Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6690–6699
Marseille, 11–16 May 2020

A Data Set for the Analysis of Text Quality Dimensions in Summarization Evaluation

Margot Mieskes[1], **Eneldo Loza Mencía**[2], **Tim Kronsbein**[2]
[1] Darmstadt University of Applied Sciences, Germany
[2] Knowledge Engineering Group, Technische Universität Darmstadt, Germany

margot.mieskes@h-da.de, research@eneldo.net, tim-kronsbein@gmx.de

Abstract

Automatic evaluation of summarization focuses on developing a metric to represent the quality of the resulting text. However, text quality is represented in a variety of dimensions ranging from grammaticality to readability and coherence. In our work, we analyze the dependencies between a variety of quality dimensions on automatically created multi-document summaries and which dimensions automatic evaluation metrics such as ROUGE, PEAK or JSD are able to capture. Our results indicate that variants of ROUGE are correlated to various quality dimensions and that some automatic summarization methods achieve higher quality summaries than others with respect to individual summary quality dimensions. Our results also indicate that differentiating between quality dimensions facilitates inspection and fine-grained comparison of summarization methods and its characteristics.
Keywords: Text Quality, Summarization Evaluation, Multi-document Summarization Data Set

1. Introduction

Work on text summarization evaluation focused strongly on creating automatic methods for evaluation, which correlate well with human judgements. Early work on ROUGE showed a good correlation between ROUGE 1 and manual quality evaluations (Lin and Hovy, 2003). But those evaluations focused on only few quality dimensions. Text quality has to be measured in a range of dimensions, which capture various aspects of a text. The question that arises is, which of the various quality dimensions are represented by the various automatic evaluation metrics when used without reference summaries. An additional question is what type of automatic evaluation method is best suited for summarization of large-scale heterogeneous data sets and which method achieves the highest scores on the various quality dimensions. We aim at answering these questions by a set of experiments in order to determine in what way various quality dimensions are represented by evaluation measures without reference summaries and which automatic summarization method gives good results on the various quality dimensions on a large-scale heterogeneous data set.
Our contributions are as follows:

- A data set of manual annotations of summaries – using both Likert scale evaluations as well as pairwise comparisons – of a range of various automatic summarization methods.
- An analysis on both evaluations with respect to the various quality dimensions and the various summarization systems.
- We make the set of evaluations and automatic summaries available to the research community.[1] This can, e.g., used to train evaluation systems.

To the best of our knowledge, it is the first data set of judgements of automatic multi-document summarization systems on a large variety of quality dimensions.

2. Related Work

This work draws on a range of areas and previous work in automatic summarization. The first area is evaluation. Early evaluation of summarization was done manually, i.e., by Okumura et al. (2003) and further extended, e.g., by Liu et al. (2011) and Li et al. (2012). Later, automatic methods such as ROUGE were developed, which is by now a standard evaluation method (Lin and Hovy, 2003). ROUGEhas various flaws (see for example (Graham, 2015), (Sjöbergh, 2007) and (Zopf, 2019)) and during the AESOP track other methods for evaluation were proposed, but no method became widely used.[2] As ROUGE is based on counting n-grams, a method to more closely evaluate content was developed. The PYRAMID method (Nenkova and Passonneau, 2004) is a manual procedure that uses content as a basis for evaluation. Implementations such as PEAK (Yang et al., 2016) aim at automatically performing the PYRAMID evaluation, removing the necessity for the time-consuming manual annotation procedure.
A common criticism regarding ROUGE is that it often does not correlate as well to human judgements as originally claimed (see (Chaganty et al., 2018; Zopf, 2019; Böhm et al., 2019) for recent analyses). Therefore, the question how to learn a summary quality evaluation function directly from human judgements obtained increased attention recently. While Peyrard and Gurevych (2018) and Gao et al. (2019a) still relied on human annotations simulated by ROUGE scores in their experiments, Peyrard et al. (2017) used annotations from TAC-2008 and TAC-2009 and Böhm et al. (2019) used the evaluations of summaries of 500 topics of a single-document news summarization corpus (Chaganty et al., 2018). In the latter experiment, five human annotators (crowd workers) were asked to rate for each of the

[1] Available at https://github.com/keelm/DIP-SumEval and https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2303.

[2] https://tac.nist.gov/2011/Summarization/AESOP.2011.guidelines.html

topics the fluency, redundancy level and overall quality of four automatically generated summaries and one reference summary on a 3-point Likert scale. The annotation experiment in this work similarly uses a 5-point Likert scale but on a larger number of criteria (11, and 6 for the pairwise comparisons) and for multi-document summaries.

The third area is based on automatic summarization methods, which are primarily based on extracting sentences or parts of sentences from the original document(s). Methods such as MEAD (Radev et al., 2004) are not new, but are still used as baselines. Recently, abstractive summarization methods have been developed with increasing success. Details about other extractive summarization methods used in our work are given in Section 3.2 below. One of the first approaches for abstractive multi-document summarization which uses neural encoder-decoder architecture, PG-MMR, was proposed by Lebanoff et al. (2018). It is based on a special type of pointer networks and the maximal marginal relevance algorithm, which was already very successful for extractive summarization (Carbonell and Goldstein, 1998). We refer to Lin and Ng (2019) for an overview over recent abstractive summarization systems.

The fourth area concerns the data usually used in automatic summarization research. Most work has been done on newswire articles (i.e., in the context of the DUC/TAC[3]). There has been work on other genres as well, partially also in the context of the DUC/TAC tasks (i.e., scientific publications) or in the context of the MultiLing task (i.e., encyclopaedic documents based on Wikipedia[4]). Most of the data sets have in common that they only contain documents from one genre. Very little work has been done on summarizing (large-scale) heterogeneous data sets. Zopf et al. (2016b) introduced a semi-automatically produced corpus (hMDS) which takes Wikipedia articles as seed for the retrieval of relevant documents regarding the topic. This idea was further extended (Zopf, 2018a) in order to automate the whole process, which lead to a large-scale multi-document summarization corpus (auto-hMDS). Liu et al. (2018) follow a similar approach and bootstrapped a MDS corpus (WikiSum) with more than 2.2 million topics, which is not publicly available.

3. DIP-SumEval: A Data Set of Human Summary Evaluations

In the following we describe the data, the summarization systems and the evaluations we used in our experiments.

3.1. Multi-document Summarization Data Set

The basis for our experiments is the DIP 16 Corpus presented by Habernal et al. (2016). This contains heterogeneous, topically clustered documents dealing with educational topics crawled from the web including blogs, forums, scientific articles and more. In total, we have 49 topics containing a total of 3984 documents with 40 to 100 English documents per topic. Approximately 1/4 of the originally 628,026 sentences were marked as relevant for the respective topic. This corpus was initially presented in the context of summarization by Tauchmann et al. (2018), but did not contain manual reference summaries. Instead, it contained topically and hierarchically grouped *information nuggets*, which ranged from very general information to very specific information. These annotations were crowdsourced for 10 of the 49 topics, resulting in 4,983 nuggets.

3.2. Summarization Systems

Following the idea of using hierarchically structured information for automatic summarization (Tauchmann et al., 2018) we implement five different systems exploiting this type of structure. We follow the reasoning as outlined by Tauchmann et al. (2018) with respect to large-scale heterogeneous sources, where hierarchical ordering of extracted information nuggets can either be regarded as a generic summary – in the case of the top-most levels – or as a collection of aspect-based summaries – in the case of individual trees. Details about the approaches and differences among the five systems (named *H1* to *H5*) are described in detail below. The documents and annotated entities from all topics are used for training the systems. There is no restriction on the type of summary generated, except the limitation of at most 600 characters. However, all systems but one generate summaries by extracting single sentences from the documents. In addition, systems can generate a title, which does not count towards the total number of characters.

In the first quality evaluation phase (see Section 3.3.1) we use two baselines in addition to the five summary systems, namely the beginning of a random document from the document base as summary (*LeadFirst*) and a classic approach based on maximal marginal relevance (*MMR*). We extend the set of baselines for our second annotation experiment (Section 3.3.2) by a stronger extractive approach based on submodular functions (*Submodular*) and a recent abstractive summarization system (*PG-MMR*).

All hierarchical systems consist of three steps: First, identify and select relevant and informative nuggets with respect to the topic. For training purposes all systems primarily use the sentences labeled as relevant by the original corpus as described in (Tauchmann et al., 2018). Various other techniques are employed which we describe in the following. Information nuggets are hierarchically structured in order to model relationships between topics and sub-topics. Finally, the systems generate an extractive summary out of a subset of the hierarchically ordered nuggets. The resulting hierarchies offer both diversity, when considering nodes close to the roots. Or they maintain focus, namely when focusing on individual sub-trees which contain information on a specific aspect contained in the data.

H1 The first hierarchical architecture uses 300-dimensional Word2Vec embeddings (Mikolov et al., 2013). An ensemble of neural networks with different hyper parameters each calculate a relevance score for each sentence. The sentences with the highest mean score are taken into further consideration as relevant sentences. A set of heuristics based on phrases indicating a summarizing or contrasting sentences such as "on the other hand" or "however" are used to identify sentences for the resulting summary. To avoid redundancy only sentences with a cosine distance of more than 0.2 to the sentences already selected are used.

[3] https://duc.nist.gov/, https://tac.nist.gov//

[4] http://multiling.iit.demokritos.gr/pages/view/1532/task-mss-single-document-summarization-data-and-information

H2 uses pairwise learning for nugget selection. Two sentences are analysed for their relevance by a GRU and afterwards compared. To build the hierarchy, the universal sentence encoder (USE) (Cer et al., 2018) calculates sentence embeddings, which are then clustered agglomeratively. The clustering algorithm also weights sentences by their TF-IDF scores regarding the corpus. The hierarchy then simply connects the closest most central sentences.

To summarize a topic, H2 combines the previously calculated relevance score and clustering. Sub-trees are ranked based on size and mean relevance. Then, the nuggets closest to the query according to the clustering are added to the summary. Finally, the selection of sentences is ordered by centroid centrality.

H3 first groups sentences via LDA topic modeling. Sentences in the same topic are ranked by their semantic similarity to each other based on their sentence embedding according to the USE. The highest ranked sentences are clustered by k-means to build the hierarchy. Afterwards, multiple summaries using different traversal and ordering modes are generated. Out of these, the one with the highest Jensen-Shannon divergence to the source documents is returned as the final summary.

H4 uses 300-dimensional GloVe embeddings (Pennington et al., 2014). The text embeddings are fed to a convolutional neural network (CNN) to classify relevant nuggets and recursively added to a tree structure with multiple sentences per node. The relevant node is found by maximizing the similarity to nearby word senses. Afterwards, short sentences are selected in a breadth-first manner and feeded to the TextRank based summarizer Summa (Barrios et al., 2016).

H5 labels each word in the corpus with the number of times it appeared in the corpus' selection of relevant sentences. This target data is used for a classifier which learns whether to add neighboring words to the information nugget. The algorithm starts with the highest score.The remaining steps are the same as for H4. The sentences containing the nuggets are organized in a tree structure according to the word sense similarity. To generate a summary from the structure, short sentences are fed to Summa.

LeadFirst This baseline system takes the first sentences of the first document in the (aleatorically ordered) collection until reaching the maximal length of the summary.

MMR and MMR* We use a variant of the maximal marginal relevance (MMR) (Carbonell and Goldstein, 1998) in which the Jaccard word similarity between the candidate sentence at hand and the source documents replaces the comparison to the query. Moreover, the variant *MMR** avoids selecting artifacts such as copyright information for the final summary.

Submodular Lin and Bilmes (2011) found that many well-established methods such as MMR and ROUGE are in fact submodular. Based on these finding, they propose a greedy sentence selection algorithm, which guarantees optimality. In a nutshell, this method selects sentences for a summary which results in a comparable coverage as the original document set. In our implementation, we weight the full coverage with the reciprocal of the total number of sentences and add a simple redundancy check which compares the candidate sentence to the summary. We exclude sentences with less than 30 characters and use the Cosine similarity as coverage function.

PG-MMR This abstractive summarization system uses a pointer-generator encoder-decoder network trained on a single-document summarization dataset as basis for summarizing a set of sentences (Lebanoff et al., 2018). The network is combined with MMR in order to pre-select relevant documents from the document base, which are then passed to the encoder-decoder to generate one sentence. This process is repeated by applying MMR with the summary so far, until the maximum length of the summary is reached. For our experiments we use a pre-trained model which the authors make available.

3.3. Manual Evaluation Experiments

We perform two sets of experiments. The first set aims at determining the quality of the five automatic summarization systems (*H1* to *H5*), described in Section 3.2 below. These summaries are evaluated using both the automatic evaluation method ROUGE and manual evaluation described in Section 3.3.1.

The second set of experiments compares automatic summaries created using a wider range of automatic summarization methods. These summaries are compared pairwise using the 2-Alternative Forced Choice (2-AFC) paradigm (see Section 3.3.1.), which means that users *must* choose between one of two selections – in this case which text performs better on a given quality dimension.

While the first set of experiments involved trained annotators, the second set of experiments was conducted using Amazon Mechanical Turk. A summary of the experiments is given in Table 1.

3.3.1. LikertAnno: Likert-Annotations by Trained Annotators

We extend the criteria used for the manual evaluation in the early DUC series by adding various aspects of text quality, as shown in the following. We basically follow the annotation guidelines of DUC2007[5] with the exception of *information content*. This criterion replaces the more specific *responsiveness*, which assesses the amount of information regarding a specific topic statement, since the broad and heterogeneous document base for each topic does not allow to provide such a specific reference.

Non-Redundancy Specific pieces of information are not repeated either once or several times.

Referential Clarity Referential expressions, such as"he" or "it", are easily resolved, i.e. it is clear from the text what or who the expression refers to.

Grammaticality The text does not contain grammatical errors, such as "He do count to ten."

Focus Does the text stay true to the topic or are various topics or aspects of the topic mentioned.

Structure Is the flow of information nice and easy to follow? Or are various aspects mixed together and their order is unclear.

[5] https://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt, https://www-nlpir.nist.gov/projects/duc/duc2007/responsiveness.assessment.instructions

LikertAnno: Summaries and Likert-annotations	
Number of criteria	11
Number of systems	7 (2 baselines)
Type of annotations	5-point Likert scale
Number of annotators	26
Number of annotations	1274 in total
per summary	4 ($\sim$ 2.67 for baselines)
per annotator	49 (1 per topic)

PairAnno: Summaries and pairwise annotations	
Number of criteria	6
Number of systems	7 (4 hierarchical, 2 extractive, 1 abstractive)
Type of annotations	binary, pairwise
Number of annotators	64
with min. annotations	8 (with 3 annotations each)
Number of annotations	43218 in total
comparisons per topic	$\binom{7}{2} = 21$
per comparison	7
per annotator	$\sim$ 675

Table 1: Key details about DIP-SumEval annotation data sets

Coherence Is the flow of the information logical, the way connections are made (i.e. *But* considering that ...) understandable etc.

Readability Is the text easy to read, in terms of chosen words etc. or is it rather hard to read.

Information Content Denotes the gain of knowledge you obtained by reading the summary, assuming you had little prior information about the topic.

Spelling Are there any spelling errors, as "Count to tne."

Length Is the text complete or does it feel like it was cut somehow, as if something is missing and it should be longer or is it too long and it should be shorter.

Overall Quality General impression of the text.

A total of 26 trained annotators evaluate the automatic summaries created by the systems *H1* to *H5*, MMR and LeadFirst on a 5-point Likert scale (very poor, poor, barely acceptable, good, very good). Each annotator receives one previously unseen summary per topic randomly and in such a way that the summaries of systems *H1* to *H5* are evaluated by 4 different annotators. The baseline systems MMR and LeadFirst are annotated by less annotators, as we were mainly interested in assessing the new hierarchical approaches.In addition to the scores for each of the 11 quality dimensions, annotators also indicate how confident they are in judging a respective quality dimension and which relevance weight they give the respective dimension of the particular summary with respect to the overall quality of the summary.

3.3.2. PairAnno: Pairwise Annotations by the Crowd

In the first phase of our annotation experiments, the goal was to identify better systems among very similar systems. In the second step, we are interested in how these hierarchical summarization systems relate to other summarization systems following different paradigms, such as pure extractive or abstractive summarization. The comparison between the second set of systems is performed using the crowd-sourcing environment Amazon Mechanical Turk. The questions to the annotators are replaced by 2-AFC experiments, i.e., pairwise comparisons between summaries, which is more suitable for a crowd-sourcing environment than judgements on a Likert-scale, as will be discussed in the following.

Foremost, trained annotators or even domain experts are required for reliable annotations on a Likert-scale because of the actual complexity of the task. Indeed, deciding the quality dimension of a set of summaries on an k-scale actually comes down to ordering the summaries into a k-partite ranking. In order to be consistent and coherent in her annotations, the annotator has to consider a large number of relations between summaries when adding a new summary into the ranking. This problem gets aggravated when summaries can only be inspected one by one. Second, as argued, Likert scores are an ordinal assessment, rather than a quantitative measurement, since they mainly serve to put summaries into relation to each other. In consequence, the same Likert score may have a different meaning for different annotators. This may require to perform a calibration of the scores, which is not a trivial task.

Therefore, pairwise comparisons instead of individual, point-wise judgments are often employed in annotation tasks, especially when only untrained annotators are available (e.g. (Gao et al., 2018; Fan et al., 2018)). A pairwise comparison between two objects only requires a binary decision about the preference, which in addition can be performed independently of previous decisions, reducing the cognitive load of the annotators. Kreutzer et al. (2018) found empirically that ordinal and pairwise ratings on a translation task have similar intra- and inter-annotator agreements. Zopf (2018b) even proposes to replace the writing of reference summaries by the annotation of reference pairwise comparisons between sentences. However, Gillick and Liu (2010) also warn about the difficulties of pairwise comparisons, such as unsatisfied transitivity relations, and especially about the caveats of using non-experts such as the increased noise. As argued before, consistency issues also arise for ordinal judgements but are less visible since they appear during the annotation process on the side of the annotators. The increased noise is usually tackled by increasing the number of annotations per annotated object, which is more feasible due to the higher availability of non-experts and lower costs per annotation.

In comparison to the first annotation experiment, the set of quality criteria is reduced to only consider Non-Redundancy, Structure, Referential Clarity, Readability, Information Content and Overall Quality in order to reduce effort and costs. We add automatic summaries created by

Symptoms of depression in children may not be obvious. Treatment of major depression is as effective for children as it is for adults. Untreated depression can have important consequences for the child's well-being. Depression is harmful whether or not a child has a chronic disease. Therapy combined with antidepressants is thought to produce the best outcomes in children with depression. Depression is not just an illness of the mind. Your child's doctor will rule out any other physical causes of your child's symptoms. Talk to your child to see how he or she is feeling.

Non-redundancy 4.0, Structure 3.75, Referential Clarity 4.0, Readability 3.75, Information Content 4.25, Overall Quality 4.25

adolescent depression : depression in children and adolescents. when that " down " mood, along with other symptoms of depression, lasts for more than a couple of weeks. symptoms of depression in children there are several depression symptoms in children to be aware of. depressive disorders during childhood. self-report measures of depression for children and adolescents. not every child who is depressed experiences every symptom. the following factors may be associated with childhood depression : treating depression in children and adolescents antidepressants for children : important information for parents childhood depression.

Non-redundancy 2, Structure 0, Referential Clarity 5, Readability 1, Information Content 1, Overall Quality 1

Figure 1: Example summaries from H2 (left, with average Likert judgments on the bottom) and PG-MMR (right, with votes out of 7 in favor of PG-MMR from direct comparison to H2's summary).

		LikertAnno			PairAnno	
Criteria	Scores	Conf.	K's α	pairw.	K's α	pairw.
Non-Redundancy	4.12	4.12	0.154	0.788	0.017	0.664
Referential Clarity	3.47	4.27	0.327	0.833	0.022	0.670
Grammaticality	4.07	4.25	0.229	0.809	–	
Focus	3.27	4.01	0.247	0.813	–	
Structure	3.00	3.75	0.255	0.814	0.122	0.713
Coherence	3.01	3.70	0.271	0.818	–	
Readability	3.51	4.21	0.233	0.810	0.180	0.735
Information Content	3.25	4.04	0.256	0.818	0.094	0.699
Spelling	4.32	4.16	0.200	0.801	–	
Length	3.70	3.96	0.167	0.791	–	
Overall Quality	3.05	4.11	0.356	0.841	0.195	0.743

Table 2: Average scores and confidence scores for all quality criteria on LikertAnno and inter-annotator agreements for both data sets w.r.t. Krippendorf (K's α) and percentage agreement (pairw.).

the Submodular and the PG-MMR methods, while removing LeadFirst and H5, which perform poorly. All summaries specific to a topic are compared in sets of two, i.e., given the 7 systems each topic leads to 21 pairwise comparisons between summaries. Each pairwise comparison is performed by 7 annotators in order to reduce the influence of noise and bad quality of annotations and annotators. Each comparison is based on one of the quality criteria using the 2-AFC-setup. Annotators are given the following instruction: *Please choose for each pair of summaries whether the left or the right text is better regarding [criterion]* for each pair of presented summaries and for each quality criterion. See also Figure 1 for an example.

4. Analysis of Annotations

In the following we describe our analysis and results of the two experiments with a focus on our research questions.

4.1. LikertAnno Evaluation

Analysis of quality dimensions In the first evaluation trained annotators evaluate the automatic summaries using a 5-point Likert scale.

Table 2 shows the average scores for each quality dimension, as well as the average confidence score. We observe that while the average scores vary between 3.0 and 4.3 the confidence is fairly stable between 3.7 and 4.3, indicating that the annotators are fairly confident in their scoring. The extractive systems receive the highest average scores for Non-Redundancy, Grammaticality and Spelling. Much more difficult criteria, as concerned with semantics, seem to be Structure, Coherence and to some extent also Focus and Information Content.

Table 2 also shows the computed inter-annotator agreements. Interestingly, Overall Quality reveals the highest Krippendorff's α value though the annotators did not agree in the same manner for the other quality criteria, except for Referential Clarity. The subjectivity in assigning and calibrating the values on the Likert scale can become an issue when computing α. Therefore, we also computed the percentage of pairwise agreement. More precisely, for the annotators who coincided in judging a pair of summaries we counted the number a and b of annotators who agreed on each of the two summaries and averaged over the ratios $max(a, b)/(a+b)$. Hence, the agreement scores are at least 0.5. Despite the fact that only a small subset of annotations can be used for computing this statistic, we observe almost the same ranking on the criteria as for α.

Intra-correlation and correlation to automatic evaluation measures Figure 2 shows the correlation between the quality criteria. For each summary and criterion we calculate the Spearman correlation between the Likert score annotations of any given annotator. Additionally, we look into the correlation of the quality criteria to various automatic evaluation methods such as PEAK (cf. Section 2), ROUGE , Jensen-Shannon divergence (JSD) and Kullback–Leibler divergence (KL) (see, e.g., (Louis and Nenkova, 2013)) in Figure 3. More specifically, in the absence of reference summaries, we use the document set for each topic as our reference text.

We find that especially Coherence and Structure are correlated to each other. This is reasonable, as a well-structured text is in theory also more coherent. The Overall Quality is correlated to Information Content, Readability, Structure, Coherence and Focus. Grammaticality, Spelling and Readability highly correlate with each other as well, which is also reasonable, as texts that suffer many grammatical and/or spelling errors are harder to understand.

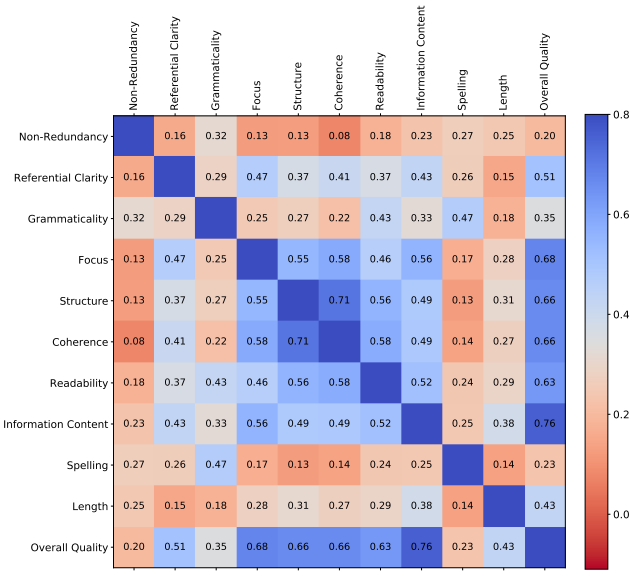

Figure 2: Heatmap of Spearman correlation between quality dimensions on LikertAnno.

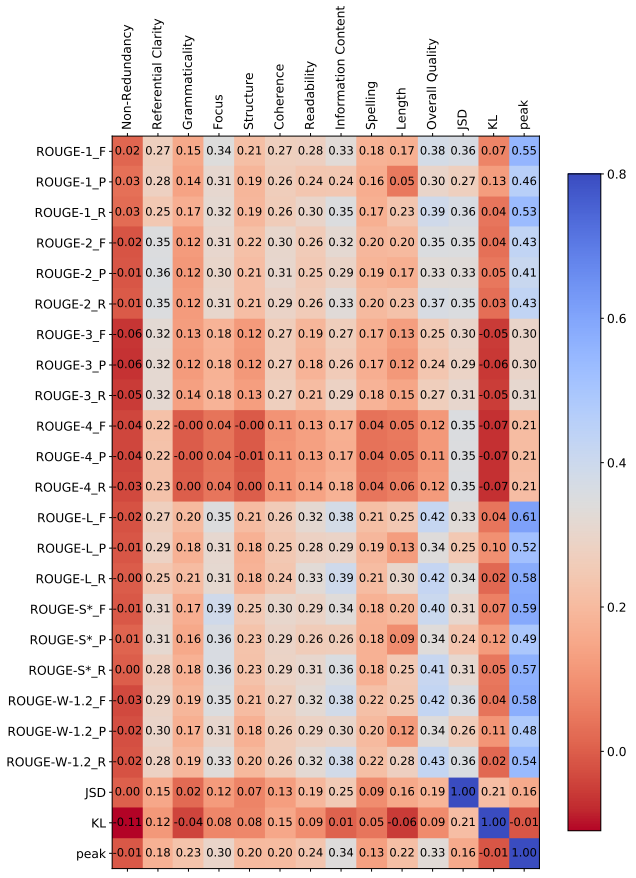

Figure 3: Heatmap of Spearman correlation between quality scores and automatic summary evaluation methods on LikertAnno.

Spelling, Length and Non-Redundancy are not correlated to any of the other quality criteria. This is also reflected in the correlations to ROUGE (see Figure 3), which does not correlate to Non-Redundancy. As ROUGE only observes n-grams, this is easily justified, as redundant information, as well as non-redundant information is nevertheless counted towards the final count of n-grams. Additionally, we see that ROUGE also correlates with Overall Quality, Information Content, Focus and Referential Clarity. This answers the first of our research questions, indicating that ROUGE can represent a range of different quality criteria, even when used without reference summaries.

PEAK, an implementation to automatically perform the Pyramid evaluation, is also correlated to Overall Quality, Information Content and Focus. Therefore, the results suggest that PEAK and ROUGE can be exchanged for one another. However, ROUGE also covers Referential Clarity, which is an important feature of a well-formed and understandable summary. In this respect, ROUGE is superior to PEAK.[6] JSD and KL both perform significantly worse than ROUGE in most quality criteria. Similar to ROUGE they both correlate the least with Non-Redundancy.

Looking at individual ROUGE scores and quality dimensions, we observe that ROUGE-L-Recall and ROUGE-W-Recall correlate strongly with Overall Quality and Focus. For Referential Clarity ROUGE-2 in general correlates strongly. Grammaticality, Structure, Spelling and Length are hardly represented by ROUGE or any of the other evaluation metrics. Coherence is somewhat correlated to ROUGE-2-Precision, while Readability is slightly represented in ROUGE-L-F.

4.2. PairAnno Evaluation

To analyze PairAnno we calculate the correlation differently since for each summary comparison we have 7 votes regarding which summary is considered better with respect to a specific quality dimension. We compute rankings for the systems for each topic, based on the number of times they were picked by the annotator. For each dimension we determine the Spearman rank correlation. The inter-annotator agreements are much lower for the PairAnno annotations (Table 2), but as the agreement percentage values indicate they are also considerably higher than chance (i.e. $4/7 = 0.57$). Again, the highest agreement is on the Overall Quality, but apparently the crowd-workers had a worse common understanding of Referential Clarity than the trained annotators. Instead, Readability was compared often equally between pairs of summaries.

Analysis of quality dimensions Figure 4 shows the correlation between various quality criteria as observed in the crowdsourcing annotation. It shows the Spearman correlation between the scores achieved by each of the $49 \cdot 7$ summaries on two quality criteria. The score corresponds to the number of favourable annotator preferences within each topic. We see that Overall Quality correlates well with Structure, Readability and to some extent to Information Content, which is similar to the annotation using trained annotators. Also similarly, Non-Redundancy does not exhibit strong correlations to any of the other quality criteria. Referential Clarity shows a more mixed picture. While it is somewhat correlated to Overall Quality for the first annotation, it does not show strong correlations with any other

[6] Recent re-implementations of an automatic PYRAMID method have been presented in (Gao et al., 2019b). However, testing this was beyond the scope of this paper up to this point.

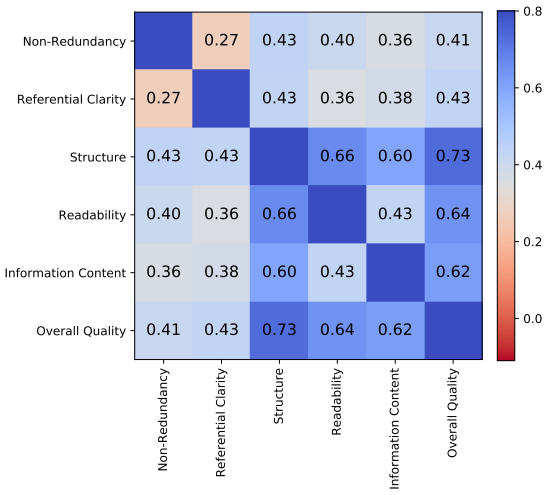

Figure 4: Heatmap of Spearman correlations between quality dimensions for the pairwise comparisons.

quality criteria in the second annotation phase. Additionally, while Referential Clarity is weakly represented in the ROUGE scores, especially ROUGE-2, Non-Redundancy is not represented in any of the automatic evaluation methods.

4.3. Comparison of Systems

Figure 5 shows the analysis of the systems in relation to the respective text quality dimensions we evaluate in both setups. The most notable observation is that the abstractive summarization method performs poorly across all dimensions. The second observation is that, on a high level, the rankings of the systems seem to correlate quite well between the trained and untrained annotators for most dimensions. For instance, they coincide in the first ranked system, or at least in the top two systems (e.g. Referential Clarity). The agreement on the best systems is rather low for Structure and Readability. Perhaps it is possible for a summary's structure to be too simple or too complicated, as opposed to information content. So the summarization system has to find a balance between short sentences, which might sound unsatisfying, and longer unnecessarily complicated sentences missing proper context. With respect to Overall Quality H2 achieves the best results in both evaluations, indicating that this method indeed produces good quality summaries. The system also achieves the best position for Information Content. In terms of Readability H2 achieves the best results in the first evaluation, while H4 achieves the best result in the second evaluation phase. The result is reversed for Referential Clarity. We observe similar results favourable to H2 for most quality criteria, except for Non-Redundancy, where H1 achieves the best result. One explanation could be that H1 integrates a mechanism which requires all extracted sentences to have a certain minimum cosine distance to each other, which explains its first position regarding this dimension. This mechanism is similar to the one used by established MMR* and Submodular, which rank after H1. A reason for this could be that H1 employs a set of post-processing rules, which improve Non-Redundancy, but harms Structure and Information Content. On Structure H3 achieves the best result on the first evaluation, while H2 receives the highest scores in the second evaluation. H4 and H2 also receive a higher score in the second evaluation. One reason could be that H3 uses sentence embeddings to select sentences with the highest mean semantic similarity, which are made up of frequently used sentence structures, i.e., simpler sentence structures. As the systems perform quite differently on the different quality criteria, it might be an interesting future research direction to choose and combine different summarization system in order to improve the desired criteria.

5. Discussion

The analysis in the first part of our experiments indicates that while ROUGE correlates well with manual quality evaluation with respect to Overall Quality, various ROUGE variants also represent other text quality dimensions, which have so far not been analysed in detail. Note, however, that these results have to be taken with care, since as mentioned in Section 4.1 we use the documents to be summarized as reference summaries. Our experiments also show that while some quality criteria correlate strongly with each other, some do neither correlate to other criteria, nor are they well represented by any of the automatic evaluation methods we used. This is especially true for Non-Redundancy, which would need to be covered otherwise. Our experiments also showed that some automatic summarization methods produce higher quality summaries with respect to selected text quality dimensions than others. We conclude that there is not only a need to replace ROUGE as a standard measure for the overall quality of a summary. Rather, the diversity and complex interdependencies between the different dimensions of quality calls for a fine-grained evaluation of summary quality. Obviously, comparing quality dimensions and systems using human annotators, like in this work, is only feasible on a small scale. One way to tackle the task without handcrafting an automatic evaluation measure for each of the criteria is to train evaluation functions (see Section 2).

With this work, we contributed a new data set which can be used to train such evaluation systems for a wide range of quality dimensions, especially compared to previous annotated data sets which consider at most three criteria (Chaganty et al., 2018). Moreover, DIP-SumEval provides two different types of judgements (Likert-scale and pairwise comparisons) which might be suitable for training different systems. For instance, (Kreutzer et al., 2018; Gao et al., 2019b; Böhm et al., 2019; Zopf, 2018b; Zopf et al., 2016a) explicitly require or allow pairwise comparison signals. Pairwise feedback is also especially relevant for personalized summarization (P.V.S and Meyer, 2017) which assumes that judgements vary according to the needs and preferences of the user. Our corpus provides an identification of annotators across annotations for this particular purpose.

Interestingly, a recently proposed abstractive method performed worst in our experiments, suggesting that despite the advances in the field, extractive techniques still represent the state of the art since this techniques can ensure a certain level of quality regarding individual sentences.

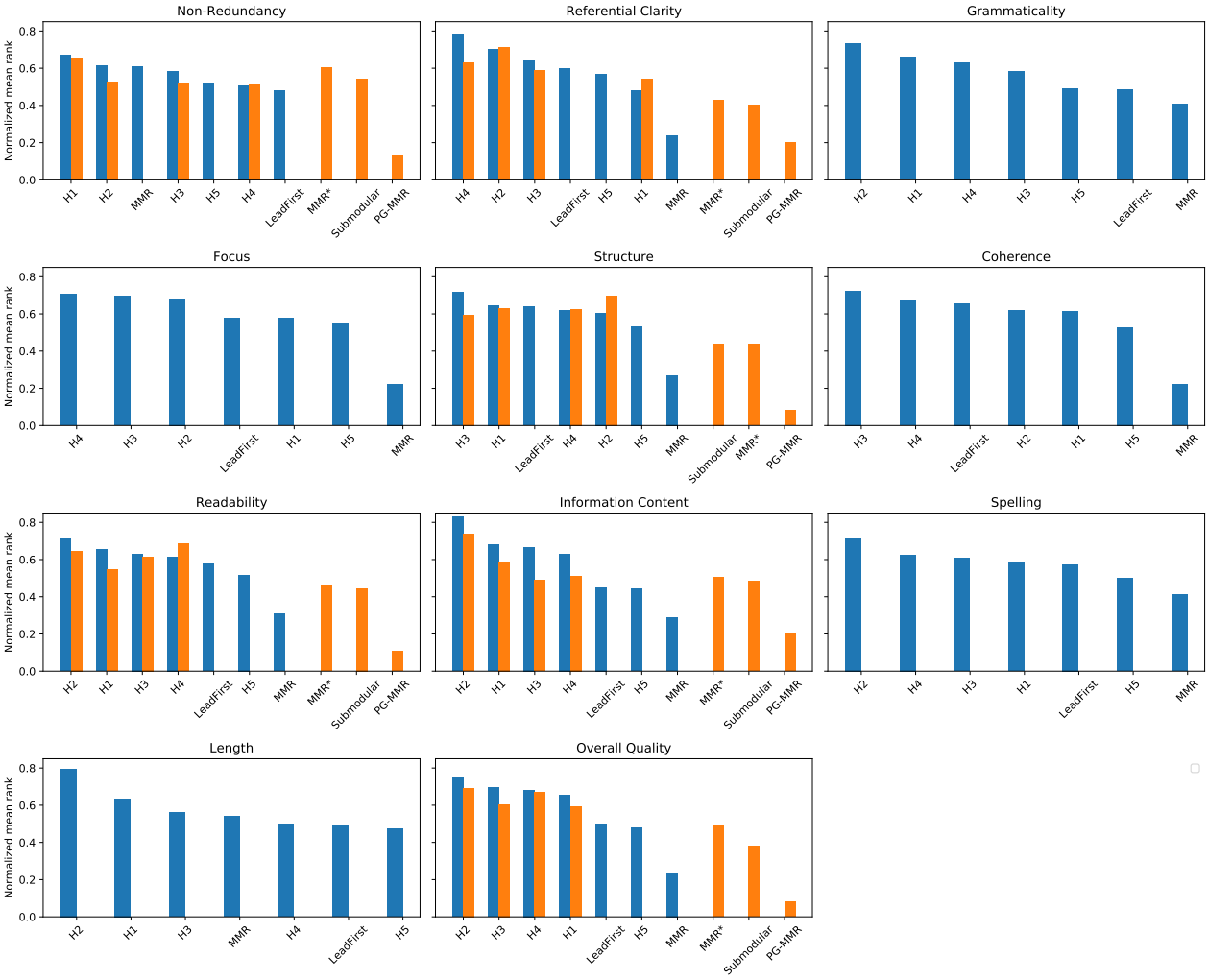

Figure 5: Comparison of systems: The normalized mean rank (the highest rank equals to 1) over the different systems and topics according to Likert and pairwise comparisons for the various quality dimensions. Systems are ordered according to best LikertAnno ranks.

Comparing hierarchical and extractive approaches, we can observe an advantage for the hierarchical approaches. Especially system H2 shows good results on a wide range of quality criteria. This indicates that the combination of various strategies as employed in H2 are beneficial to create high quality summaries on a range of criteria. The mechanisms for redundancy avoidance in the extractive approaches work fine, as the comparison demonstrates. On the other hand, the better control of the diversity of covered information of the hierarchical approaches seems to pay off in our scenario of multi-document summarization.

6. Conclusions

In this work, we presented two annotation experiments on the evaluation of automatically generated summaries on a multi-document summarization task. The analysis revealed that there are complex relationships between the different summary quality criteria. In particular, we added evidence to the insight that automatic evaluation measures such as ROUGE and PYRAMID/PEAK are not sufficient to cover all relevant aspects of a summary's quality. Therefore, we advocate for the use of a variety of automatic evaluation mechanisms covering the different quality dimensions, in particular the use of trainable evaluation functions. The resulting data set DIP-SumEval, which is freely available, provides an excellent resource for training and testing such evaluation mechanisms since it covers a large variety of different quality criteria.

We also used the human annotations from trained as well as non-trained annotators in order to explore the potential of hierarchical automatic summarization. Our analysis suggests that the hierarchical organization is beneficial for trading-off diversity and focus of the covered information especially in a multi-document environment. On the other hand, well-established extractive approaches are well suited to avoid redundancies. A recently introduced abstractive summarization system clearly suffered from the added difficulty of self-composing the texts.

Potential next steps include an analysis of individual document sets. We observed that the quality scores per topic varied considerably. An analysis into why specific document collections showed lower quality scores or annotators indicated a lower confidence in judging them might provide insight into summarization and/or evaluation difficulty.

Acknowledgements

We would like to thank Thomas Arnold, Markus Zopf and M. Braei, A. Filighera, S. Gutsch, F. Helfenstein, M. Höhn, J. Hoppe, J. Jäger, M. Juschak, F. Metzler, V. Pfanschilling, L. Raymann, S. Reynolds, M. Rothermel, F. Scharf, R. Scheffler, S. Seipp, S. Stenger, S. Feudjo, S. Thiem, N. Thoma, J. Vatter, L. Vogel, H. Wieland for their contributions to this work. This work has been supported by the German Research Foundation (DFG) as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) under grant No. GRK 1994/1.

References

Barrios, F., López, F., Argerich, L., and Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Böhm, F., Gao, Y., Meyer, C. M., Shapira, O., Dagan, I., and Gurevych, I. (2019). Better rewards yield better summaries: Learning to summarise without references. In *The 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pages 3101–3111.

Carbonell, J. G. and Goldstein, J. (1998). The use of mmr and diversity-based reranking for reodering documents and producing summaries. In *Proceedings of the 21st meeting of International ACM SIGIR Conference*, volume 335, page 336.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chaganty, A., Mussmann, S., and Liang, P. (2018). The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653.

Fan, A., Grangier, D., and Auli, M. (2018). Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54.

Gao, Y., Meyer, C. M., and Gurevych, I. (2018). APRIL: Interactively learning to summarise by combining active preference learning and reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4120–4130.

Gao, Y., Meyer, C. M., Mesgar, M., and Gurevych, I. (2019a). Reward learning for efficient reinforcement learning in extractive document summarisation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pages 2350–2356.

Gao, Y., Sun, C., and Passonneau, R. J. (2019b). Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418.

Gillick, D. and Liu, Y. (2010). Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151. Association for Computational Linguistics.

Graham, Y. (2015). Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137.

Habernal, I., Sukhareva, M., Raiber, F., Shtok, A., Kurland, O., Ronen, H., Bar-Ilan, J., and Gurevych, I. (2016). New collection announcement: Focused retrieval over the web. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 701–704.

Kreutzer, J., Uyheng, J., and Riezler, S. (2018). Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, July.

Lebanoff, L., Song, K., and Liu, F. (2018). Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141.

Li, J., Li, S., Wang, X., Tian, Y., and Chang, B. (2012). Update summarization using a multi-level hierarchical Dirichlet process model. In *Proceedings of COLING 2012*, pages 1603–1618.

Lin, H. and Bilmes, J. (2011). A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* Edmonton, May-June 2003, pages 71–78.

Lin, H. and Ng, V. (2019). Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9815–9822.

Liu, F., Liu, Y., and Weng, F. (2011). Why is "SXSW" trending? exploring multiple text sources for twitter topic summarization. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 66–75.

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., and Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*. arXiv:1801.10198.

Louis, A. and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*. arXiv:1301.3781.

Nenkova, A. and Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technolgy Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May, 2004, pages 145–152.

Okumura, M., Fukusima, T., and Nanba, H. (2003). Text summarization challenge 2 - text summarization evaluation at NTCIR workshop 3. In *Proceedings of the HLT-NAACL 2003 Text Summarization Workshop*, pages 49–56.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543.

Peyrard, M. and Gurevych, I. (2018). Objective function learning to match human judgements for optimization-based summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 654–660.

Peyrard, M., Botschen, T., and Gurevych, I. (2017). Learning to score system summaries for better content selection evaluation.

In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84.

P.V.S, A. and Meyer, C. M. (2017). Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1353–1363.

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celibi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., and Zhang, Z. (2004). MEAD – a platform for multi-document multilingual text summarization. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26-28 May, 2004.

Sjöbergh, J. (2007). Older versions of the ROUGEeval summarization evaluation system were easier to fool. *Information Processing Management.*, 43(6):1500–1505.

Tauchmann, C., Arnold, T., Hanselowski, A., Meyer, C. M., and Mieskes, M. (2018). Beyond generic summarization: A multi-faceted hierarchical summarization corpus of large heterogeneous data. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, May 2018, pages 3184–3191.

Yang, Q., Passonneau, R. J., and de Melo, G. (2016). Peak: Pyramid evaluation via automated knowledge extraction. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI 2016)*, pages 2673–2679.

Zopf, M., Loza Mencía, E., and Fürnkranz, J. (2016a). Beyond centrality and structural features: Learning information importance for text summarization. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 84–94. Association for Computational Linguistics, August.

Zopf, M., Peyrard, M., and Eckle-Kohler, J. (2016b). The next step for multi-document summarization: A heterogeneous multi-genre corpus built with a novel construction approach. In *Proceedings of COLING 2016*, pages 1535–1545.

Zopf, M. (2018a). auto-hmds: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, May 2018, pages 3228–3233.

Zopf, M. (2018b). Estimating summary quality with pairwise preferences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 1687–1696.

Zopf, M. (2019). *Towards Context-free Information Importance Estimation*. Ph.D. thesis, Technische Universität Darmstadt. Section 7.3.