Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6700–6708
Marseille, 11–16 May 2020

Summarization Beyond News: The Automatically Acquired Fandom Corpora

Benjamin Hättasch, Nadja Geisler, Christian M. Meyer, Carsten Binnig
TU Darmstadt
{benjamin.haettasch, nadja.geisler, carsten.binnig}@cs.tu-darmstadt.de

Abstract

Large state-of-the-art corpora for training neural networks to create abstractive summaries are mostly limited to the news genre, as it is expensive to acquire human-written summaries for other types of text at a large scale. In this paper, we present a novel automatic corpus construction approach to tackle this issue as well as three new large open-licensed summarization corpora based on our approach that can be used for training abstractive summarization models. Our constructed corpora contain fictional narratives, descriptive texts, and summaries about movies, television, and book series from different domains. All sources use a creative commons (CC) license, hence we can provide the corpora for download. In addition, we also provide a ready-to-use framework that implements our automatic construction approach to create custom corpora with desired parameters like the length of the target summary and the number of source documents from which to create the summary. The main idea behind our automatic construction approach is to use existing large text collections (e.g., thematic wikis) and automatically classify whether the texts can be used as (query-focused) multi-document summaries and align them with potential source texts. As a final contribution, we show the usefulness of our automatic construction approach by running state-of-the-art summarizers on the corpora and through a manual evaluation with human annotators.

Keywords: corpus construction, multi-document summarization, query-focused summarization

1. Introduction

Motivation: Abstractive summaries help users to understand new text collections efficiently but writing these summaries is time-consuming and complex. Automatic summarization aims to eliminate or reduce the manual process. Since the advent of deep learning, automatic summarization methods require huge corpora to properly train neural architectures. The CNN/DailyMail dataset (Hermann et al., 2015), Gigaword (Napoles et al., 2012), and the New York Times (NYT) corpus (Paulus et al., 2018) are currently the largest summarization corpora. They have been used successfully for training a wide range of neural architectures, including recurrent neural networks (Nallapati et al., 2017), pointer-generator networks (See et al., 2017), attention mechanisms (Paulus et al., 2018; Gehrmann et al., 2018), and approaches based on reinforcement learning (Narayan et al., 2018; Gao et al., 2018).

All of these corpora are limited to the news genre where texts are typically too short to qualify as general-purpose summaries. For example, CNN/DailyMail provides only bullet-point summaries, Gigaword contains headlines as the summary of an article's first sentence, and the NYT corpus pairs news articles with their abstracts. To break new ground in the automatic summarization of other genres, we require new corpora that can cover other text genres and summary types on the one side but are large enough to train neural networks on the other side. However, constructing summarization corpora is still a manual task today and thus requires excessive resources which limits the variety of available corpora significantly.

Contributions: In this paper, we propose a novel approach to automatic construction of large summarization corpora. The main idea behind our approach encompasses the use of existing large text collections (e.g., thematic wikis) and automatically classifying whether the texts can be used as (query-focused) multi-document summaries as well as aligning them with potential source texts.

As an important first step for developing such an automatic construction approach, we use the Fandom wikis (formerly known as wikia). `Fandom.com` is a community page dedicated to providing a place for enthusiasts to discuss and share knowledge about their favourite entertainment content. It currently consists of more than 385,000 communities on different franchises (movies, television series, games, books, and more) with over 50 million pages in total. The sizes of the different communities range from only a few pages to well over 100,000 content pages. Most of those wikis use an open *Creative Commons Attribution Share-Alike license* allowing us to use and redistribute their articles.

The Fandom wikis often contain articles describing the same topic in multiple levels of detail—there are articles giving a general overview of a character, event or place as well as articles focusing on a single aspect of it (e.g., a relationship, scene or time) in detail. Those articles normally reference each other through links. Our main idea is to automatically identify such overview articles or sections that qualify as a summary and align them with the potential source documents (i.e., the detailed articles) if the supposed alignment quality is high enough.

We show that it is possible to generate multiple different corpora with user-defined properties using this idea. For example, it is possible to vary the target length of the summaries, but also the difficulty of the summarization task which we control by the ratio between the sizes of summary and the source documents. Finally, we also allow users to choose whether the contents of a constructed corpus should be retrieved from a single community or whether a more general corpus is constructed from multiple communities at once.

To summarize, in this paper we make the following contributions: (1) We present a framework that can be used to create new summarization corpora and discuss reasonable choices for the parameters. (2) We provide three

new sample corpora created with our automatic construction pipeline. (3) We provide a comprehensive evaluation based on these corpora, using traditional and neural network based methods to validate that our pipeline is able to automatically create corpora of use for state-of-the-art summarizers. (4) We make our code available as open source under the MIT License. It can be found along with the data and a user documentation at `https://datamanagementlab.github.io/fandomCorpus`.

Outline: We first give an overview of our automatic corpus construction pipeline and report important properties of our approach. Then we show the results of our evaluation before discussing potential future work and wrapping up our contribution.

2. Automatic Corpus Construction

In this Section, we describe the steps of our automatic approach to create topic-specific multi-document summarization corpora. The essential stages of our approach are: (1) parsing and cleaning of input documents, (2) selecting potential candidates for abstractive summaries from those input documents and assigning summary candidates to them, and (3) choosing the final set of abstractive summaries based upon a newly developed quality threshold and splitting the selected summaries into training, validation, and test set if needed. An overview can be found in Figure 1.

As mentioned before, in this paper we use the Fandom wikis as an example set of source documents, but we believe that our approach can be easily extended to other sources: while step (1) is source specific and has to be implemented for each new set of sources, steps (2) and (3), which are the core of our automatic construction approach, are implemented in a general way.

2.1. Overview of the Pipeline

Parsing and Cleaning the Sources: The first step of our pipeline encompasses parsing the sources and cleaning the data for the automatic corpus construction.

As already mentioned, we use the the Fandom wikis as a document source in this paper. Database dumps can be downloaded from the "Special:Statistics" page of each Fandom community. Each dump consists of a large xml-file containing all contents of that wiki. In addition to the articles, this covers metadata on media files, discussion pages, category overviews, special pages and other sites not relevant to our task. For example, the English Star Wars wiki[1] contains about 150,000 content pages but over half a million pages in total. Hence, all non-article pages have to be discarded. This can be done by specifying an article namespace to use or by using rules to ignore certain page title prefixes (e.g., "Help:" or "Talk:").

Afterwards, the contents are preprocessed: We split the pages into sections including their respective titles, extract the links between pages and convert the content into plain text. This includes removing link texts, tables, templates and other kinds of wiki markup.

[1]Wookieepedia, `https://starwars.fandom.com`

Min. summary length [words]	150
Max. summary length [words]	400
Extractive summary length [words]	250
Target-source-ratio	2
Min. source doc count	5
Min. bigram overlap	50%

Table 1: Parameter settings for corpus creation used for the sample corpora of this paper.

Finding Summary Candidates: The identification of summary candidates is the most crucial step for creating high-quality corpora automatically. At a high-level, a corpus that is useful for abstractive summarization should group a set of documents with at least one possible summary of these documents. In addition, many of the automatic summarization approaches take a "query" as input that represents the information needs of the user i.e., describes what aspects are important.

Hence, in this step, we aim to select triples (i.e., a set of source documents, a summary, and a query) that represent good candidates from a given cleaned data dump for the final corpus. For both the *source documents* and *summaries* our pipeline uses sections of the wiki articles since they are coherent and self-contained portions of text. As a *query* describing the section, we combine the title of an article with the section title, e.g., "Luke Skywalker: Birth".

To identify sections that qualify as possible summary candidate triples we use the following heuristics: (1) Only sections with a length between certain threshold values are considered as summaries. These thresholds can be adapted based on the task at hand. The default values for all parameters used for the sample corpora in this paper can be found in Table 1. (2) We discard summary candidates having only few linked documents (i.e., potential source documents). Again, the number of source documents is a parameter that can be set by the user. Higher values increase the difficulty of the summarization task since the summary content has to be extracted from more input documents, but may also drastically decrease the number of candidates overall. (3) After applying these purely statistical heuristics, we compute the content alignment between summary and source documents as the overlap between sources and summary candidates. The required minimal overlap, too, is a parameter that can be set by the user for creating a corpus; the lower the value, the more candidate summaries and source documents will be selected but the difficulty increases. In this paper, we use the number of shared bigrams to approximate the similarity. The quantity of overlap shows how much the summary and source texts contain similar concepts but it can only be a first hint as to whether the information in the sources is sufficient to re-create the abstractive summary given a particular user-query. Therefore, in addition to the overlap, we create extractive summaries from the selected candidate sources based on the abstractive summary. An automatically calculated quality score for the extractive summary is used to select the set of summaries and source documents to form the final corpus.

In addition to the user-tuneable parameters of the fully-

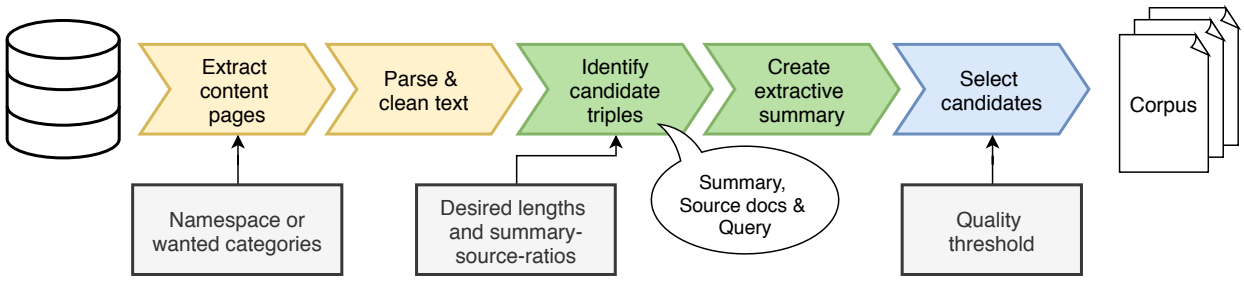

Figure 1: Fandom Corpus Construction Pipeline

automatic process, users can also specify preferences as to which contents are particularly relevant. The Fandom wikis, for example, use a category system, as most wikis do. As a default, articles from all categories are extracted, but it is possible to restrict the categories, e.g., to discard all articles about non-fictional characters (i.e. actors, directors, film crew, ...) from a corpus about a movie franchise.

Selecting Summaries for the Corpus: The heuristics mentioned above help to identify possible candidates for triples consisting of a summary, a list of source documents, and a query—however, their quality can vary significantly as we show in our evaluation in Section 4. For some of them, the summary is indeed a high-quality summary of the extracted documents complying with the query, while for others it is hardly possible to find the information of the summary in the source documents. Hence, in a final step, we need to identify the usefulness of each triple and select only those which exceed a predefined quality threshold.

The selected summaries can optionally be split into training, validation, and test set. The split sizes—like all other parameters of the pipeline—can be adapted by the user of the framework according to their needs.

2.2. Extractive Summaries for Final Selection

Building an extractive summary involves choosing the best subset of sentences from the sources that form a summary of their content. In this paper, the extractive summarization procedure is modelled as an Integer Linear Program (ILP) based on the ideas of Boudin et al. (2015), and Avinesh P. V. S. and Meyer (2017). The main intuition is that the ILP extracts the sentences with the most important concepts from the source documents to form a summary within a maximal length. To model the importance of sentences, we weight concepts according to their frequency in the human-written text (i.e., the selected candidate summary from the Fandom wiki). By doing so, we reward the system for a summary that contains many concepts of the abstractive reference summary. We use bigrams as concepts and ignore those consisting solely of stopwords.

To find good candidate triples, we use the objective score of the ILP for extractive approximation of the summary. This score is high if the extractive summary contains many concepts from the reference summary, hence resembling it well. For the final corpus, we only use summaries with a score higher than a certain threshold. In our evaluation in Section 4.2, we show how different values for this threshold impact the overall corpus quality.

In this paper, we use two different optimization objectives for the ILP. In both formulations, c_i refers to the individual concepts and L to the maximal summary length (which we set according to the selected range of the target length for the abstractive summaries). Moreover, sentences are referred to as s_j with length l_j and Occ_{ij} meaning that concept c_i occurs in that sentence. The first ILP formulation, as shown below, intends to maximize the overall sum of weights for distinct covered concepts, while making sure that the total length of all selected sentences stays below a given threshold and the weight of a concept is only counted if it is part of a selected sentence.

$$\max \sum_i w_i c_i$$
$$\forall j. \sum_j l_j s_j \leq L$$
$$\forall i, j. s_j Occ_{ij} \leq c_i$$
$$\forall i. \sum_j s_j Occ_{ij} \geq c_i$$
$$\forall i. c_i \in \{0, 1\}$$
$$\forall j. s_j \in \{0, 1\}$$

The second objective is simpler and tries to maximize the weight of the distinct selected sentences. Therefore, it is rewarded if an important concept appears in multiple sentences.

$$\max \sum_j s_j \sum_i w_i Occ_{ij}$$
$$\forall j. \sum_j l_j s_j \leq L$$
$$\forall j. s_i \in \{0, 1\}$$

In our experiments, we evaluate both of these ILP formulations with regard to the final corpus quality. Both approaches use only syntactical features and no semantic ones (e.g., embeddings). They do not require time-intensive training and can be computed within a few seconds. Yet, it would be easy to exchange this component of the pipeline, if needed for a certain application.

3. Properties of Our Corpora

In the previous section, we have presented our new approach for automatically constructing summarization corpora. Using this approach, we have created three different

sample corpora (one for Harry Potter, two for Star Wars) using the Fandom wikis as input. In this section, we will now discuss the unique properties of these corpora which differentiate them from other available corpora and, thus, make them a valuable contribution on their own. These sample corpora are all available for download with the sources of our construction pipeline.

First of all, our corpora do not feature news texts with their typical peculiarities (e.g., all important sentences at the beginning) but a mix of encyclopedic and narrative (storytelling) texts. In contrast to other sources, in Fandom wikis there are not a few dozens but thousands of articles about a certain topic. If the corpus is constructed from a single community, all articles are from the same domain (i.e., a closed world). However, it is also possible to utilize the common structure of the different communities and build a corpus containing texts of different domains, e.g., to train more general summarizers.

Additionally, new corpora are fast and cheap to construct with just a minimum of manual work needed. There are many communities with lots of articles (e.g., Star Trek with 47,181 articles, Dr. Who with 71,425 articles) and the wikis are still growing. Moreover, communities are available in many different languages, hence this approach can be used to create corpora for various languages (e.g., one of our sample corpora is in German). The Creative Commons License of the texts allows us to offer the resulting corpora for download instead of only publishing tools for re-creating the corpora. This is in contrast to many existing news-based corpora such as Zopf (2018) which depend on crawling and thus the availability of external resources.

Last but not least, the abstractive texts in our corpora are of high quality since they are written by volunteers with intrinsic motivation and not by poorly paid crowd workers rushing through the task. A sample for such an abstractive text which shows the high-quality can be seen in Figure 4.

4. Analysis & Results

In our analysis, we show the validity of our pipeline and the usefulness of the generated data using three sample corpora created with our approach (two in English, one in German). We start by analyzing the properties of the automatically constructed corpora, then discuss the design decisions and validity of our pipeline steps, and finally run state-of-the-art summarizing systems on the data and evaluate their performance.

4.1. Statistics of Corpora

As a first analysis, we computed several statistics about the three sample corpora we constructed using our pipeline. The goal is to show whether, from a purely statistical perspective, the automatically constructed corpora are similar to manual (human-created) ones.

The results can be found in Table 2. The abstractive summaries have an average length of 260–270 words, the extractive summaries were created using a target length of 250 words, hence they have an average length little below that value. This length is similar to traditional multi-document summarization corpora like the DUC '06 and DUC '07 datasets[2]. It is a lot longer than the average length of 50 words of the live blog corpus (Avinesh P. V. S. et al., 2018) and drastically longer than the headline summaries of multiple news-based corpora such as Gigaword (Napoles et al., 2012).

The average number of source documents per summary lies between 19 and 25 documents. This as well is similar to the DUC '06 and DUC '07 datasets, higher than the ten documents considered in the DUC '04 and TAC '08[3] challenges, and about one half to one fourth of the amount of snippets per summary for the live blog corpora. The average length of the source documents is one to two magnitudes higher than for the live blog corpora, resulting in a higher overall source length to extract the important concepts from. Especially for Harry Potter, the overall length is two to three times higher than for the other two corpora, making this task especially hard.

The size of the final corpus varies depending on the size of the Fandom community and the quality threshold. For our sample corpora, it ranges from 250 topics, which is similar to the DUC '06 dataset used for traditional summarization approaches, to 1,300 topics, which is a size that can be used to train deep learning approaches. Additionally, it is possible to combine topics from multiple communities into a single training corpus.

This has an effect on the domain distribution and topic heterogeneity as well. A corpus constructed from a single community covers topics from only one domain, with the main difference between documents being whether they are about an event, a place, a being or a thing. Mixed corpora may contain texts from totally different domains (e.g., about a movie, a video game and baking recipes). The heterogeneity of writing styles, levels of detail, narrating styles and more, comes from the nature of the wiki itself and is inherently contained in all of the corpora.

In summary, it can be seen that, from a statistical perspective, it is possible to generate corpora with various properties matching typical needs of current (multi-)document summarization tasks.

4.2. Validation of the Pipeline

Our pipeline requires some parameters. Most of them are straightforward and can be adapted directly, according to the task at hand (e.g., the target length of the summaries) or have a direct impact on the difficulty of the dataset (e.g., the range of the amount of source documents or the length ratio between source and target). The most important parameter is the quality threshold (and connected to it the method to generate a score for the *extractability* of the summary from the sources). In this Section, we evaluate how this parameter influences the overall corpus quality.

First, we compare the two extraction modes (i.e., the two different ILPs described in Section 2). Figure 2 shows the correlation between the scores of both methods. It can be seen that the score of the sentence-based method is always equal or higher than the one of the concept-based method on the same data. The reason is that the sentence-based

[2] https://duc.nist.gov/

[3] https://tac.nist.gov/2008/

Corpus	Star Wars (en)	Star Wars (de)	Harry Potter	
Quality Threshold	50	50	50	20
# Articles	148,348	39,356	15,993	15,993
Candidates	5,659	999	1,466	1,466
Selected Summaries (train/valid/test)	882 / 109 / 107	221 / 28 / 28	205 / 23 / 26	1,171 / 146 / 147
Avg. Summary Length	261.14	270.79	270.06	245.47
Avg. # of Source Docs per Summary	24.69	20.15	19.79	17.11
Avg. Source Length per Doc	1,143	855	3,087	3,400
Avg. Overall Source Length per Summary	28,236	17,241	61,111	58,188

Table 2: Properties of the three sample corpora. For each, the amount of textual documents, the amount of candidates for a topic (target summary and matching source documents), and the amount of documents selected by the quality threshold (split into train, validation and test sets) are reported. For the Harry Potter corpus, the sizes of a second variant with a lower quality threshold are listed. For the selected summaries, the average length of the summary (in words) as well as the average size of the input documents (in words) and their average number per summary are reported.

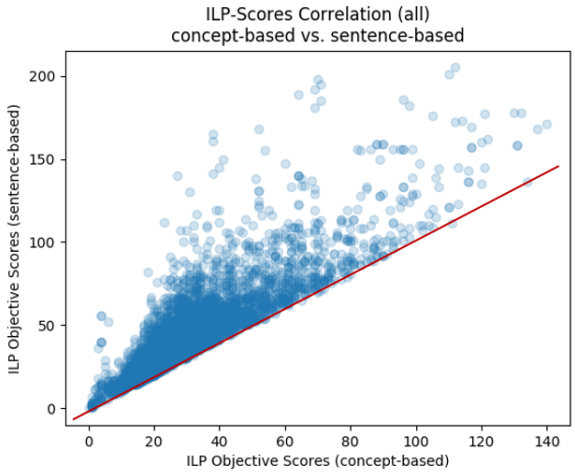

Figure 2: Correlation between the objective scores of concept- and sentence-based extraction method on the same data.

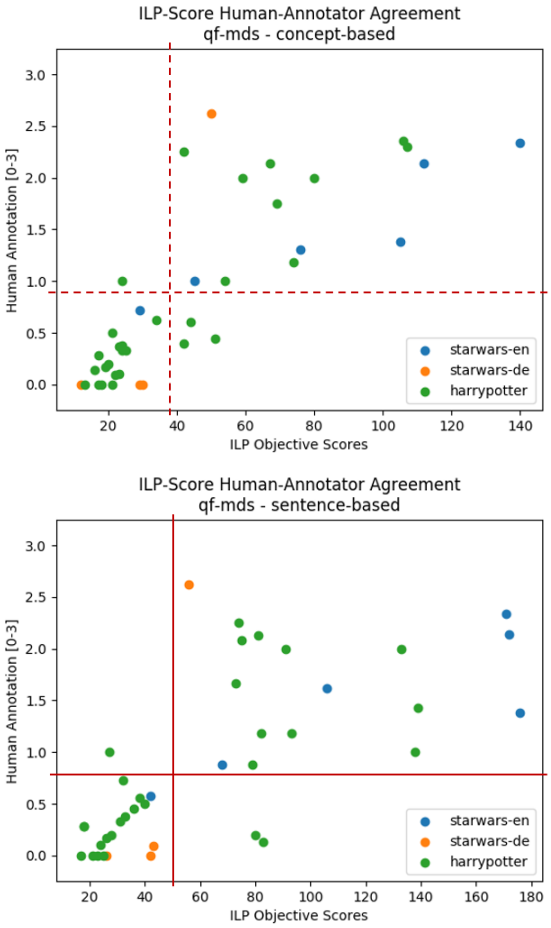

Figure 3: Human agreement with automatically generated extractive summaries for concept-based and sentence-based creation method. Average values for the sentences of 39 annotated documents, possible values between 0 and 3 (best).

method will always get at least the same score for the summary selected by the concept-based method, but the score will be higher as soon as at least one duplicate concept is in it. However, normally both approaches produce similar but nonetheless different summaries, and the quality of a summary produced by the concept-based method might be better than a sentence-based one with a higher score, because for the first one duplicate concepts were not rewarded.

The correlation justifies using either of the two methods as a quality indicator. However, the question how the quality of the summary really correlates to the score of the extraction remains. To assess this, we asked human annotators to evaluate the quality of 39 equally distributed summaries. We asked them to decide for each sentence in the human abstract if it is covered by the extractive summary (0) not at all, (1) partially, (2) mostly, or (3) fully. The human decision is averaged for the full summary and correlated to the score of the extraction. The results can be found in Figure 3. It can be seen that a higher ILP score does indeed correlate with a better human evaluation. Based on this we have chosen the ILP-thresholds for the selection of the summaries. Our experiments suggest a value of about one fifth of the target length ($250/5 = 50$) for the sentence-based method, while for the concept-based method the corresponding threshold would be slightly lower since duplicates are not counted.

In addition we show a sample for a human abstract and the corresponding extractive summary in Figure 4, to in-

Human-written abstract

At Hogwarts School of Witchcraft and Wizardry, sixth years are typically 16 to 17 years of age, although some may be older, if they have had to repeat a year like Marcus Flint. The sixth year is the year in which students advance to N.E.W.T.-level classes. [...] Neither the core classes nor the elective courses are available to any student who does not meet said requirements. While students do have the opportunity to choose whether they wish to continue in particular subjects, those who begin studying N.E.W.T.-level subjects in their sixth year are expected to carry on with the subject into the seventh year, and sit the N.E.W.T. exam in that subject. [...] Students in the sixth year may also elect to take part in Apparition lessons for a fee of twelve Galleons. [...]

Extractive summary

The fifth year is also the year in which students receive career counselling from their Heads of House. Sixth years are typically sixteen to seventeen years of age, although some may be older, if they have had to repeat a year like Marcus Flint did. Sixth years may also elect to take part in Apparition lessons for a fee of twelve Galleons. [...] Depending on the minimum requirements of the professor teaching the subject at that time, students are allowed to sit any number of classes as long as they meet said requirements. While students do have the opportunity to choose whether they wish to continue in particular subjects, those who begin studying N.E.W.T.-level subjects in their sixth year are expected to carry on with the subject into the seventh year and sit the N.E.W.T. exam in that subject. [...]

Figure 4: Sample human abstract and the corresponding auto-generated extractive summary (concept-based) for topic "Sixth year: During the sixth year" (00943) from the Harry Potter corpus. Due to spacing reasons, only excerpts are shown. From 10 sentences in the human abstract, 7 are fully covered by the extractive summary, one sentence mostly and two sentences not at all, leading to a human evaluation score of 2.3 with an optimization score of 107.6 for the extraction ILP.

tuitively demonstrate that extractive summaries are a good stub to judge the quality of source documents for abstractive summarization. While we can see that a good quality of the extractive summary implies that the source documents are useful for abstractive summarization of the given documents, there is still room for improvement: First, some sentences might be missing in the extractive summary simply because the length of the extractive summary is typically lower than the abstractive one (since not the full length could be exploited). A second problem can be found in the Harry Potter wiki, but it is likely that it will frequently occur in other domains as well: in many cases all names of people or places in a summary are linked to articles about them, adding these articles to the source documents. Yet, without co-reference resolution and explicit query handling, the system is prone to selecting sentences about the wrong entity as input. More generally speaking, there is a lack of real understanding of the extracted contents in our construction pipeline. The approach works solely on a syntactic level and does not use any semantic features such as synonyms at the moment. We want to address this as an extension of our pipeline in future work.

4.3. Corpora Quality

In these experiments, we ran multiple well-known techniques that were successfully used for single and multi-document summarization. The goal is to show that the quality of the automatically created corpora is high enough that state-of-the art summarizers can perform reasonably well on those corpora. Our implementations are based upon the implementation by Avinesh P. V. S. et al. (2018). For the assessment of summary quality based upon a reference summary, we compute and report the ROUGE metrics. Owczarzak et al. (2012) show that these metrics strongly correlate with human evaluations of this similarity. We report the ROUGE-1 (R1) and ROUGE-2 (R2) metrics (without stemming or stopword removal) as well as ROUGE-SU4 (SU4) as the best skip-gram matching metric.

Baseline Summarizers: As state-of-the-art summarizers, we use the following systems:

*TF*IDF* (Luhn, 1958): The sentences are scored with the term frequency times the inverse document frequency for all their terms, ranked by this score and greedily extracted.

LexRank (Erkan and Radev, 2004): This well-known graph-based approach constructs a similarity graph $G(V, E)$ for all sentences V with an edge between them if their cosine-similarity is above a certain threshold. The summary is built by applying the PageRank algorithm on this graph and, again, extracting greedily.

LSA (Steinberger and Jezek, 2004): This approach uses singular value decomposition to reduce the dimensions of the term-document matrix to extract the sentences containing the most important latent topics.

KL-Greedy (Haghighi and Vanderwende, 2009): This approach tries to minimize the Kullback-Leibler (KL) divergence between the word distributions of the summary and the source documents.

ICSI (Gillick and Favre, 2009): This approach is based on global linear optimization. It extracts a summary by solving a maximum coverage problem that considers the most frequent bigrams in the source documents. Hong et al. (2014) found this to be among the state-of-the-art systems for multi-document summary.

We applied all of these approaches to all topics of our corpora. Due to large input sizes, LexRank, LSA, KL-Greedy and ICSI did not terminate in a reasonable time on some topics. The affected topics varied for each approach.

In addition, to judge the quality of the baselines, we also computed the upper bound that an extractive summarizer could achieve in the best case. An extractive summarization system normally cannot re-create the human-written abstractive text exactly, since the abstractive sentences differ from the sentences of the source texts that can be extracted. Hence, the best overlap between an abstractive and the best extractive text is usually below 100%. To take this into consideration, we compute and report those upper bounds

Systems	Harry Potter			Star Wars (en)			Star Wars (de)		
	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4
Luhn	0.1669	0.0308	0.1366	0.2440	0.0523	0.2045	0.1725	0.0357	0.1378
LexRank	**0.3702**	**0.0729**	**0.2850**	**0.3845**	**0.1049**	**0.3103**	**0.3579**	**0.0784**	**0.2711**
LSA	0.3113	0.0421	0.2454	0.3135	0.0533	0.2550	0.3081	0.0512	0.2350
KL	0.2407	0.0528	0.1897	0.3087	0.0808	0.2546	0.2213	0.0524	0.1742
ICSI	0.2224	0.0360	0.2041	0.3041	0.0423	0.2507	0.2199	0.0353	0.1984
UB1	**0.5585**	0.1744	0.3802	**0.5793**	0.2341	0.4210	**0.6095**	0.3354	0.4859
UB2	0.5465	**0.2609**	**0.4137**	0.5700	**0.3050**	**0.4491**	0.6089	**0.3847**	**0.5111**

Table 3: Average scores (ROUGE-1, ROUGE-4, ROUGE-SU4) for different baseline systems on all candidates of all three sample corpora. Values between 0 and 1, higher is better.

Systems	Harry Potter			Star Wars (en)			Star Wars (de)		
	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4
Luhn	0.1791	0.0365	0.1475	0.2605	0.0560	0.2195	0.1830	0.0412	0.1491
LexRank	**0.3855**	**0.0881**	**0.3053**	**0.3929**	**0.1083**	**0.3227**	**0.3662**	**0.0849**	**0.2849**
LSA	0.3267	0.0545	0.2635	0.3293	0.0584	0.2722	0.3226	0.0624	0.2541
KL	0.2753	0.0655	0.2176	0.3116	0.0780	0.2609	0.2321	0.0617	0.1902
ICSI	0.2440	0.0419	0.2245	0.3223	0.0496	0.2683	0.2350	0.0412	0.2125
UB1	0.6261	0.2885	0.4742	0.6830	0.4115	0.5656	0.7513	0.5811	0.6815
UB2	**0.6265**	**0.3746**	**0.5122**	**0.6835**	**0.4726**	**0.5939**	**0.7569**	**0.6164**	**0.7027**

Table 4: Average scores (ROUGE-1, ROUGE-4, ROUGE-SU4) for different baseline systems on selected summaries (score of sentence-based extraction ILP greater or equal to the quality threshold of 50) of all three sample corpora. Values between 0 and 1, higher is better.

for extractive systems as suggested by Peyrard and Eckle-Kohler (2016). This is done using the first ILP from Section 2.2 with slightly adapted concepts and weights: we compute one upper bound based on unigrams (UB1) and one upper bound based on bigrams (UB2). For both of them, the concepts are not weighted but the maximum coverage of distinct n-grams is counted. As for the baselines, the ROUGE scores for the created extractive summary compared to the abstractive text are computed.

Neural Summarizers: In addition to the baseline systems mentioned above, we also evaluate the data using learned models. To do so, we use the best scoring model combination for extractive summarization by Kedzie et al. (2018), a combination of a Seq2Seq model as extractor and an Averaging Encoder. Yet, our datasets use a compatible data format, hence all other models evaluated in that paper can be used on our data as well[4]. For training, the extractive summary provides a binary decision for every input sentence, i.e. whether it should be part of the summary or not. For generation, a probability is inferred for every sentence and then used to rank them and extract greedily.

We benchmark all three corpora with both extraction methods and a quality threshold of 50. Additionally, we run the benchmark on the Harry Potter corpus with sentence-based extraction and a threshold of 20, and on a combined dataset (Star Wars, Harry Potter and Star Trek[5], all English). All experiments use 200-dimensional GloVe vectors to represent words.

Analysis of the Summarization Quality: Table 4 shows the benchmark results of the selected summaries for the three sample corpora. We report the ROUGE-1, ROUGE-2 and ROUGE-SU4 scores for the different baseline systems. All experiments use a target length of 250 words, if not stated otherwise. This corresponds to the length of the commonly used DUC '06 and DUC '07 datasets. When compared to the benchmark runs on all candidates of the corpora (see Table 3), one can see that the average scores for all systems are higher on the selected summaries, proving that these are, on average, better pairs of summary and source documents. However, in relation to the upper bounds (UB1 and UB2), even the best performing baseline (LexRank) can only reach one third to one fifth of the upper bound on ROUGE-2 (for ROUGE-1 and SU4 it is at least half or better). This reflects our findings from other papers, e.g., Avinesh P. V. S. and Meyer (2017), and thus we believe that the quality of our automatically constructed corpora is on par with the manually created ones used in previous evaluations. Moreover, the fact that state-of-the-art summarizers can only reach one third to one fifth of the upper bound on ROUGE-2 also emphasizes that multi-document summary is still a challenging task in general and needs further research which we hope to stimulate with this paper.

This is also stressed by the following findings: Table 5 shows the results of training multiple sequence-to-sequence models with the training data from the corpora. We tested them both on the original human abstracts and the extrac-

[4]Kedzie et al. (2018) provide reference implementations with an unified interface for all evaluated models at `https://github.com/kedz/nnsum/`

[5]`https://memory-alpha.fandom.com/`

Corpus	Extraction Mode	Quality Threshold	Validation R2	Test: Human Abstracts R1	Test: Human Abstracts R2	Test: Human Abstracts SU4	Test: Auto-Extractive R1	Test: Auto-Extractive R2	Test: Auto-Extractive SU4
Star Wars (en)	concept	50	0.0876	0.4461	0.1501	0.4148	0.4729	0.1807	0.3369
	sentence	50	0.0864	0.4375	0.1476	0.4042	0.4721	0.2096	0.3310
Harry Potter	concept	50	0.0655	0.3557	0.0714	0.3321	0.3754	0.0876	0.2528
	sentence	50	0.0692	0.3528	0.0699	0.3290	0.3657	0.0802	0.2473
	sentence	20	0.0460	0.3673	0.0690	0.3384	0.3775	0.0888	0.2562
Star Wars (de)	concept	50	0.1127	0.4197	0.1700	0.3984	0.4189	0.1606	0.3066
	sentence	50	0.1294	0.4365	0.1852	0.4146	0.4456	0.2086	0.2957
Combined (en)	concept	50	0.0749	0.4136	0.1151	0.3825	0.4500	0.1501	0.3126
	sentence	50	0.0753	0.3927	0.0947	0.3629	0.4019	0.1079	0.2729

Table 5: Average ROUGE values for Seq2Seq models (neural baseline) trained on the different training sets and tested on the original human abstracts and the auto-generated extractive abstracts of the respective test sets. Rouge values between 0 and 1, higher is better.

tive summaries that were automatically created based on them to see the effect of the abstraction. The scores on the extractive test set are higher for all models, as expected. The test on the human abstracts can be compared to the results of the non-neural baselines from Table 4. We can see that the neural approach outperforms the other baselines on the Star Wars corpora. Especially for the German variant the result is surprisingly good even though we are not using German embeddings but rather standard GloVe vectors. However, for the Harry Potter corpus, the neural baseline cannot even outperform the LexRank baseline. We find three reasons for that: first, the total length of the source documents (which is two to three times higher than for the other two corpora), second the linking style of the wiki (see Section 4.2) and third the comparatively low amount of training data. It can be seen that scores for the model trained on the variant with lower quality threshold (leading to a five times higher corpus size) are similar or even higher. Getting similar results from training data of a worse quality supports the assumption that the amount of training data is a problem here.

For those cases, we test a combined corpus, where texts from multiple domains are combined. We can see that this can be used to handle the lack of training data, but that a specialized model will outperform this more general model when there is enough training data available.

5. Future Work

With this paper, we present ready-to-use data and an approach to generate more. Of course, there is still room for improvement and extensions of the pipeline:

One important point is the generation of extractive summaries. As discussed in Section 4.2, our pipeline does not exploit semantic features yet. The use of semantic word representations, word sense disambiguation, co-reference resolution, or entity linking could create better extractive summaries and serve as a better basis for quality threshold decisions.

A second important point is the length of the source documents to be summarized. Since wiki authors are encouraged to add a lot of hyperlinks between the texts, the list of source documents might contain articles not entirely relevant for the topic, making it very hard to solve the summarization task. Future work should focus on developing methods to choose more relevant subsets of the source texts. Semantic features could play an important role here as well.

Finally, we think that our approach can also work as a basis to generate data for other tasks. One example is hierarchical summarization (Christensen et al., 2014): Fandom communities about television series often contain articles about every single episode, about each season and articles about certain aspects of the full series. These articles all have different levels of detail and form a hierarchy that can be extracted using some simple manual rules.

6. Conclusion

In this paper, we presented a novel automatic corpus construction approach and three open-licensed corpora for multi-document summarization based on this approach. All corpora are available online to be used directly by other researchers, together with a ready-to-use framework to create custom corpora with desired parameters like the length of the target summary and the amount of source documents to create the summary from. We verified the pipeline and showed the usefulness of the corpora, including the fact that state-of-the-art summarizers cannot yet solve all challenges posed by our new corpora. Our data could contribute to the further development of systems for (semi-)automatic multi-document summarization, especially those exploiting the query or relying on user feedback. The framework can be used to generate further training and test data for these systems or serve as basis to generate data for other tasks, such as hierarchical summarization.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group *Adaptive Preparation of Information from Heterogeneous Sources* (AIPHES) under grant No. GRK 1994/1. Thanks to Aurel Kilian and Ben Kohr who helped with the implementation of the first prototype and to all human annotators.

7. Bibliographical References

Boudin, F., Mougard, H., and Favre, B. (2015). Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMLNP)*, pages 1914–1918, Lisbon, Portugal.

Christensen, J., Soderland, S., Bansal, G., and Mausam. (2014). Hierarchical summarization: Scaling up multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 902–912, Baltimore, Maryland, June. Association for Computational Linguistics.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

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, Brussels, Belgium.

Gehrmann, S., Deng, Y., and Rush, A. (2018). Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4098–4109, Brussels, Belgium.

Gillick, D. and Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics.

Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1684–1692, Montréal, Canada.

Hong, K., Conroy, J. M., Favre, B., Kulesza, A., Lin, H., and Nenkova, A. (2014). A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*, pages 1608–1616.

Kedzie, C., McKeown, K., and Daumé III, H. (2018). Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium, October-November. Association for Computational Linguistics.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Avinesh P. V. S. and Meyer, C. M. (2017). Joint optimization of user-desired content in multi-document summaries by learning from user feedback. In *ACL*, pages 1353–1363. ACL.

Avinesh P. V. S., Peyrard, M., and Meyer, C. M. (2018). Live blog corpus for summarization. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 3197–3203, Miyazaki, Japan, May.

Nallapati, R., Zhai, F., and Zhou, B. (2017). SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *Proceedings of the Thirty-First Conference on Artificial Intelligence (AAAI)*, pages 3075–3081, San Francisco, CA, USA.

Napoles, C., Gormley, M., and Van Durme, B. (2012). Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC/WEKEX)*, pages 95–100, Montréal, Canada.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Ranking Sentences for Extractive Summarization with Reinforcement Learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT): Long Papers*, pages 1747–1759, New Orleans, LA, USA.

Owczarzak, K., Conroy, J. M., Dang, H. T., and Nenkova, A. (2012). An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9. Association for Computational Linguistics.

Paulus, R., Xiong, C., and Socher, R. (2018). A Deep Reinforced Model for Abstractive Summarization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

Peyrard, M. and Eckle-Kohler, J. (2016). Optimizing an approximation of ROUGE - a problem-reduction approach to extractive multi-document summarization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1836, Berlin, Germany, August. Association for Computational Linguistics.

See, A., Liu, P. J., and Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers*, pages 1073–1083, Vancouver, Canada.

Steinberger, J. and Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100.

Zopf, M. (2018). Auto-hmds: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*.