The Next Step for Multi-Document Summarization: A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach

Markus Zopf, Maxime Peyrard and Judith Eckle-Kohler
Research Training Group AIPHES / Knowledge Engineering Group / UKP Lab
Department of Computer Science, Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt, Germany
www.aiphes.tu-darmstadt.de, www.ukp.tu-darmstadt.de

Abstract

Research in multi-document summarization has focused on newswire corpora since the early beginnings. However, the newswire genre provides genre-specific features such as sentence position which are easy to exploit in summarization systems. Such easy to exploit genre-specific features are available in other genres as well. We therefore present the new ***h***MDS corpus for multi-document summarization, which contains heterogeneous source documents from multiple text genres, as well as summaries with different lengths. For the construction of the corpus, we developed a novel construction approach which is suited to build large and heterogeneous summarization corpora with little effort. The method reverses the usual process of writing summaries for given source documents: it combines already available summaries with appropriate source documents. In a detailed analysis, we show that our new corpus is significantly different from the homogeneous corpora commonly used, and that it is heterogeneous along several dimensions. Our experimental evaluation using well-known state-of-the-art summarization systems shows that our corpus poses new challenges in the field of multi-document summarization. Last but not least, we make our corpus publicly available to the research community at the corpus web page https://github.com/AIPHES/hMDS.

1 Introduction

Multi-document summarization (MDS) is the task of creating a summary from a topically related document collection. Existing corpora for the evaluation of MDS systems, most notably from the Document Understanding Conference (DUC) (Over et al., 2007) and from the Text Analysis Conference (TAC),[1] cover mostly MDS of news documents (Nenkova et al., 2011). While research on MDS has also considered genres other than newswire (e.g., opinionated blog posts in TAC 2008 or biomedical research papers in TAC 2014), MDS has almost exclusively focused on *homogeneous* document collections that belong to the same genre.

However, this homogeneous nature of the existing MDS benchmark corpora does not reflect application scenarios where topically related documents from different genres need to be summarized. An exemplary scenario of increasing importance is MDS on the web, where a user retrieves multiple online documents for a particular topic (cf. Rosner and Camilleri (2008), Nenkova and McKeown (2011)). These online documents may comprise news articles, blog posts, encyclopedic texts, or even scientific articles.

A related issue is the very high effort needed to create a new MDS corpus. Summarizing a set of documents is a demanding task for humans and requires expert abstractors, e.g., information analysts as in the DUC competitions. Attempts to obtain human-written summaries via crowdsourcing have failed (Lloret et al., 2013). When the set of documents to be summarized grows, the summarization task even becomes infeasible for humans.

We address these gaps and present (i) a large heterogeneous MDS corpus in English as a new challenging benchmark for summarization systems, and (ii) a novel approach for constructing such a corpus

[1] http://www.nist.gov/tac

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1535–1545, Osaka, Japan, December 11-17 2016.

at a large scale. Our methodology combines summaries from Wikipedia featured articles with human-retrieved source documents and can thus be generally applied in all languages where Wikipedias exist. Our detailed analysis and evaluation of the new MDS corpus shows that MDS from heterogeneous genres poses a new challenge and calls for future research.

2 Related Work

Related to our work are methods to create MDS corpora, i.e., pairs of topically related documents and reference summaries. In this section, we focus on the particular aspect of obtaining summaries of a given set of documents. We summarize the traditional approaches used in DUC and TAC, and previous work on alternative approaches.

Approach in DUC and TAC How have multi-document summaries been written in DUC and TAC? For example, the 2005 DUC Summary-Writing Task states that a human should first read the topic and all the 25 - 50 documents in the topic cluster. While reading the documents, information relevant for the topic should be highlighted and then used in a second step to write a 250-word summary of the documents. As observed in early DUC competitions, the agreement between different reference summaries for the task of generic summarization was low (cf. Harman and Over (2004)). Therefore, DUC 2005 offered 10 reference summaries for a subset of the 50 topics in order to cover a more representative sample of the diverging reference summaries.

Another approach to reduce the high variation between reference summaries was taken in TAC 2010, where the summarization task was made more specific and accompanied by guidelines describing a list of required aspects to be covered, e.g., summaries in the category accidents should cover aspects such as what, when, where why (Owczarzak and Dang, 2011).

Alternative Approaches Having trained humans, e.g., professional abstractors, write multidocument summaries is expensive, and for laymen, the task is highly demanding and time-consuming. Therefore, a recent project on the construction of a MDS corpus for European Portuguese took a semi-automatic approach where the documents within a topic were first filtered for relevance using a summarization system and then processed by human annotators who only had to remove sentences until the summary length was reached (Almeida et al., 2014). However, this approach is problematic in our context, because existing summarization systems have not been developed and evaluated on heterogeneous text types.

Crowdsourcing is another approach to reduce the effort for creating annotated data and it has been increasingly used in NLP for a range of different tasks. However, for multidocument summaries, it has been shown to lead to poor results (Lloret et al., 2013).

Exploiting user generated content on the web as a source of reference summaries is another option. For example, Aker and Gaizauskas (2010) used content available on the VirtualTourist platform to construct reference summaries for the task of summarizing location-related images. Even more appealing are collaboratively edited platforms such as Wikipedia where the content is refined and consolidated over time through a well-documented discussion and revision process. Recently, Wikipedia featured articles[2] have been used to create an evaluation dataset for the task of multilingual single-document summarization (Giannakopoulos et al., 2015). In Wikipedia featured articles, the first paragraph constitutes a summary of the article; hence Giannakopoulos et al. (2015) used it as a reference summary for the body of the Wikipedia article.

While Giannakopoulos et al. (2015) address the issue in DUC and TAC datasets of topic bias towards news-specific content, they consider only single-document summarization and sources from a single text type (i.e., encyclopedic text). In contrast to the DUC and TAC datasets, their multilingual SDS dataset provides only a single reference summary for each source document. Although the use of featured articles ensures a certain quality level of the summaries, there might be important aspects of the topic which are missing, in particular for topics that are discussed differently in different languages and cultures. For example, Filatova (2009) suggests to combine the Wikipedia articles across multiple languages for a specific topic in order to obtain a comprehensive summary.

[2] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

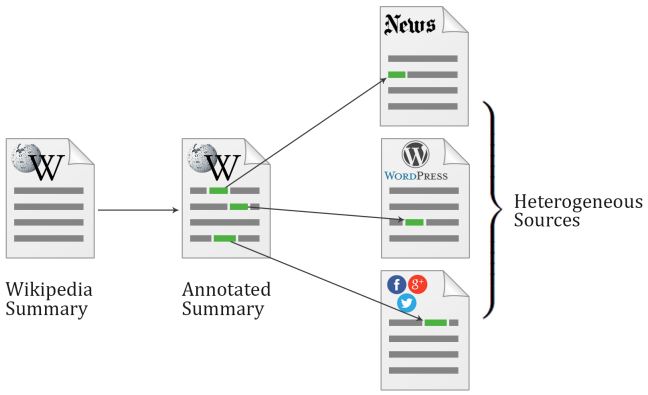

Figure 1: Illustration of the novel corpus construction approach. Left: already available Wikipedia summary; middle: Wikipedia lead with annotated information nuggets; right: a set of heterogeneous source documents which contain the information nuggets.

Our new corpus construction method addresses the bottleneck of human summary creation and the limitation of Giannakopoulos et al. (2015) to be restricted to the single-document summarization task.

3 Reversing Corpus Construction

Previous approaches for constructing MDS corpora start with specifying topics and collecting topic-related source documents; in a second step humans write a summary of the source documents for each of the topics. This approach has several disadvantages, in particular, the reading effort (a human has to read all source documents), the writing effort (the summary has to be written), and the subjectivity of the resulting summary (since the summary is written by only one human). Furthermore, good topics and suitable source documents that cover the topics have to be found in the first place, which can be a laborious and expensive task.

To address these issues, we propose to build MDS corpora by reversing the usual process: Instead of writing a summary for previously gathered source documents, we propose to use an already available high-quality summary and just search for suitable source documents which contain information about the topic. This simplifies the corpus construction process and reduces the effort to create pairs of source documents and summaries. The process is illustrated in Figure 1. In the rest of this section, we describe our novel methodology in detail.

3.1 Methodology

We propose to use the first section of Wikipedia featured articles (the so called *lead*) for this purpose. According to the Wikipedia featured article criteria,[3] these articles are (i) well-written, (ii) comprehensive, (iii) well-researched, (iv) neutral, and (v) stable. In particular, the lead of a featured article is supposed to summarize the topic (according to the guidelines). Since Wikipedia articles are written by many authors, we can consider the lead of a featured article as the consensus of many people regarding the important information about a particular topic. We can therefore consider the leads as high-quality summaries which are representative, because they combine contributions from many authors.

It is important to emphasize that our approach can easily be transferred to other languages where the respective Wikipedias contain featured articles, or even to completely different sources of already existing high-quality summaries.

Given a particular topic (i.e., Wikipedia article), the corpus construction process consists of two steps. First, we annotate topic-specific information nuggets in the summary, in order to be able to retrieve source documents in a systematic way. The information nuggets are supposed to represent atomic information units, i.e., omitting words would change the meaning of the nugget significantly. In the second

[3] https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

step, we search for heterogeneous source documents given the information nuggets. For each annotated information nugget, we search for one source document, which contains the information represented by the nugget, and we keep the text genre as document metadata (the set of possible text genres has been defined in advance, see section 3.2 for details). During our search for source documents, we try to find sources from as many text genres as possible, since we aim to build a heterogeneous corpus. Furthermore, we collect only source documents which are not too similar to the summary. This means that the sources should (i) not contain all information nuggets about the topic, and (ii) contain information which is irrelevant to the topic.

The first property enforces a *content distribution* across all source documents, which means that a summarization system has to consider all source documents to create a proper summary. This is different from previous corpora consisting of documents from a single text type (e.g., newswire), where each individual source document already provided most of the relevant information. In particular, this property reduces the redundancy in the source documents which is often used by multi-document summarization systems to detect important information. Consider as an example homogeneous newswire corpora where documents on a given topic typically have a high degree of redundancy due to the way journalists write news articles: in a typical hard news report, the reported news is first summarized in the lead section, and then specified and elaborated on in the body of the article (Thomson et al., 2008), which creates redundancy even in a single article. As a further consequence of the low redundancy in our collected source documents, the heuristics that the most frequent information is also the most important information (making frequency usually a strong feature in newswire MDS corpora (Nenkova et al., 2006)) is weakened.

The second property, the *presence of unimportant information*, is also different from prior corpora, where the sentences provide usually at least some information about the topic. Selecting the right sentences becomes therefore even more important in order to achieve a good scoring.

Our corpus construction approach is valid, since it is reasonable to assume that the retrieved documents do not introduce new important information about the topic. Wikipedia featured articles already contain the most important information related to a topic, and are furthermore continuously updated by the Wikipedia community. We therefore retrieve the version timestamp of each article and store it in the corpus metadata; the manual retrieval of sources for this version of the Wikipedia article was performed within a short time frame.

3.2 Corpus Construction

We applied our new corpus construction methodology to build ***h***MDS, a new, heterogeneous, multi-genre corpus for MDS. In this section, we describe details of the actual corpus construction process.

Since Wikipedia provides a large number of featured articles, we first selected a subset of the articles for our project. Based on the featured article overview page,[4] we selected the three broad domains

- *Art, Architecture, and Archaeology* (D1),
- *History* (D2), and
- *Law, Politics, and Government* (D3).

We asked three annotators to perform the steps described in the previous section. As part of the first step, they should tag and extract roughly 10 to 20 information nuggets from the lead of each Wikipedia article. For the topic *California Gold Rush*, examples of extracted nuggets are *1848-1855*, *period in American history*, *Sutter's mill*, *gold was found by James W. Marshall*, and *300,000 gold-seekers*.

In the second step, the annotators searched for source documents as described above, using well-known web search engines. Since they tried to find the nugget text verbatim in the source documents, ROUGE scoring is assumed to perform well in our corpus, which is confirmed in our experiments in section 5. The retrieved documents were archived in the Wayback Machine.[5] We provide more detailed

[4] `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles`
[5] `http://archive.org/web/`

corpus information as well as the annotator guidelines and download scripts to retrieve the documents at the corpus web page `https://github.com/AIPHES/hMDS`. For the retrieval step, we used a predefined set of 10 text genres, which are shown in Table 1 along with short descriptions.

Since the source documents are web pages and not ready-to-use raw text documents, we asked the annotators to extract and store the relevant part of the web page, which leads to rather compact source documents. We also generate two larger versions of the corpus by performing boilerplate content removal with Boilerpipe (Kohlschütter et al., 2010). A second version contains all visible web page content. We use the shorthand notation ***h***MDS-M, ***h***MDS-A, and ***h***MDS-V to denote the manually extracted, the automatically extracted and the version with all the visible content, respectively. We also provide a version of the corpus where sentence splitting has already been applied, because most extractive summarizing systems extract sentences. This improves the reproducibility of summarization experiments since it removes one noisy preprocessing step. We use version 1.7.0 of the Stanford segmenter in the DKPro Core software (Eckart de Castilho and Gurevych, 2014) to produce the sentence segmentation.

Although the main purpose of our corpus is MDS, it can also be used for various other tasks. Since we store the genre of each source document, it can be used for text genre classification, or as dataset for training and evaluating boilerplate removal systems. Research in automatic source document retrieval can also use our new corpus.

4 Corpus Analysis

In the following, we summarize and analyze the result of our corpus construction effort. In particular, we analyze whether our corpus actually is as heterogeneous as we designed it to be via our corpus construction methodology. For this, we also compared the ***h***MDS corpus with two common MDS corpora, namely DUC2004[6] and TAC2008A.[7]

The ***h***MDS corpus contains 91 summary-source documents pairs (topics). In total, the annotators retrieved 1,265 source documents. We obtained 13.90 ± 3.09 (where $\pm$ indicates the standard deviation) source documents per topic in comparison to DUC2004 and TAC2008A which both have exactly 10 source documents per topic (i.e. 10 ± 0).

4.1 Text Genres

Since we asked the annotators to classify each source document according to the text genre it belongs to, we are able to provide a detailed analysis of the distribution of texts genres in our corpus. Table 1 provides an overview of the text genres which are present in our corpus, as well as their distributions.

		across domains			
Text Genre – Description (Example)	Count	D1	D2	D3	Avg. length (in words)
article – well-written text (high-quality blog post, news article)	524	0.37	0.42	0.47	1452.70 ± 1751.28
forum post – lack text structure (QA site, Youtube comment)	115	0.10	0.08	0.09	964.10 ± 1726.89
microblog – short, contains abbreviations (Twitter)	33	0.03	0.03	0.02	53.61 ± 14.44
organization – announcement, press release (any org./company)	99	0.11	0.06	0.06	749.29 ± 1119.21
encyclopedic short – encyc. source (Urban Dictionary, IMDB)	115	0.12	0.07	0.08	400.45 ± 362.88
encyclopedic long – encyc. source (Wikipedia)	137	0.11	0.14	0.07	3434.15 ± 5077.32
social media – post in social network (Facebook, Google+)	11	0.01	0.01	0.01	270.45 ± 250.67
scientific – contain citations and bibliography	119	0.07	0.08	0.14	5394.03 ± 9118.11
education – text book, tutorial	79	0.05	0.09	0.05	1568.76 ± 3020.62
dialogues – opinionated (interview, transcript, discussion)	33	0.02	0.03	0.03	3759.79 ± 4897.97
Total	1265	0.36	0.35	0.28	1863.59 ± 3928.91

Table 1: List of genres present in the ***h***MDS corpus along with their fractions in the different domains. The length details are computed for the M-version of our corpus.

We obtained a large amount of source documents for the "article" genre. The distribution of the other genres is rather uniform with most documents belonging to the encyclopedic, scientific, and forum post categories. The fraction of microblog documents, dialogues, and social media is considerably smaller.

[6] `http://duc.nist.gov/duc2004`
[7] `https://tac.nist.gov/2008`

On average, we obtained 5.39 $\pm$ 1.54 different genres per topic with a minimum of 3 and a maximum of 9 different genres per topic. These results substantiate our claim that the ***h***MDS corpus contains sources from very diverse genres. We also observe variations of the distributions of text genres across the three different domains.

4.2 Length of Source Documents and Summaries

Another property which we analyze is the length of both source documents and summaries. Table 1 provides information about the distribution of lengths across the different genres. We see that we obtained a wide variety of different lengths across the genres. Since Table 1 only provides information for ***h***MDS-M, we provide more details of the source document lengths in Table 2 where we can see that the variation of lengths increases strongly in the versions A and V of the ***h***MDS corpus. Compared to DUC2004 and TAC2008A, we obtained much longer source documents, as well as a much higher variance in length.

Corpus	Avg. length (in words)	Relative std
hMDS-M	1863.59 $\pm$ 3928.91	2.11
hMDS-A	2192.53 $\pm$ 8196.75	3.74
hMDS-V	2973.06 $\pm$ 8429.32	2.84
DUC2004	672.14 $\pm$ 506.32	0.75
TAC2008A	589.20 $\pm$ 480.33	0.82

Table 2: Length comparisons of source documents.

Corpus	Avg. length (in words)	Relative std
hMDS	245.55 $\pm$ 132.94	0.54
DUC2004	118.11 $\pm$ 6.38	0.05
TAC2008A	109.33 $\pm$ 7.01	0.06

Table 3: Length comparisons of summaries.

Regarding the summaries, we achieved a large difference to prior work as well, see Table 3. Our summaries are on average about twice as long as the summaries in DUC2004 and TAC2008A. The major difference, however, is the huge variance of lengths in our corpus which can be observed by both standard- and relative standard deviation. The minimum length of a summary in our corpus equals 72 words and the maximum length equals 657 words.

4.3 Textual Heterogeneity

The heterogeneity of our corpus also results from other textual properties. Heterogeneous documents are expected to use different wording and to have some topics shifts. In order to measure this textual heterogeneity, we use information theoretic metrics on word probability distributions.

In our experiments, we use the Jensen-Shannon (JS) divergence, which is a symmetric and always defined version of the Kullback Leibler (KL) divergence (Kullback and Leibler, 1951). It incorporates the idea that the distance between two distributions cannot be very different from the average of distances from their mean distribution. Its expression is $JS(P\|Q) = \frac{1}{2}KL(P\|A)+\frac{1}{2}KL(Q\|A)$ where $A = \frac{P+Q}{2}$ is the mean distribution of P and Q. Based on the JS divergence, we can define a measure of textual heterogeneity TH for a topic T composed of documents $d_1, \cdots, d_n$ as

$$TH_{JS}(T) = \frac{1}{n}\sum_{d_i \in T} JS(P_{d_i}, P_{T\setminus d_i}) \quad (1)$$

where P_{d_i} is the probability distribution of words in document d_i and $P_{T\setminus d_i}$ is the probability distribution of words in all other documents of the topic except d_i. TH_{JS} is the average divergence of documents with all the others and provides therefore a measure of diversity among documents of a given topic.

	DUC2004	TAC2008A	***h***MDS-M	***h***MDS-A	***h***MDS-V
TH_{JS}	0.3019	0.3188	0.3815	0.3358	0.3252

Table 4: Average TH_{JS} scores of classical corpora and our new datasets.

Table 4 shows the results of the analysis according to the TH metric. DUC2004 and TAC2008A have similar source documents in comparison to ***h***MDS-M according to TH. ***h***MDS-A and ***h***MDS-V are closer to the classical datasets than ***h***MDS-M. This is due to the fact that these versions contain much more boilerplate content compared to ***h***MDS-M, which makes them more similar again. Nevertheless, the heterogeneity of the main content in the source documents is verified by TH.

4.4 Distribution of Content

	ROUGE-1 Recall				ROUGE-2 Recall			
Corpus	full	n-1	n/2	1	full	n-1	n/2	1
hMDS-M	0.68	0.67 ± 0.03	0.63 ± 0.06	0.42 ± 0.12	0.48	0.46 ± 0.04	0.40 ± 0.07	0.17 ± 0.09
DUC2004	0.43	0.43 ± 0.01	0.43 ± 0.02	0.36 ± 0.05	0.16	0.15 ± 0.01	0.15 ± 0.01	0.09 ± 0.03
TAC2008A	0.46	0.46 ± 0.02	0.44 ± 0.02	0.35 ± 0.05	0.20	0.19 ± 0.01	0.17 ± 0.02	0.10 ± 0.03

Table 5: Results of the content distribution experiment. We present results of using all, n-1, n/2, and only 1 source document. The omitted documents were selected randomly.

As mentioned in section 3.1, our corpus construction aims for a distribution of content across all source documents. To evaluate this property, we conduct an experiment in which we use different fractions of the source documents as input for a greedy optimal summarization systems. Since we use a greedy optimal summarizer due to performance reasons, the optimal results differ slightly from the results in Table 6. We observe in Table 5 that the omission of one document has already an effect in our corpus, whereas the optimal performance stays nearly constant in DUC2004 and TAC2008A. The effect increases when only half of the source documents is considered. A rather large difference can be observed when only one source document is available to the summarizer. In DUC2004 and TAC2008A, the summarizer is still able to achieve 83.7% and 76.1% of the optimal score, whereas in our corpus, only 61.8% can be achieved according to ROUGE-1. For ROUGE-2, we observe 56.3% and 50.0% in DUC2004 and TAC2008A, compared to 35.4% in our corpus.

As we did not manually annotate the information nuggets in the source documents (due to the high annotation effort), we can investigate the differences in content distribution only regarding the ROUGE scores, which also considers high-frequent function words in the default setup. This might explain why the variation is not even higher.

4.5 Distribution of ROUGE Scores

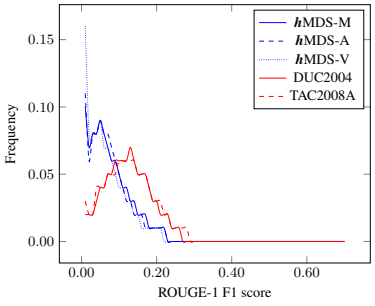

Figure 2: F1 score

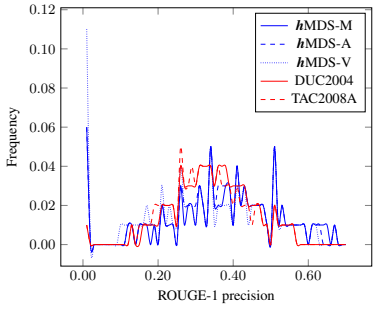

Figure 3: Precision

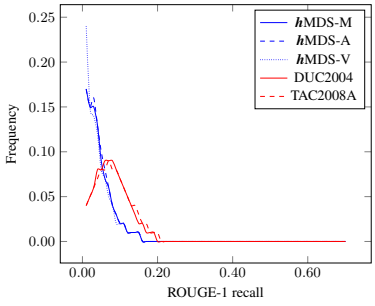

Figure 4: Recall

In this section, we investigate the distribution of ROUGE-1 scores of single sentences in our corpus compared to the DUC2004 and TAC2008A corpora. Figures 2, 3, and 4 provide the distribution of ROUGE-1 F1 measure, ROUGE-1 precision, and ROUGE-1 recall, respectively. The evaluation shows that the distribution according to ROUGE-1 precision is not much different, except for a large number of sentences with very low precision in our corpora. The ROUGE-1 recall curve shows that single sentences in DUC2004 and TAC2008A on average provide a higher recall compared to ***h***MDS. In combination, we see that there are a lot of sentences in ***h***MDS with both very low precision and very low recall. Thus, we can conclude that we indeed constructed a corpus containing sentences which do not contribute much to a good summary (see section 3.1, *presence of unimportant information*).

5 Summarization Experiments

In this section, we conduct experiments with well-known baselines and summarization systems to analyze our new corpus further. As evaluation metric, we apply the commonly used ROUGE scoring (Lin,

2004). All experiments were evaluated with ROUGE version 1.5.5 with standard parameters `-a -m -n 2 -x -c 95 -r 1000 -f A -p 0.5 -t 0 -l` l, which includes stemming without removing stopwords. Since our summaries have different length, we use a variable length parameter l for the evaluation, where l denotes the length of the reference summary. For the DUC2004 and TAC2008A datasets, we use $l = 100$. All results are averaged across all topics.

5.1 Summarization Systems

First, we describe several well-known extractive summarization approaches. By applying them to both the classical datasets and to our new corpus, we can test whether or not we succeeded in creating a new challenge for this research area. Furthermore, we can investigate different properties of our corpus in more detail, such as the strength of the sentence position and centrality features which are used by different summarizers.

Measure	Dataset	Optimal	Random	Lead	LexRank	ICSI	LSA	TF-IDF
ROUGE-1	DUC 2004	0.4535	0.2955	0.3424	0.3450	0.3778	0.2904	0.3318
	TAC 2008A	0.5067	0.2963	0.3315	0.3466	0.3675	0.3154	0.3236
	hMDS-M	0.6992	0.3754	0.4069	0.4192	0.5401	0.3447	0.3671
	hMDS-A	0.6962	0.3242	0.1041	0.4083	0.5370	0.3391	0.3439
	hMDS-V	0.7019	0.2847	0.0050	0.3133	0.5033	0.3228	0.3302
ROUGE-2	DUC 2004	0.1876	0.0435	0.0766	0.0715	0.0900	0.0430	0.0657
	TAC 2008A	0.2540	0.0458	0.0765	0.0773	0.1107	0.0696	0.0572
	hMDS-M	0.4960	0.0732	0.1237	0.1273	0.2293	0.0689	0.0939
	hMDS-A	0.4845	0.0594	0.0318	0.1192	0.2267	0.0652	0.0805
	hMDS-V	0.5018	0.0450	0.0018	0.0797	0.2082	0.0603	0.0766

Table 6: Performance according to ROUGE recall of various summarization approaches for DUC 2004 and our new MDS corpus.

Optimal provides an upper bound for the performance of the summarization systems by searching for the best combination of sentences that achieves the highest ROUGE score. It is no competitive summarization system, since it uses oracle knowledge (i.e. the reference summaries) to generate the best possible summary with an ILP solver. **Random** selects sentences randomly. Although it is a quite simple approach, it helps to compare the absolute ROUGE scores in different datasets. **Lead** is a simple, but quite strong baseline in newswire documents. It iteratively selects the first sentences of the source documents until the desired summary length is reached. Its strength in classical datasets derives from the fact that most important information are usually written first in news articles.

TF-IDF was introduced by Luhn (1958). It uses the typical word frequency assumption as a proxy for importance, which is a very strong feature in multi-document newswire corpora (Nenkova et al., 2006). In **LexRank** (Erkan and Radev, 2004), a similarity graph is constructed with sentences as nodes and the Cosine similarity between them as edge weights. Sentences are scored according to their PageRank score. LexRank relies on centrality to measure relevance. **LSA** performs single value decomposition on a terms-sentences matrix weighted by TF-IDF scores. The summary is made up of the sentences that represent well the most important topics. LSA is a topic-model approach which relies on frequency of co-occuring patterns. **ICSI** (Gillick and Favre, 2009) treats summarization as global linear optimization problem. It extracts a summary by solving a maximum coverage problem considering the most important bi-grams. The importance of bi-grams is estimated via their frequency in the source documents.

In our experiments, we use our own implementations of Optimal, Random, and Lead and implementations of TF-IDF, LexRank, and LSA provided by the sumy package.[8] We use the Python implementation of ICSI released by Boudin et al. (2015).

5.2 Results

The results of the summarization experiment are displayed in Table 6. We first observe that the optimal achievable score is much higher in our corpus, which means that a better fit of source documents and summary is possible. This might be due to the fact that we asked the annotators to search for documents

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

containing the information nuggets verbatim. Having only one reference summary compared to multiple reference summaries in DUC2004 and TAC2008A might also have an impact. Since we observe different optimal scores, we present in the following relative ROUGE-1 scores according to the respective optimum. For ***h***MDS-V and ***h***MDS-A, we use the optimal score of ***h***MDS-M, which is a lower bound for the true score.

Selecting sentences randomly yields lower results in ***h***MDS compared to DUC2004 and TAC2008A in terms of the maximal possible result (e.g. in DUC2004, selecting sentences randomly yields 65.2% of the optimal score, in ***h***MDS-M we only achieve 54.7%). The lead baseline, which is strong in newswire, does not perform well in ***h***MDS-M (58.2%) and particularly bad in ***h***MDS-A (14.9%) and ***h***MDS-V (0.7%). Sentence position is therefore no longer a good feature in these corpora. ROUGE-2 scores support this result even stronger.

TF-IDF, LSA, LexRank, and ICSI use different approaches to model centrality. We observe that the performance of all approaches decreases when more noise is added to the corpus (versions A and V compared to M). ICSI, a strong state-of-the-art approach (Hong et al., 2014), performs best in our new corpus. We observe, similarly to Random, that it is better suited to summarize DUC2004 (83.3%) than ***h***MDS-M (77.2%), ***h***MDS-A (76.8%), and ***h***MDS-V (72.0%). LSA and TF-IDF have the lowest performance across the four non-baseline systems, and the LexRank results are between them and ICSI.

Our results show that the relative difference to the optimum is quite large in our corpora compared to the classical datasets.

6 Conclusion

In this paper, we presented ***h***MDS: a new, heterogeneous, multi-genre MDS corpus which provides new challenges for summarization systems and is therefore suited to drive research in new directions. To build the corpus, we proposed a novel corpus construction approach which reverses the classically applied approach and reduces the construction effort.

We provide detailed analyses of ***h***MDS which verify that our corpus is inherently different from DUC2004 and TAC2008A in various ways. ***h***MDS contains topics from three broad domains, the reference summaries and source documents have varying lengths, and the source documents belong to different genres, which results in the important information being distributed across all source documents. Thus, a system has to be able to deal with this high degree of heterogeneity in order to generate a proper summary. Furthermore, summaries in the corpus are not only written by one person, but are the consensus of a whole community and therefore provide a representative view of the importance of information. We provide results of several baselines and well-known summarization systems which indicate that our corpus poses new challenges for summarization systems. Last but not least, we make our corpus publicly available to the research community.

In future work, we are going to add another dimension of heterogeneity to the corpus by adding pairs of summary and source documents in German, based on German Wikipedia featured articles. Furthermore, we want to investigate if we can generate a very large corpus based on the proposed construction approach by automatically retrieving source documents instead of manually collecting them. The corpus built manually would then serve as a gold standard.

Acknowledgments

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, and via the German-Israeli Project Cooperation (DIP, grant No. GU 798/17-1). We thank Kamel Chelly, Matthias Hanreich, and Hannah Wieland for their contributions to the corpus construction.

References

Ahmet Aker and Robert Gaizauskas. 2010. Model Summaries for Location-related Images. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3119–3124, Valletta, Malta. European Language Resources Association (ELRA).

Miguel B. Almeida, Mariana S. C. Almeida, André F. T. Martins, Helena Figueira, Pedro Mendes, and Cláudia Pinto. 2014. Priberam Compressive Summarization Corpus: A New Multi-Document Summarization Corpus for European Portuguese. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 146–152, Reykjavik, Iceland. European Language Resources Association (ELRA).

Florian Boudin, Hugo Mougard, and Benoît Favre. 2015. Concept-based Summarization using Integer Linear Programming: From Concept Pruning to Multiple Optimal Solutions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1914–1918, Lisbon, Portugal. The Association for Computational Linguistics.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

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

Elena Filatova. 2009. Multilingual Wikipedia, Summarization, and Information Trustworthiness. In *Proceedings of the SIGIR 2009 Workshop Information Access in a Multilingual World*, pages 19–24, Boston, Massachusetts USA.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual Summarization of Single and Multi-Documents, On-line Fora, and Call-center Conversations. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 270–274, Prague, Czech Republic. Association for Computational Linguistics.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

Donna Harman and Paul Over. 2004. The Effects of Human Variation in DUC Summarization Evaluation. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 10–17, Barcelona, Spain.

Kai Hong, John Conroy, benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 441–450, New York City, NY USA.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of ACL workshop on Text Summarization Branches Out*, pages 74–81, Barcelona, Spain.

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

Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2:159–165.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 573–580, Seattle, Washington, USA. ACM.

Ani Nenkova, Julia Hirchberg, and Yang Liu, editors. 2011. *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*. Association for Computational Linguistics, Portland, Oregon.

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

Karolina Owczarzak and Hoa Dang. 2011. Who wrote What Where: Analyzing the content of human and automatic summaries. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 25–32, Portland, Oregon. Association for Computational Linguistics.

Mike Rosner and Carl Camilleri. 2008. MultiSum: Query-Based Multi-Document Summarization. In *Proceedings of the workshop Multi-source Multilingual Information Extraction and Summarization at Coling 2008*, pages 25–32, Manchester, UK.

Elizabeth A. Thomson, Peter R. White, and Philip Kitley. 2008. "Objectivity" and "hard news" reporting across cultures: comparing the news report in English, French, Japanese and Indonesian journalism. *Journalism Studies*, 9(2):212–228.