Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources

Darina Benikova[§†], **Margot Mieskes**[§*], **Christian M. Meyer**[§‡], **Iryna Gurevych**[§‡]
[§]Research Training Group AIPHES
[†]Language Technology Lab, University of Duisburg-Essen
[*]Information Science, University of Applied Sciences, Darmstadt
[‡]Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt
https://www.aiphes.tu-darmstadt.de

Abstract

Coherent extracts are a novel type of summary combining the advantages of manually created abstractive summaries, which are fluent but difficult to evaluate, and low-quality automatically created extractive summaries, which lack coherence and structure. We use a corpus of heterogeneous documents to address the issue that information seekers usually face – a variety of different types of information sources. We directly extract information from these, but minimally redact and meaningfully order it to form a coherent text. Our qualitative and quantitative evaluations show that quantitative results are not sufficient to judge the quality of a summary and that other quality criteria, such as coherence, should also be taken into account. We find that our manually created corpus is of high quality and that it has the potential to bridge the gap between reference corpora of abstracts and automatic methods producing extracts. Our corpus is available to the research community for further development.

1 Introduction

The sheer amount of information available via news agencies, social media, scientific publications, etc. overwhelm information seekers. Not only the information overload itself, but also the heterogeneity of the data poses an obstacle to the aggregation and assessment of information, as text structures, styles, and forms of information presentation vary across different genres and domains of text. Therefore, automatic summarization of multiple heterogeneous sources is a key research demand.

Previous efforts such as DUC[1], TAC[2], and MultiLing[3] have mostly focused on homogeneous data. Heterogeneous data poses many new challenges to Multi-Document Summarization (MDS) that have not been extensively covered so far, such as abstracting from deviating text structures, maintaining topic diversity, resolving potentially opposing opinions, and adapting a text for different target audiences. In our work, we aim at creating a novel MDS corpus of heterogeneous sources.

Besides the lack of heterogeneous MDS corpora, we consider the type of summaries available in MDS corpora in general as an even more severe problem: Existing corpora contain primarily *abstractive summaries*, whereas automatic systems typically utilize extractive methods yielding *extractive summaries* (i.e., bags of sentences taken from the source documents). Producing a coherent, human-understandable text, however, requires steps beyond mere extraction. These steps include, among others, compression, co-reference resolution, paraphrasing, and content structuring. Though there are multiple works striving for abstractive summaries, it is very hard to evaluate them, since current MDS corpora do not explicitly address these tasks.

Although ROUGE is the primary evaluation metric used so far, it only compares n-gram overlap between at least one model summary and a reference summary and thus gives a mere *quantitative* impression of the lexical coverage, leaving the coherence of a summary largely unconsidered.

[1] http://duc.nist.gov
[2] http://www.nist.gov/tac
[3] http://multiling.iit.demokritos.gr

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1039–1050, Osaka, Japan, December 11-17 2016.

Corpus	Summary type	Genre	Lang.	Topics	Doc/Topic	Summary size
DUC (2001–2003)	Abstracts	News	en	30–60	10	50–400 words
DUC (2004)	Abstracts	News	en, ar	50	10	10–100 words
DUC (2005–2007)	Abstracts	News	en	50	≥ 25	250 words
TAC (2008–2011)	Abstracts	News, blogs	en	44	10	100 words
TAC (2014)	Abstracts	Scientific	en	50	10	250 words
MultiLing (2011; 2013)	Abstracts	News	7	10	10	240–250 words
MultiLing (2015)	Abstracts	News	10	15	10	240–250 words
Loupy et al. (2010)	Abstracts	News	fr	20	20	≈ 200 words
Hirao et al. (2004)	Abstracts	News	ja	N/A	N/A	5–10 % characters
Ulrich et al. (2008)	Abstracts & Extracts	e-Mails	en	30	≈ 11	250 words†
Goldstein et al. (2000a)	Extracts	News	en	25	10	≥ 10 sentences†
Zechner (2002)	Extracts	Trans. speech	en	23	N/A	N/A
Carenini et al. (2007)	Extracts	e-Mails	en	20	≥ 4	30 % sentences
Nakano et al. (2010)	Extracts	Heterogeneous	en	24	352	$\approx$ 2,560 characters
Lloret et al. (2013)	Extracts	Heterogeneous	en	310	10	100–200 words
Our work	Coherent Extracts	Heterogeneous	de	10	4–14	≈ 500 words†

Table 1: Comparison of manual multi-document corpora († = no predefined summary size)

In contrast, there are corpora containing only extracts, which are suitable for evaluating extractive MDS systems. However, this yields an overly artificial evaluation, since neither their creation guidelines nor the evaluation metrics employed take coherence, fluency, and organization of the summary into account. We argue that such bag-of-sentences-based extracts are of limited use for real users, who prefer a coherent, meaningful text over an unordered collection of sentences.

In order to bridge the gap between the system results (primarily extractive) and the reference data (primarily abstractive), we need high-quality corpora that allow for training and evaluation of automatic systems and their intermediate steps, such as the identification of important nuggets or the removal of redundancy. To this end, we introduce and analyze a novel type of reference summary: *coherent extracts*. Coherent extracts are based on information explicitly taken from the source documents allowing to create a straight-forward evaluation setup based on the selection of important content. At the same time, coherent extracts have a meaningful order and are redacted to ensure a coherent text, which is substantially more helpful for humans than purely extractive summaries. Our vision is to use the data collected during structuring and redaction in order to research improved extractive MDS systems.

In this paper, we introduce our idea of coherent extracts and we create and evaluate a novel MDS corpus of coherent extracts from heterogeneous source documents. In Section 2 we discuss previous work in manual summarization and corpus construction. We describe coherent extracts and our summarization workflow in Section 3 and we evaluate our work in Section 4. Finally, we conclude the paper in Section 5.

2 Related work

Our work is inspired by two main aspects of multi-document summarization: the manual creation of MDS corpora and the evaluation of the resulting summaries. There are two main approaches to summarization: abstractive and extractive. Humans typically produce abstractive summaries by paraphrasing and condensing information from the sources. Automatic systems primarily generate extractive summaries by copying information from the sources without modifying and ordering them. For evaluating an automatic system, however, we require gold standard summaries that fit well to the system output, cf. (Endres-Niggemeyer, 2012). This major discrepancy between the human-generated abstractive gold standard and the system-generated extractive summaries has largely been neglected in previous work.

2.1 Existing multi-document summarization corpora

Endres-Niggemeyer (2012) states that the process of summarization includes three subtasks: *analysis of the input information*, *performance of the core summarization task (condensation, abstraction)* and *representation of the results in an appropriate form*. Schlesinger et al. (2003, p. 46) observe that MDS "lacks complementary documentation of procedures and methodologies for human performance".

Table 1 gives an overview over various MDS corpora in comparison to our work. Most of them are based on homogeneous data such as news (DUC, TAC, MultiLing, Goldstein et al. (2000a), Hirao et al. (2004)), e-mails (Ulrich et al., 2008), and transcribed speech (Zechner, 2002). Lloret et al. (2013) take the top ten results from a regular web search engine as the source documents for their summaries, but do not specify the types of documents used. Therefore, it is unclear whether their data is heterogeneous or not. To the best of our knowledge, only Nakano et al. (2010) specifically use a collection of heteregeneous documents and discuss the corresponding challenges, such as subjectivity and contradictions.

Although the majority of the works presented in Table 1 mention guidelines for producing the summaries, most of them neither describe these guidelines in detail, nor divide the summarization process into clear-cut subtasks. Zechner (2002) gives descriptions for individual annotation tasks, but they are not necessarily subsequently applied. Additionally, their process is targeted towards the summarization of transcribed speech, which requires additional tasks, such as the annotation of topic boundaries, which are not necessary in thematically clustered document collections.

Lloret et al. (2013) use crowdsourcing and therefore the instructions are very limited. Their results show that summaries created via crowdsourcing are unsuitable for reference or gold standard summaries. Although the crowd authors are able to create summaries very quickly, if high-quality results are a priority, it is still advisable to employ trained annotators, despite the slower progress. Hirao et al. (2004) and Nakano et al. (2010) put an extraction task before the actual summarization, but there was no further division of the summarization process. They allow the editing of the text extracts without limitation, thus their summaries cannot be regarded as purely extractive. However, as they do neither provide any instructions on the summary creation process nor evaluate the coherence of the results, it is unknown whether their results are coherent summaries. Hirao et al. (2004) realize a three-step annotation process for producing summaries: (1) extracting important sentences, (2) minimizing redundant sentences, and (3) rewriting the result of the previous step in order to match the size constraint. Their goal is a summary that consists of a collection of informative sentences from the source text.

Another issue in the creation of summaries is coherence and cohesion. None of the methods mentioned above specifically addresses this. Although Zechner (2002, p. 456) states that the "overall goal [...] was to create a concise and readable summary", they do not specify how they reach the goal of coherent summaries. Neither Hirao et al. (2004) nor Zechner (2002) explicitly perform co-reference resolution, although the latter states that it would "certainly be desirable, for the sake of increased coherence and readability, to employ a well-working anaphora resolution component" (ibid., p. 451). Lloret et al. (2013, p. 346) observed that summaries based on crowdsourcing have a poor quality, as they "frequently present lost anaphoric and pronominal references". To our knowledge, none of the mentioned works address this issue in their evaluation.

2.2 Previous summary evaluation

Summaries are primarily evaluated intrinsically, either based on *qualitative* or *quantitative* measures. Early DUC evaluations (Over, 2001; Over and Liggett, 2002; Over and Yen, 2003) used Likert scales, which are psychometric response scales that are often used in questionnaires, for summary evaluation. This evaluation focuses on *qualitative* aspects of a summary. To perform a *quantitative* evaluation of the summaries, model units (Over and Liggett, 2002; Over and Yen, 2003) were used. These are manually annotated semantic units in human summaries, which are compared to automatic summaries according to the frequency of occurrence. Lloret et al. (2013) uses the overall quality for evaluation, based on a five-point Likert-type scale. Later on, Nenkova et al. (2007) introduced the Pyramid method. So-called *Summary Content Units (SCUs)*, which are semantically motivated, subsentential units, serve as the basis for the evaluation. They are variable in length, but not bigger than a sentential clause (Nenkova et al., 2007), and are closely related to model units. In both cases, system summaries are evaluated based on the content overlap rankings. A widely used *quantitative*, automatic evaluation metric is ROUGE (Lin, 2004), which has also been used to evaluate manual summaries, for instance, by Nakano et al. (2010), Lloret et al. (2013), and in DUC evaluations (up to 2007).

In our work, we focus on manual qualitative evaluation using various aspects of text quality, similar to

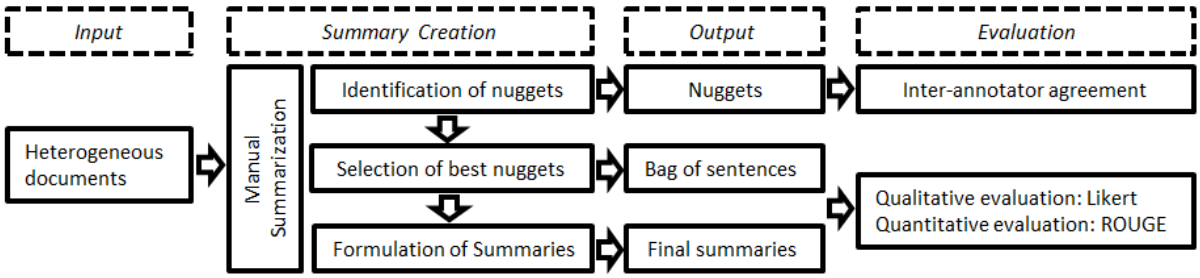

Figure 1: Process of creation and evaluation of coherent extracts and the intermediate processing steps.

the evaluation approach taken in the early DUC evaluations (Over, 2001; Over and Liggett, 2002; Over and Yen, 2003). More specifically, we focus on coherence in summary evaluation and statistically prove that this feature, which is very important for human understanding of text, has not only been neglected in the creation, but also in the evaluation process. In particular, we focus on coherence in our evaluation, which has previously been neglected in the creation process. Our results indicate that this feature is very important for human understanding and should therefore gain more importance.

3 Coherent Extracts

We propose *coherent extracts* as a new type of reference summary, bridging the gap between human-optimized abstracts and (automatic) extracts lacking coherence and fluency. Coherent extracts consist of important nuggets extracted from (multiple) source documents, which are meaningfully ordered and minimally redacted to ensure a coherent text that is close to a completely manually written text.

To substantiate this idea, we create a novel MDS corpus of coherent extracts, which we make freely available.[4] Figure 1 shows an overview of our corpus creation and evaluation procedure discussed in the remaining paper. First, we introduce the *Input* in form of the source document collection and topic selection procedure (Section 3.1). Our MDS process consists of three steps: In the first step the most important units are selected (so-called *Identification of Nuggets*). Next, the selected nuggets are clustered into groups with similar content and for each cluster the best nugget is selected (*Selection of best nuggets*). Finally, during the *Formulation of Summaries*, the best nuggets are co-reference resolved, grammaticalized and sorted coherently. Details for each step are given in Section 3.2 and in Meyer et al. (2016). In Section 4, we describe the *Evaluation* of our work using inter-annotator agreement measures, automatic summarization scoring with ROUGE, and human judgments based on Likert scales.

3.1 Heterogeneous Document Collection

Deutscher Bildungsserver[5] (DBS) is a large web portal that collects links to educational resources in German. As a publicly funded project, it fulfills an important information need of different stakeholders, including teachers, students, parents, politicians, and educational researchers. DBS contains links to highly heterogeneous text types, such as interviews, book reviews, teaching material, NGO profiles, newspaper and scientific articles. Although all topics focus on the educational domain, they vary enormously in terms of audience and specialized subject area (e.g., educational reforms, social impact, educational research, political decisions). Domain experts organize the links according to different topics, different audiences (teachers, students, parents, etc.) and strive for introducing each topic with a brief overview text. However, writing a manual overview for each topic is hardly feasible, as there are hundreds of topics, which are continuously updated by the editors. As we consider the processing of heterogeneous documents as a particularly important research challenge, the DBS makes an excellent basis and use case for research in automatic MDS of heterogeneous sources.

For our corpus, we select 10 topics from DBS and crawl their linked pages yielding between 4 and 14 documents per topic. Table 2 shows the properties of each topic and its documents. Our choice

[4] https://github.com/AIPHES/DBS
[5] http://www.bildungsserver.de

Topic	Source documents #	Sentences	Tokens	Genres News-like	Other	Summaries #	∅ Sentences	∅ Tokens
1. Preventing violence	10	1,446	12,276	2	8	3	39 ± 22	625 ± 343
2. Environmental education	10	184	2,446	2	8	3	12 ± 7	216 ± 37
3. Healthy Nutrition	13	894	6,551	2	11	4	23 ± 14	431 ± 278
4. Lateral entry teachers	5	134	1,950	4	1	4	20 ± 11	482 ± 266
5. Reform of the dual education system	10	432	5,972	2	8	3	26 ± 17	485 ± 263
6. Mediation	7	253	2,651	1	6	3	30 ± 13	453 ± 187
7. German Qualifications Framework	4	127	2,089	3	1	3	16 ± 3	310 ± 71
8. Reading	12	1,270	20,212	9	3	4	78 ± 56	1509 ± 1056
9. Short-term secondary school diploma	14	522	5,275	6	8	3	28 ± 18	576 ± 393
10. Right-wing extremism and racism	7	176	1,933	2	5	3	18 ± 7	336 ± 93
Average per topic	9.2	544	6,136	3.3	5.9	3.3	30 ± 20	566 ± 378
Total sum	92	5,438	61,355	33	59	33	987	18,694

Table 2: Overview of the topics, documents, genres, and summaries in our MDS corpus

of topic and document sizes is guided by the parameters of similar corpora (e.g., the DUC corpora) and by practical constraints of the DBS portal. The genre columns highlight the heterogeneity of the individual topics and the corpus overall. Although there is at least one news-like article in each topic, the source documents are highly heterogeneous, covering brief introductory texts, syllabi, term definitions, presentations of interest groups, and many other text genres.

Once we identify the source documents, we shorten those with more than 40 pages and topics with more than 30,000 tokens in order to reduce the annotation effort. We do not remove documents in order to conserve the full diversity of subtopics and genres. Conversely, we define a minimum of 3 documents and 1,500 tokens to ensure that there is enough information to summarize. We use the boilerplate removal tool by Habernal et al. (2016) to remove HTML markup and non-content (e.g., advertisement, navigation menu).

3.2 Creating Coherent Extracts

We divide the task of creating coherent extracts into three subsequent steps, which in turn consist of further substeps which we explain in an extensive, publicly available annotation guidebook. These detailed instructions ensure a reliable, reproducible annotation workflow and allow us to produce informative, non-redundant, grammatically correct coherent extracts.

Step 1: Identification of nuggets We first instruct the annotators to read each text *before* annotating, to ensure that they understand the full context of the topic. Afterwards, they identify important *nuggets* in the source documents. Our notion of nuggets includes a grammatical and a semantic constraint. The grammatical constraint is that a nugget has to contain at least one verb and one corresponding noun and can maximally span over one sentence. The minimum restriction is based on the assumption that this would facilitate the reformulation of nuggets to grammatically correct sentences as part of the third annotation step (see below). Our semantic constraint of nuggets is similar to previous work on *model units* (Over and Liggett, 2002), *factoids* (Halteren and Teufel, 2003), and *SCUs* (Nenkova et al., 2007) in that they should be important in the context of the given topic.

Step 2: Selection of best nuggets *Redundancy* is a more severe issue in MDS (Goldstein et al., 2000b; Goldstein et al., 2000a) compared to single-document summarization, as recurring facts are more likely to occur in multiple different documents about the same topic than within a single one. Thus, we ask the annotators to group the selected nuggets from step 1 if they "convey the same meaning using different wording" (Bhagat and Hovy, 2013, p. 463). For selecting the *best nugget* of a group, the annotators are required to prefer declarative, objective, and co-reference-free nuggets wherever possible.

Step 3: Formulation of summaries We divide the formulation of the coherent extract into several substeps: First, we ask the annotators to resolve co-references in the best nuggets in order to prevent references to context which are not part of any nugget. This is an important requirement for producing

coherent extracts, although it has been largely ignored in previous works.

Subsequently, the annotators reformulate the nuggets into full, grammatically correct sentences, which form the basis for our coherent extract. The annotators perform as few changes as necessary to prevent them from adding additional semantic content. Typical transformations are the introduction of function words or clarifying indirect speech (e.g., explicitly adding phrases, such as *"according to John Doe"*).

Once the best nuggets are transformed into grammatically correct sentences, the annotators order them in a meaningful way to structure the extract. Depending on the topic, annotators may choose a topic-based or argumentation-based information flow. Based on this structure, the annotators formulate the final extracts. They may add discourse connectives and conjunctions, such as *after*, *however*, or *as well as*. They may re-introduce co-references, if the referring object, person or event is available in the context, to create a fluent and readable text that closely resembles manually written texts. We do not restrict the length of the resulting summaries.

Annotation It is important to note that each annotation step builds upon the results of the previous steps (e.g., the structuring of the sentences builds upon the set of reformulated best nuggets). To implement this annotation setup, we use the MDS*Writer* tool recently introduced by Meyer et al. (2016).

Our annotation team consists of four German native speakers, who are graduate students in linguistics or computational linguistics. We trained the annotators using two smaller topics with 2 and 5 documents. For the actual corpus creation, we assign at least two annotators to each topic and we instruct them to finish all annotation steps of a specific topic before moving to the next one.

Final Corpus Our final corpus consists of 92 input documents with over 61,000 tokens and 33 multi-document summaries with over 18,000 tokens (566 per summary on average). As we did not restrict the length of the summaries, their texts range from 6 to 127 sentences and from 173 to 2,494 tokens. The average token compression rate is 12.2 %. During the summarization process, the annotators identify between 15 and 228 nuggets per topic (48 on average, 1,596 in total), which they process in the later steps in order to remove redundancy and co-references, formulate complete sentences and ultimately the coherent extracts. Our corpus is freely available from our GitHub project page (`https://github.com/AIPHES/DBS`).

4 Evaluation

We evaluate our corpus creation procedure in three experiments: In Section 4.1, we measure the inter-annotator agreement to assess the quality of the identification of nuggets. In Section 4.2, we conduct a quantitative evaluation of our coherent extracts in comparison to bag-of-sentence-based and automatically created summaries using ROUGE. In Section 4.3, we present results on a qualitative evaluation of the three summary types using a range of quality criteria judged on a five-point Likert scale. Finally, Section 4.4 compares the results of the qualitative analysis with the results from the quantitative analysis.

4.1 Inter-Annotator Agreement of the Nugget Identification Step

To ensure the reproducibility of our corpus creation procedure, we first assess how well the annotators agree on the results of the nugget identification step. The agreement scores reveal to what extent the annotation experiment can be repeated with different annotators or data. Furthermore, they yield insight into the degree of subjectivity of the task. Besides the average proportion of observed agreement (A_{O}) of nuggets jointly identified by two annotators, we use the standard metrics from content analysis, namely Fleiss' κ and Krippendorff's α, as defined by Artstein and Poesio (2008).

We first model the nugget identification step as a coding task (i.e., assigning categories to predefined units) by aggregating each annotator's nuggets to the levels of sentences and paragraphs. That is, a sentence or paragraph is considered important, if it contains at least one nugget. This setup allows us to compute the usual A_{O}, κ and α agreement scores. Although κ has been used in some previous works, it is not well-suited for this setup, as it has no explicit notion of missing values. Rather, it assumes that all annotators processed the entire corpus, which is neither true for our work nor for many other previous

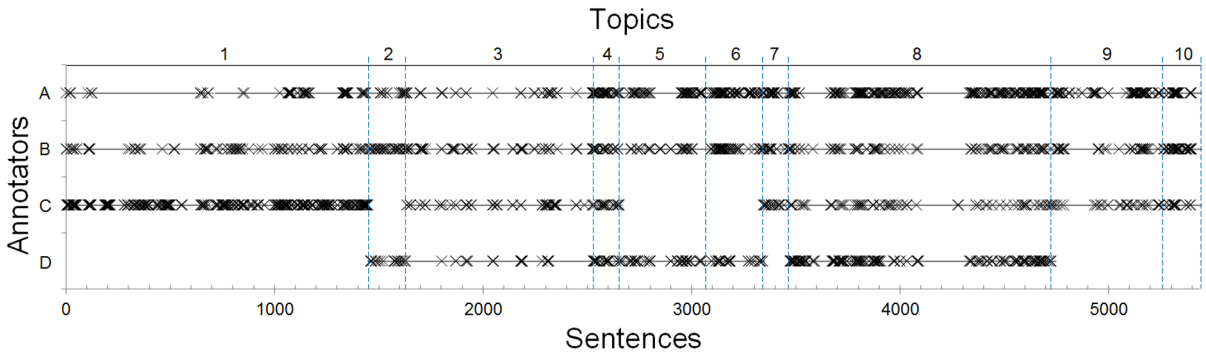

Figure 2: Plot marking all sentences, in which an annotator identified at least one nugget

works, such as Zechner (2002). We therefore recommend to always report Krippendorff's α as well, which uses a similar scale, but explicitly deals with missing annotations.

Since our nugget identification procedure is not limited to selecting entire sentences but to mark spans of varying lengths starting at arbitrary positions, we also model the nugget identification step as a unitizing task (i.e., finding the margins of the unit to be annotated) in order to take the inter-annotator agreement on the nugget boundaries into consideration. Krippendorff's α_{U} is a suitable metric for such setups (Krippendorff, 1995), as it uses the same scale as the regular α and therefore allows for comparison. All measures were calculated using DKPro Statistics (Meyer et al., 2014).[6]

Figure 2 illustrates the resulting annotation. Every $\times$ marks a sentence (bottom x-axis; pooled over all topics indicated by the top x-axis) containing at least one nugget as annotated by one of the four annotators (y-axis). The gray line indicates sentences, which do not contain any nugget, although they have been assigned to a particular annotator. Empty space thus counts towards the missing values mentioned above – i.e., sentences that have not been assigned to the corresponding annotator. The plot shows that the annotators largely vary in the number of nuggets they consider important: Annotator C identified more nuggets in the sentences with low index, while annotator A more extensively marked nuggets in sentences with higher index (Note that the annotators did not perform the task strictly in order of this sentence index). When zooming in, we can observe a fair overlap of sentences considered important or unimportant by a majority of annotators, despite the fact that annotators rarely fully agree on what is important. There are, however, sentences that are very consistently considered unimportant for writing a summary, for example, in topic 8 around sentence index 4200, which is a scientific article about reading with overly detailed information and a list of references.

Table 3 shows the inter-annotator agreement scores in the coding and unitizing setups. Our annotators agree that 85 % of the sentences either contain an important nugget or are not important at all. This is substantially higher than the 56–62 % reported by Goldstein et al. (2000a) and in line with what could be observed in Figure 2. Although agreement scores for nugget identification are rarely reported, we find that our κ score of 0.23 exceeds the $\kappa = 0.13$ reported by Zechner (2002). It is not surprising to find rather low overall κ and α scores, since judging importance is known to be a subjective task. Additionally, we face a highly skewed class distribution between important and unimportant sentences. Disagreement on the smaller class has a high impact on the final κ and α scores. For the unitizing setup, the discrepancy between observed agreement and statistically corrected agreement becomes extreme: Given the large parts of the documents for which no annotator identified a nugget, the observed unitizing disagreement is only 1 %, whereas α_{U} is only 0.19. Comparing α and α_{U} is highly interesting, because the small difference indicates that the annotators mostly agree on a nugget's boundaries. The main decision is therefore *what* to judge as important, not the specific nugget boundaries.

Table 3 also shows the inter-annotator agreement scores individually for each topic. We find a higher agreement for topics with less sentences and tokens (e.g., topic 4), where the annotators mostly agree on the important nuggets. Instead, we observe much lower agreement scores for large topics, in which

[6] https://dkpro.github.io/dkpro-statistics/

IAA setup / Topic ID		1	2	3	4	5	6	7	8	9	10	Overall
Paragraph level	A_O	0.92	0.91	0.97	0.86	0.92	0.89	0.93	0.91	0.95	0.93	0.93
	κ	0.17	0.30	0.34	0.43	0.33	0.39	0.38	0.44	0.41	0.49	0.35
	α	0.17	0.31	0.34	0.43	0.33	0.39	0.38	0.44	0.41	0.49	0.34
Sentence level	A_O	0.82	0.79	0.93	0.74	0.86	0.75	0.75	0.84	0.86	0.81	0.85
	κ	0.10	0.16	0.28	0.29	0.31	0.25	0.18	0.19	0.22	0.31	0.23
	α	0.10	0.16	0.28	0.29	0.31	0.25	0.18	0.19	0.23	0.31	0.22
Nugget level	A_O	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	α_U	0.01	0.07	0.25	0.43	0.17	0.21	0.04	0.16	0.13	0.19	0.19

Table 3: Inter-annotator agreement for our 10 topics at the paragraph, sentence, and nugget levels

α	A	B	C	D
A	—	0.26	0.17	0.28
B	0.26	—	0.17	0.28
C	0.17	0.17	—	0.14
D	0.28	0.28	0.14	—

α_U	A	B	C	D
A	—	0.29	0.18	0.24
B	0.29	—	0.15	0.22
C	0.18	0.15	—	0.05
D	0.24	0.22	0.05	—

Table 4: Pairwise inter-annotator agreement using sentence-based α (left) and nugget-based α_U (right)

the annotators put different foci or in which they identified an overly high or low number of nuggets. Especially for topic 1, the α_U drops extremely, because of the different number of identified nuggets: 49 by annotator A, 76 by B, and 228 by C. Moreover, we find that the specificity of the topic has a large impact on the inter-annotator agreement. Broad topics such as *Reading* (topic 8) tend to have lower agreement than specific topics (e.g., topic 5, *Reform of the dual education system*) with a very clear focus.

In Table 4, we additionally assessed the agreement for each pair of annotators individually. We find the sentence-based α to range from 0.14 to 0.28 and the nugget-based α_U to range from 0.05 to 0.29 across the annotator pairs. Annotator C achieved consistently lower agreement with the others. When looking into the data, we note that this annotator persistently chose more nuggets per topic (58 on average) than the other annotators (45 on average). The qualitative evaluation of the annotator's coherent extracts revealed high quality. Given that our agreement figures are clearly higher than reported for previous corpora, we consider our identified nuggets reliable and our procedure reproducible. However, we also note that the identification of important nuggets is a fairly subjective task. Thus, producing *at least* 3 reference summaries is essential to cover the different notions of what is important.

4.2 Quantitative Evaluation

Quantitative evaluations based on ROUGE have been used both for the evaluation of automatic systems (e.g., in the context of DUC) and for manual summaries, for instance by Nakano et al. (2010) and Lloret et al. (2013). In our evaluation, we compare results on three different summary types: First, we create automatic summaries using various standard summarization methods, including TextRank (Mihalcea and Tarau, 2004), LexRank (Erkan and Radev, 2004), and LSA (Steinberger and Ježek, 2004) as implemented in the SUMY[7] tool.[8] Second, we use *bag of sentences (BoS)*, which are an intermediate result of our manual summarization process, namely the reformulated best nuggets (see Section 3.2). These BoS essentially correspond to purely extractive MDS summaries, as they contain full sentences, which then serve as the basis for coherent extracts. They are already ordered in a meaningful way and annotators are required to not add content, but to only add (for example) connectives in order to create a coherent text (see Section 3.2 for details).[9] Third, we use the coherent extracts, which are the final results of our summarization process. In order to evaluate automatic summaries and BoS we use the manual summaries as references. For the manual summaries, we take individual manual summaries and use the remaining

[7] `https://github.com/miso-belica/sumy`

[8] We used standard settings, except for the length, which we wanted to keep comparable across all summary types.

[9] Examples for the various summary types (automatic, BoS and manual) are available for download.

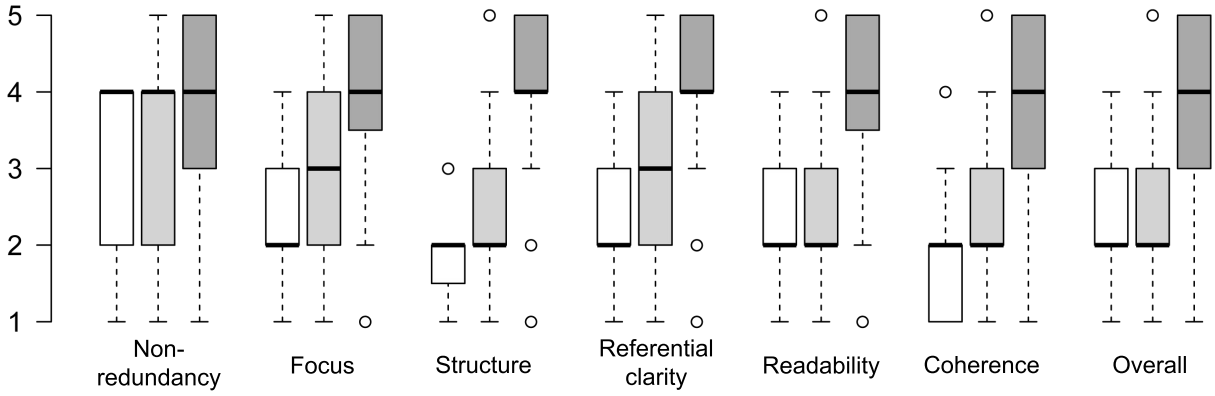

Figure 3: Boxplot of our qualitative judgments of automatic summarization systems (white), bags of sentences (light gray), and coherent extracts (dark gray)

manual summaries as references. The motivation for the latter step is to gain an intuition about the upper bound for the ROUGE scores, but also about the deviation based on individual differences.

We compute the ROUGE-1 recall metric (R-1) for each summary of the three types, in which we use the coherent extracts as peers. For BoS and coherent extracts, we remove the summary of the same author from the set of peers. Our coherent extracts achieve R-1 = 0.42, which is very similar to the R-1 = 0.43 reported by Nakano et al. (2010), but considerably higher than the R-1 = 0.35 discussed by Lloret et al. (2013) for their crowdsourcing setup. BoS achieve the best results with R-1 = 0.49, followed by the coherent extracts (R-1 = 0.42) and then the automatic summaries ($0.32 \leq$ R-1 ≤ 0.35). The scores of automatic methods are comparable to Lloret et al. (2013), but significantly lower than the BoS of coherent extracts. BoS, which are the basis for the coherent extracts, achieve the best results, as they vary less. For the coherent extracts, we asked our annotators to minimally revise the text for readability, fluency, and coherence, which causes a slightly lower n-gram overlap and thus lower R-1 score.

4.3 Qualitative Evaluation

We complement our ROUGE-based evaluation by manually rating the three different summary types on a five-point Likert scale. Our quality criteria are based on early DUC evaluations, but slightly extended for our purposes: We judge *non-redundancy*, *focus*, *structure*, *referential clarity*, *readability*, *coherence*, *length*, *grammaticality*, *spelling*, *layout*, and *overall quality* for 25 automatically produced summaries (as introduced in the previous section) and 16 BoS and coherent extracts from our corpus. We asked six human raters, of mixed gender and age, with different academic degrees in linguistics or computational linguistics, to participate in the evaluation. To prevent a bias, we randomly shuffled the summaries, such that the participants did not know how a summary was created.

Figure 3 shows a boxplot of the average scores for each summary type according to the seven most relevant quality criteria. *Non-redundancy* achieves good results across all summary types. This means that the automatic approaches as well as the two manual approaches (BoS and coherent extracts) manage to compile a mostly redundancy-free summary. However, both automatic approaches and the BoS-based approach fail to achieve high ratings for the other six quality criteria. The criteria *structure* and *readability* vary the most with coherent extracts achieving significantly higher scores than automatic summaries and BoS. *Referential clarity* is – at least for the manual summaries – the most consistently marked evaluation criterion. Additionally, we see that both BoS and automatic methods achieve considerably lower results than the coherent extracts. Especially for the automatic methods, the distribution of values is broadest, whereas for the BoS the distribution is more even. For *focus*, *structure*, and *referential clarity*, we observe a clear pattern with the lowest scores for automatic summaries, slightly better results for BoS, and the best scores for our coherent extracts. In terms of *readability* and *coherence*, both BoS and automatic summaries have similarly low scores, whereas the coherent extracts clearly achieve adequate results. The *overall* score reflects the results on individual quality criteria, in that automatic methods achieve the lowest scores, BoS are slightly better and coherent extracts are by far the best.

4.4 Discussion

Comparing the qualitative and the quantitative evaluation results, we find substantial differences between the automatic summaries, BoS, and coherent extracts. According to ROUGE, the BoS are better than coherent extracts, whereas the manual judgments suggest the opposite. This indicates that even high-quality, redundancy-free BoS with reasonable ROUGE scores are still not good enough to provide reasonable summaries for humans, as they lack in coherence, structure, and focus. We compute the correlation between ROUGE and all criteria judged by human raters, but no rater showed significant results, except for one. This annotator achieved very high scores on the qualitative evaluation, which also shows a significant correlation with ROUGE. It has been shown in the past by Liu and Liu (2010) that ROUGE does not reflect human judgements. Our results support this. But in addition, we show that: (a) BoS as mostly produced by automatic systems are not considered high-quality by humans, (b) although they do achieve very good ROUGE scores and (c) that quality criteria such as *coherence* and *structure*, which are important factors in human judgements, are not covered by a state-of-the-art ROUGE-based evaluation.

The coherent extracts we propose, take these quality criteria into account, as they achieve overall high scores in the manual evaluation. At the same time, they are not restricted to a ROUGE-based evaluation as is the case for abstractive summaries. Rather they allow for evaluating the identification of nuggets (i.e., content selection), redundancy removal, and the formulation of a coherent and fluent summary separately using the intermediate results collected during our corpus creation process. We render this highly important for taking extractive summarization methods to the next level.

5 Conclusion and Further Work

We introduced coherent extracts as a novel type of summary and presented a corresponding MDS corpus based on heterogeneous sources from the educational domain. To compile this corpus, we defined a workflow addressing the identification of important nuggets, the removal of redundancy, and the formulation of a coherent text in separate steps. For each step, we collect the individual results, which allows us to evaluate the corpus creation steps individually as well as their overall quality. We used inter-annotator agreement measures and ROUGE to quantitatively analyze the corpus and we relied on the human judgments for 11 quality criteria on a five-point Likert scale to assess the quality of our corpus and the value of coherent extracts. Our results show that ROUGE overestimates the quality of purely extractive summaries, which especially lack coherence, readability, and structure.

At the same time, our corpus provides data for evaluating automatic MDS systems beyond the typical ROUGE-based setup, because we have an explicit notion of important elements in the input documents and we can trace which sentence of the summary belongs to which source document. Though the latter is also possible with purely extractive summaries, they are not well-suited for human use and thus yield a rather artificial evaluation setup. At the same time, our detailed workflow allows researchers working on MDS systems to reconstruct how humans go beyond mere BoS-type summaries and incorporate this into automatic systems and evaluate the resulting quality. To this end, coherent extracts bridge an important gap between fluent, but difficult-to-evaluate abstracts and low-quality extracts lacking coherence and structure.

In future work, we plan to increase the size of the corpus, such that we can investigate the effects of certain summarizers, topics, genres, or even multiple languages. To complement our human judgments, we also consider to use the Pyramid method (Nenkova et al., 2007) in order to understand different means of summary evaluation for MDS of heterogeneous document collections. Furthermore, we plan to enrich our corpus with abstractive multi-document summaries written by experts. On the one hand this would give us the possibility to compute a qualitative comparison with the new kind of summary described in this paper, on the other hand it would provide data for more accurate evaluation and training.

Acknowledgments

This work has been supported by the DFG-funded research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES, GRK 1994/1). We would like to thank our annotators for their valuable contribution.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472.

Giuseppe Carenini, Raymond T. Ng, and Xiaodong Zhou. 2007. Summarizing email conversations with clue words. In *Proceedings of the 16th International World Wide Web Conference (WWW)*, pages 91–100, Banff, AB, Canada.

Brigitte Endres-Niggemeyer. 2012. *Summarizing Information*. Springer.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Jamie Callan. 2000a. Creating and evaluating multi-document sentence extract summaries. In *Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM)*, pages 165–172, McLean, VA, USA.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000b. Multi-Document Summarization By Sentence Extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization*, pages 40–48, Seattle, WA, USA.

Ivan Habernal, Omnia Zayed, and Iryna Gurevych. 2016. C4Corpus: Multilingual Web-size corpus with free license. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 914–922, Portorož, Slovenia.

Hans van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the 2003 HLT-NAACL Workshop on Text Summarization*, pages 57–64, Edmonton, AB, Canada.

Tsutomu Hirao, Takahiro Fukusima, Manabu Okumura, Chikashi Nobata, and Hidetsugu Nanba. 2004. Corpus and Evaluation Measures for Multiple Document Summarization with Multiple Sources. In *Proceedings of the 20th international conference on Computational Linguistics (COLING)*, pages 535–541, Geneva, Switzerland.

Klaus Krippendorff. 1995. On the reliability of unitizing contiguous data. *Sociological Methodology*, 25:47–76.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL-2004*, pages 25–26, Barcelona, Spain.

Feifan Liu and Yang Liu. 2010. Exploring correlation between ROUGE and human evaluation on meeting summaries. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):187–196.

Elena Lloret, Laura Plaza, and Ahmet Ake. 2013. Analyzing the capabilities of crowdsourcing services for text summarization. *Language Resoures and Evaluation*, 47(2):337–369.

Claude de Loupy, Marie Guégan, Christelle Ayache, Somara Seng, and Juan-Manuel Torres Moreno. 2010. A French Human Reference Corpus for Multi-Document Summarization and Sentence Compression. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 3113–3118, Valletta, Malta.

Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych. 2014. DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING): System Demonstrations*, pages 105–109, Dublin, Ireland.

Christian M. Meyer, Darina Benikova, Margot Mieskes, and Iryna Gurevych. 2016. MDS*Writer*: Annotation tool for creating high-quality multi-document summarization corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 97–102, Berlin, Germany.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain.

Masahiro Nakano, Hideyuki Shibuki, Rintaro Miyazaki, Madoka Ishioroshi, Koichi Kaneko, and Tatsunori Mori. 2010. Construction of Text Summarization Corpus for the Credibility of Information on the Web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 3125–3131, Valletta, Malta.

Ani Nenkova, Rebecca Passonneau, and Kathleen Mckeown. 2007. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4.

Paul Over and Walter Liggett. 2002. Introduction to DUC: An Intrinsic Evaluation of Generic News Text Summarization Systems. In *Proceedings of DUC 2002*, Philadelphia, PA, USA.

Paul Over and James Yen. 2003. An Introduction to DUC 2003 Intrinsic Evaluation of Generic News Text Summarization Systems. In *Proceedings of DUC 2003*, Edmonton, AB, Canada.

Paul Over. 2001. Introduction to DUC-2001: an Intrinsic Evaluaton of Generic News Text Summarization Systems. In *Proceedings of DUC 2001*, New Orleans, LA, USA.

Judith D. Schlesinger, John M. Conroy, Mary Ellen Okurowski, and Dianne P. O'Leary. 2003. Machine and Human Performance for Single and Multidocument Summarization. *IEEE Intelligent Systems*, 18(1):46–54.

Josef Steinberger and Karel Ježek. 2004. Using latent semantic analysis in text summarization and summary evaluation. In *Proceedings of the 5th International Conference on Information Systems Implementation and Modelling (ISIM)*, pages 93–100, Rožnov pod Radhoštěm, Czech Republic.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A Publicly Available Annotated Corpus for Supervised Email Summarization. In *Enhanced Messaging: Papers from the 2008 AAAI Workshop*, Technical Report WS-08-04, pages 77–82. Menlo Park, CA: AAAI Press.

Klaus Zechner. 2002. Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics*, 28(4):447–485.