its thesis emphasized in bold.

Pilot Study. To test and revise our guideline, we carried out another pilot study to evaluate the 25 editorials and their summaries from the summary acquisition pilot study. Regarding thesis-indicativeness,

[2]Reading titles only instead of the whole editorials significantly reduced the time taken for a HIT.

for each summary's thesis, workers judged its relevance to the shown title, rather than reading the whole editorial. This design decision was backed by manually inspecting each title to ensure that it sufficiently indicates the issue discussed in the corresponding editorial. Acknowledging worker feedback, we included examples to help judge reasonableness and self-containedness, while only the description shown in Table 2 sufficed for judging persuasiveness.

For a sanity check, we exploit the fact that each of an editorial's five summaries is supposed to have the same or at least a similar thesis (Table 3a). Specifically, we asked workers to judge how similar in meaning is the thesis of a particular summary to that of the remaining summaries in a HIT. Then, given two summaries comprising similar thesis segments, we rejected the submissions of workers who judged them to be dissimilar.[3]

Annotator Agreement To compute annotator agreement, we first mapped all judgments to numeric scores (strongly-disagree: -2, weakly-disagree: -1, weakly-agree: 1, strongly-agree: 2). Then, we computed an overall score for a summary by averaging the numeric scores of all its quality dimensions. Thus, a summary with multiple quality dimensions gets a higher score. Table 2 shows the majority agreement for each quality dimension, as well as for the high-quality summaries. We note sufficient agreement on all quality criteria with the highest value for the reasonableness dimension.

In Table 3c, we also report significant correlations (at $p < 0.05$) between quality dimensions and overall score. We observe that: (1) A reasonable summary is also self-contained. By including justifications that build upon its thesis, a reasonable summary mitigates any distractions in its argumentation flow, thus rendering it understandable. (2) A reasonable summary is also persuasive. A key persuasion technique by authors is providing logical arguments to support their stances, i.e., reasonable summaries where a sufficient number of justifications is provided are more likely to be persuasive.

Quality Groups We distinguish summaries as being high or low quality based on the workers' assessments. We assert that a high-quality summary has at least three quality dimensions defined in our annotation scheme (Section 3) as judged by workers. As each summary is assessed by three workers, they may disagree on which dimensions it has. Thus, we use the majority vote to label a summary as "high-quality," i.e., if at least two workers agreed that it has at least three quality dimensions. We similarly distinguished the summaries per quality dimension (e.g., high/low-thesis-indicativeness). The distribution of high-quality summaries per editorial (on average, by quality dimension) is shown in Table 2. As for the quality dimensions, we observe that all dimensions are (almost) equally distributed in high-quality summaries, with reasonableness dimension favored slightly more (26.25%) (thesis-indicativeness: 24.08%, persuasiveness: 24.98%, self-containedness: 24.69%). Totally, we have 1096 high-quality and 234 low-quality summaries as per our manual evaluation.

6 Corpus Analysis

In this section, we present a thorough analysis of our corpus, exploring (1) summary content overlap (i.e., the annotator agreement), (2) distribution of evidence types, (3) adherence of summaries to the editorial's structure, and (4) annotators' profiles and their impact on the quality of summaries.

Summary Content Overlap Here, we inspect the overlap among the five summaries per editorial. We first computed the Jaccard index[4] between each pair of summaries, in which the Jaccard index measures the intersection over the union between a pair's segments. Then, we averaged the indices over all the summary-pairs for an editorial (five summaries, ten pairs). We found that the average of Jaccard indices over all editorials is 0.2, which speaks for a low overlap between summaries. Although the workers agreed on the thesis (Table 3a), they still chose different justifications, leading to diverse summaries.

Distribution of Evidence Types We examine the distribution of evidence types in our summaries by obtaining the ADU labels from our data source. Table 4 shows this distribution in the three discourse parts (i.e., lead, body, and conclusion),[5] and in high-quality and low-quality summaries according to their role

[3]This sanity check was repeated multiple times to ensure reliable judgments.

[4]Jaccard index has an interval of [0,1]; values close to 1 indicate high overlap between two sets.

[5]After inspecting the length of *online lead paragraphs* from the NYT corpus (Sandhaus, 2008), which are on average 12% of the article's length, we considered 15% of the top and bottom segments of an editorial as its lead and conclusion.

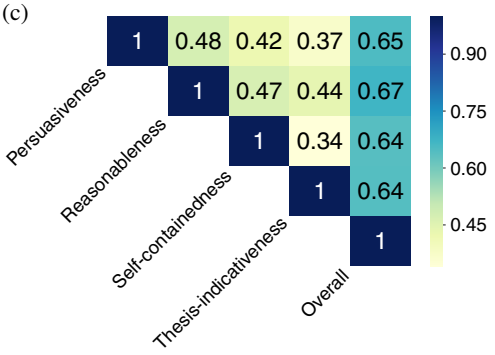
(a)

Workers	Editorials
2/5	96%
3/5	61%
4/5	25%

(b)

Length	Min	Mean	Max
Words	71	209.3	492
Segments	4	9.6	26

(c)



(d)

Position	Summary Segments		
	Thesis	Justif.	Combined
Lead	73.2%	20.2%	28.3%
Body	21.5%	67.6%	60.6%
Conclusion	5.2%	12.2%	11.1%

Table 3: (a) Agreement on thesis segment(s) among five workers. Values are computed on all 266 editorials (1330 theses). (b) Length statistics of all the summaries. (c) Correlation among judgments (Kendall's τ) for various quality dimensions including overall summary score. All values are significant at $p < 0.05$. (d) Percentages of summary segments (by function as thesis or justification and combined) extracted from lead, body and conclusion.

ADU type	Editorials				High-quality summaries			Low-quality summaries		
	Lead	Body	Concl.	Combined	Thesis	Justif.	Combined	Thesis	Justif.	Combined
Assumption	61.7	66.9	79.2	68.0	70.4	64.2	65.1	66.5	66.3	66.3
Anecdote	25.1	18.9	8.5	18.2	18.8	18.9	18.9	23.3	20.7	21.1
Common-Ground	2.0	1.8	1.4	1.7	1.1	1.4	1.4	1.6	1.1	1.2
Statistics	2.9	3.1	1.6	2.9	1.8	4.4	4.0	1.6	2.5	2.4
Testimony	7.6	8.5	6.0	8.0	7.5	10.6	10.1	6.4	8.3	8.0

Table 4: Comparison of ADU distributions (in %) in editorials (by occurrence in lead, body, or conclusion and combined) as well as in groups of high and low-quality summaries w.r.t overall quality (by function as thesis or justification and combined).

as thesis, justification, and combined.

The table reveals two key insights into the summaries' evidence types: (1) High-quality summaries have more statistics than the low-quality ones. Statistics is an evidence type stating or quoting the results or conclusions of quantitative research, studies, empirical data analyses, or similar (Al-Khatib et al., 2016). (2) High-quality summaries have more testimony than the low-quality ones. Testimony is an evidence type that either states or quotes propositions made by some expert (person or organization) other than the author (Al-Khatib et al., 2016). Although the overall percentage of statistics and testimony is less compared to other evidence types, such as assumption or anecdote, it is interesting to note that workers preferred more third-party evidence in their summaries. In conventional news summarization, these contents may been seen as extraneous details that need not be in a summary, whereas for editorials, they play a crucial role of supporting the thesis as justification.

Adherence to Editorial Structure We argue that constructing editorial summaries requires considering the specific contributions of its discourse parts to the argumentation. This means that unlike news reports where the summary is condensed primarily in the lead (Wasson, 1998), important contents are distributed throughout the editorial. As shown in Table 3d, the majority of thesis segments are extracted from the lead (73.2%), but are insufficient to fully summarize the editorial, comprising only 28.3% of the combined summary segments. Furthermore, we note that each discourse part contributes proportionally to its summary (28.3%, 60.6% and 11.1% for lead, body, and conclusion).

Worker Profiles' Impact on Quality As mentioned in Section 4, all workers employed in the acquisition task were assessed on a five-point scale regarding their background knowledge of the editorial's topic, if they agree or disagree with the editorial's stance, their interest in the topic, and if they find it persuasive.

To understand the impact of the workers' profiles on the quality of their summaries, we computed the

Model	Length (words)	Editorial Position			Thesis Coverage						Summary Coverage
		Lead	Body	Concl.	1	2	3	4	5	Maj.	
TextRank-Lex	79.9	13.6	**69.3**	**17.1**	67.3	23.1	7.7	1.9	0.0	9.6	11.9
TextRank-Entity	141.4	**40.6**	58.5	0.9	41.0	21.6	14.9	13.4	9.0	**37.3**	**20.1**
ExtSum-XLNet	151.0	23.4	60.5	16.2	34.4	29.7	23.6	8.5	3.8	35.8	16.4
ExtSum-DistilBERT	155.0	22.5	64.8	12.7	36.2	30.5	21.9	7.6	3.8	33.3	16.2
References	**209.3**	21.6	65.3	13.1							

Table 5: Average summary length in words. Average distribution of summary segments extracted by models from lead, body and conclusion in comparison to references. Percentage of (reference) theses and summary segments covered by models. For thesis coverage, we inspected if a thesis is completely included in the automatic summary (i.e., both segments). Accordingly, as each editorial has five theses, we also show coverage by number of theses completely captured in the model's summary. Summary coverage is the percentage of unique summary segments (from all five summaries of an editorial) captured.

correlation (Kendall's τ) between each aspect of their profile and the summaries' quality dimensions.[6] With a significant positive correlation ($p < 0.05$), we found that the workers who have more background knowledge of an editorial composed more persuasive summaries. On the other side, we did not find any significant correlation between the workers' stance toward a topic and the persuasiveness of a summary (as well as the overall quality) of their summaries.

7 Automatic Extractive Summarization of News Editorials

In this section, we investigate the capability of automatic summarization technology for generating high-quality summaries for editorials. Specifically, we implemented two unsupervised extractive summarization models (TextRank and ExtSum) and evaluated their output based on the Webis-EditorialSum-2020 corpus. These models emulate the manual summary acquisition setting, i.e., extracting segments within a given length budget. The input for each summarization model was the argumentative segments in an editorial (without any information about their evidence type). We set the same summary length (20%) for the automatic summaries as the ground truth ones.

Summarization Models Our first summarization model is based on TextRank (Mihalcea and Tarau, 2004), an unsupervised summarization model based on PageRank (Brin and Page, 1998). Petasis and Karkaletsis (2016) demonstrated that TextRank is able to identify argument components in a text. By comparing the connections among sentences with those between claims and premises, they established TextRank as a suitable model for argument mining.

Accordingly, we leverage this to create extractive summaries of the editorials. TextRank first constructs an undirected graph of the entire editorial with the segments as nodes. For weighing the connecting edges, we investigated two similarity functions resulting in two variants of TextRank: `TextRank-Lex` which uses lexical overlap among segments and `TextRank-Entity` which uses the number of common named entities[7] between two segments.

As our second summarization model, we adopt an extractive summarization model based on BERT (Devlin et al., 2019) that clusters (using K-Means) the contextual embeddings of an editorial's segments and selects those that are closer to its centroid as the final summary (Miller, 2019). To encode the editorial segments, we chose contextual embeddings from two distinct architectures: `ExtSum-XLNet` based on XLNet (Yang et al., 2019), an autoregressive language model that outperforms BERT on several tasks, and `ExtSum-DistilBERT` based on DistilBERT (Sanh et al., 2019). DistilBERT is an efficient language model that leverages knowledge distillation to achieve similar performance as BERT but with significantly fewer resources and increased speed compared to XLNet.

[6] We converted each judgment to a numerical score as in Section 5.

[7] We used Spacy's *en-core-web-md* model for tagging named entities.

7.1 Model Evaluation

We compare each model's summary for an editorial with its multiple references in terms of its adherence to the editorial's structure and coverage of unique summary (and theses) segments.

Regarding the structure of the automatic summaries, the distribution of segments from lead, body, and conclusion is shown in Table 5. We see that `TextRank-Lex` extracts more segments from the body and the conclusion. However, it extracts much shorter segments than those in the references. In contrast, `TextRank-Entity` extracts more segments from the lead of an editorial and produces longer summaries. This is because the actors of an editorial (named entities) are usually introduced in the beginning. Both the ExtSum variants have almost a similar distribution of extracting segments from the editorial's discourse parts; besides that, embeddings from the smaller DistilBERT produce relatively longer summaries. Still, all the automatically produced summaries are shorter than the references in terms of word count.

The coverage of the references' theses and summary segments in the summaries of each model is shown in Table 5. We observe that `TextRank-Entity` has the highest coverage of the reference summary segments. Despite producing shorter summaries than the ExtSum models, it also consistently captures a majority of theses. This reveals a plausible segment extraction strategy followed by workers in the summary acquisition task, where argumentative segments connecting different actors are often selected. Among the ExtSum models, `ExtSum-DistilBERT` has a similar distribution of segments from the discourse parts as the references, with `ExtSum-XLNet` having a slightly higher coverage of the unique summary segments from the references.

8 Conclusion

This paper takes the first steps towards summarizing news editorials, a type of long form argumentative text. We introduce an annotation scheme tailored to editorial summaries, which we employ to acquire and evaluate the Webis-EditorialSum-2020 corpus; the first corpus for news editorial summarization containing five summaries per editorial (1330 summaries in total). Our annotation scheme defines multiple quality dimensions grounded in argumentation quality studies. Through detailed corpus analyses, we find that editorial summaries have a distinct structure compared to those of news reports; that third party evidence in summaries improves their overall quality; that background knowledge of workers is positively correlated to the persuasiveness of their summaries; and, that some automatic models can at least capture an editorial's thesis.

We consider our corpus a useful resource for promoting research in automatic summarization and computational argumentation. As next steps, we plan to investigate automatically identifying evidence types in other long-form argumentative texts such as debates, social media posts, and student essays. Also, we intend to leverage our findings (i.e., adherence to an editorial's structure and the distribution of evidence types) to develop novel summarization models, tailored to argumentative texts, that capture argumentative aspects alongside salience.

References

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A News Editorial Corpus for Mining Argumentation Strategies. In *26th International Conference on Computational Linguistics (COLING 2016)*, pages 3433–3443. Association for Computational Linguistics, December.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of Argumentation Strategies across Topics. In *2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 1362–1368. Association for Computational Linguistics, September.

Aristotle. translated 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039. Association for Computational Linguistics, July.

Adriana Bolívar. 2002. The structure of newspaper editorials. In *Advances in written text analysis*, pages 290–308. Routledge.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks*, 30(1-7):107–117.

Ann L Brown and Jeanne D Day. 1983. Macrorules for summarizing texts: The development of expertise. *Journal of verbal learning and verbal behavior*, 22(1):1–14.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Charlie Egan, Advaith Siddharthan, and Adam Z. Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus. In *22nd Conference on Computational Natural Language Learning (CoNLL 2018)*, pages 454–464. Association for Computational Linguistics, October.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online, July. Association for Computational Linguistics.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2010. Using mechanical turk to create a corpus of arabic summaries.

Julie Firmstone. 2019. Editorial journalism and newspapers' editorial opinions, 03.

Slavko Gajevic. 2016. Journalism and formation of argument. *Journalism*, 17(7):865–881.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA, June 6, 2010*, pages 148–151. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7805–7813. AAAI Press.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3381–3392.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Suzanne Hidi and Valerie Anderson. 1986. Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of educational research*, 56(4):473–493.

Ernest Hynds and Erika Archibald. 1996. Improved editorial pages can help papers, communities. *Newspaper Research Journal*, 17(1-2):14–24.

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *AAAI symposium on intelligent summarization*, pages 51–59. Palo Alto, CA.

Nancy S Johnson. 1983. What do you do if you can't tell the whole story? the development of summarization skills. *Children's language*, 4:315–383.

Taeda Jovičić. 2004. Authority-based argumentative strategies: a model for their evaluation. *Argumentation*, 18(1):1–24.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard H. Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3322–3333. Association for Computational Linguistics.

Chris Kedzie, Kathleen R. McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1818–1828. Association for Computational Linguistics.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 540–551. Association for Computational Linguistics.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 137–147. The Association for Computer Linguistics.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 9815–9822. AAAI Press.

Inderjeet Mani. 2001. Summarization evaluation: An overview. In *Proceedings of the Third Second Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization, NTCIR-2, Tokyo, Japan, March 7-9, 2001*. National Institute of Informatics (NII).

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411.

Derek Miller. 2019. Leveraging BERT for extractive text summarization on lectures. *CoRR*, abs/1906.04165.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 280–290. ACL.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated english gigaword. *Linguistic Data Consortium, Philadelphia*.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1797–1807. Association for Computational Linguistics.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *IJCINI*, 7(1):1–31.

Georgios Petasis and Vangelis Karkaletsis. 2016. Identifying argument components through textrank. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1059–1073. Association for Computational Linguistics.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait Detection. In *Advances in Information Retrieval. 38th European Conference on IR Research (ECIR 2016)*, volume 9626 of *Lecture Notes in Computer Science*, pages 810–817, Berlin Heidelberg New York, March. Springer.

PurdueOWL. 2019. Journalism and journalistic writing: The inverted pyramid structure.

Carole Rich. 2015. *Writing and reporting news: A coaching method*. Cengage Learning.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. 2018. Neural abstractive text summarization with sequence-to-sequence models. *arXiv preprint arXiv:1812.02303*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510. ACL.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 110–117. Association for Computational Linguistics.

Teun A Van Dijk. 1992. Racism and argumentation: Race riot rhetoric in tabloid editorials. *Argumentation illuminated*, pages 242–259.

Teun A Van Dijk. 1995. Opinions and ideologies in editorials. In *4th International Symposium of Critical Discourse Analysis, Language, Social Life and Critical Thought, Athens*, pages 14–16.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 176–187, April.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation Synthesis following Rhetorical Strategies. In *The 27th International Conference on Computational Linguistics (COLING 2018)*. Association for Computational Linguistics, August.

Mark Wasson. 1998. Using leading text for news summaries: Evaluation results and implications for commercial summarization applications. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference.*, pages 1364–1368.

Peter N Winograd. 1984. Strategic difficulties in summarizing texts. *Reading Research Quarterly*, pages 404–425.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336.