Multilingual Text Summarization Approaches - A Case Study on Generative and Extractive Methods

Masterarbeit

im Studiengang Computing in the Humanities der Fakultät Wirtschaftsinformatik und Angewandte Informatik der Otto-Friedrich-Universität Bamberg

Lehrstuhl für Medieninformatik

Verfasser: Giulia Dal Cin

Prüfer: Prof. Dr. Andreas Henrich

Bamberg 2026

URN: urn:nbn:de:bvb:473-irb-112942x

DOI: https://doi.org/10.20378/irb-112942

Abstract

In recent years, research in the field of automatic text summarization (ATS) has mainly focused on improving model performance, but it has rarely considered the context and the purpose for which summaries are produced. Therefore, in this master's thesis, five multilingual ATS scenarios are defined, and each of them is associated with a purpose and some specific requirements. These scenarios are used to evaluate and compare the summaries produced by three ATS systems: extractive algorithm LexRank, pre-trained language model mLongT5, and large language model Mistral NeMo. Both quantitative and qualitative evaluation is performed.

Results show that LexRank often fails at writing well-structured and coherent summaries; to a minor extend, mLongT5 does as well. In some of the five scenarios, both systems also produce summaries with insufficient information coverage. Additionally, mLongT5-generated summaries often contain factually incorrect statements or hallucinations. Problems linked to factually incorrect content, hallucinations and insufficient information coverage also occur in NeMo-generated summaries, but only rarely. Additionally, NeMo often does not respect length requirements, and it sometimes switches language in its summaries. Despite these problems, NeMo has good results in almost every scenario, outperforming the other two systems. However, performance differences between the systems vary based on the scenario.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Thesis structure . . . 2

2 Background and related work **3**
- 2.1 Summarization approaches . . . 3
  - 2.1.1 Extractive summarization . . . 3
  - 2.1.2 Abstractive summarization . . . 4
- 2.2 Summary evaluation methods . . . 4
  - 2.2.1 Automatic evaluation . . . 5
  - 2.2.2 Manual evaluation . . . 6
- 2.3 Constraining the task of ATS . . . 7

3 Scenarios **8**
- 3.1 Generation of introduction paragraphs for news articles . . . 8
- 3.2 Abstract generation for scientific papers . . . 9
- 3.3 Web snippet generation . . . 10
- 3.4 School material summarization . . . 11
- 3.5 Email thread summarization . . . 11

4 Implementation **12**
- 4.1 Datasets . . . 12
  - 4.1.1 Existing datasets . . . 12
  - 4.1.2 Self-collected datasets . . . 20
- 4.2 Models and algorithms . . . 22
  - 4.2.1 LexRank . . . 22
  - 4.2.2 mLongT5 . . . 23
  - 4.2.3 Mistral NeMo . . . 24

5 Evaluation **26**
- 5.1 Evaluation metrics . . . 26
  - 5.1.1 Automatic quantitative evaluation . . . 26
  - 5.1.2 Qualitative evaluation . . . 28
- 5.2 Quantitative evaluation results . . . 30
  - 5.2.1 News scenario . . . 30
  - 5.2.2 Abstract generation scenario . . . 34
  - 5.2.3 Web snippet generation scenario . . . 34

5.2.4 School material summarization scenario 35
5.2.5 Email thread summarization scenario 36
5.3 Qualitative evaluation results 37
5.3.1 News scenario 37
5.3.2 Abstract generation scenario 48
5.3.3 Web snippet generation scenario 52
5.3.4 School material summarization scenario 55
5.3.5 Email thread summarization scenario 57
5.4 Discussion 60
5.4.1 Results per scenario 60
5.4.2 Results per ATS system 61
5.4.3 Correlation between ROUGE scores and manual evaluation 63
5.4.4 Usefulness of other metrics 64

6 Conclusions **65**

References **66**

A Datasets **73**

B NeMo Prompts **78**

C Results **79**

List of Tables

1 Languages and datasets per scenario. 14
2 Average article and reference summary length per preprocessed dataset in the news scenario. 14
3 Average article and summary length per preprocessed dataset in the paper generation scenario. 15
4 Snippet of an article containing a subtitle (highlighted in red). 17
5 Snippet of an article containing a subscribers-only hint, followed by an invitation to read another article (both highlighted in red). 17
6 Snippet of an article in which the first paragraph is missing. Please follow the link to the source to compare it to the original article. 18
7 Example of summary that does not have enough context without the article's title. 19
8 Composition of the train split of the sumstew dataset. 24
9 Average quantitative evaluation scores for news scenario. R-1, R-2 and R-L stand for ROUGE-1, ROUGE-2 and ROUGE-L F-measure scores. The abbreviations in the remaining columns stand for compression ratio, word count, sentence count, extractivity, and truncated summaries. 31
10 Length distribution (in words) of summaries in news scenario, per dataset and per method. 31
11 Example of mLongT5 summary containing several spelling, grammar and factual mistakes, as well as minor reformulations. The former are highlighted in red (while the corresponding correct information in the original text is highlighted in green), the latter in blue (both in summary and in original text). 33
12 Average quantitative evaluation scores for abstract generation scenario. . . . 34
13 Average quantitative evaluation scores for web snippet generation scenario. Compression ratio, word count, sentence count and extractivity were computed **before** truncating summaries longer than 160 characters. The last table column (*> 160 char.*) shows how many summaries exceeded the length limitation and therefore had to be truncated. 35
14 Average quantitative evaluation scores for school material summarization scenario. 36
15 Average quantitative evaluation scores for email thread summarization scenario. 36
16 Average manual evaluation scores for news scenario. 37

17 Example of LexRank summary containing a pronoun without a referent (highlighted in red) and unrelated sentences. 38
18 Example of LexRank summary containing an incomplete quote, where quotation marks (highlighted in red) are closed, even though they have never been opened. 38
19 Example of LexRank summary in which the juxtaposition of two sentences that are not close to each other in the original text (where they are highlighted in bold) results in factual correctness problems. 39
20 Example of LexRank summary with insufficient information coverage. The original text is about the work of sculpture and painter Thomas Schütte and his exhibitions in London, but the summary does not allow this to be understood. 40
21 Example of LexRank summary leveraging lead bias and scoring a 5 in each of the Q1–Q5 metrics. 40
22 mLongT5 summary containing a grammar mistakes and a wrong lexical choice (both highlighted in red). 41
23 mLongT5 summary containing redundancy ("British Britons") and lacking some context in the parts highlighted in blue. Please note that grammar mistakes are ignored here. 41
24 mLongT5 summary containing a general factual mistake, a partially wrong quote and a hallucination. Mistakes are highlighted in red, the corresponding correct information in the original text in green. Please note that grammar mistakes and coherence problems are ignored here. 42
25 Example of mLongT5 German summary containing several short, simple sentences. Grammar mistakes are ignored. 43
26 NeMo summary with misleading use of adverbs "however" and "still" (highlighted in red). Their use probably aims to express an opposition to what is written in the first summary sentence. However, since the second summary sentence already mentions some of the difficulties of the Colorado Avalanche, the use of the red-highlighted adverbs in the third sentence is misleading. 43
27 Example of NeMo summary containing some grammar mistakes (highlighted in red) and an adverb that is not semantically coherent to the rest of the text (highlighted in blue). 44
28 Example of NeMo summary containing a word used improperly (highlighted in red). 45
29 Example of NeMo summary containing a factual mistake (highlighted in red, while the corresponding correct information in the original text is highlighted in green). 45
30 Example of NeMo summary containing a sentence in English at the end (highlighted in red). 46
31 Example of moderately repetitive NeMo summary. The information that is presented several times is highlighted in blue. 47

32 Example of NeMo summary containing a long, not easily readable sentence (highlighted in blue) and using a determinative article (highlighted in red) to refer to something that has not been mention in the summary before, resulting in a lack of context. 47
33 Example of NeMo summary containing a word in English at the end (highlighted in red). 48
34 Average manual evaluation scores for abstract generation scenario. 48
35 Example of LexRank summary lacking some context and containing redundancy (highlighted in blue) and an apparent contradiction (highlighted in red). 49
36 Example of NeMo summary using impersonal forms (highlighted in red). Corresponding personal forms used in the original abstract of the article are highlighted in green. 51
37 Average manual evaluation scores for web snippet generation scenario. Before performing manual evaluation, all summaries longer than 160 characters were truncated. This is important to keep in mind when looking at data about lead bias. 52
38 Example of LexRank summary scoring 5 in Q1–Q5. The summary consists of the first sentence of the original text. 52
39 Example of LexRank summary scoring 1 in Q4 and Q5. The website to summarize (available at the link in field *Source*) is a guide on how to write an eulogy for a dead grandparent. 53
40 The first sentence in this German mLongT5 summary is a hallucination giving a (wrong) definition of what the EU is. The second sentence is redundant and written in a childish style. Moreover, the summary completely fails to convey the topic of the article, which is how the EU manages its finances. 53
41 NeMo summaries of the same text in different languages. The English summary stands out as the only one that truly gives an overview of the content of the original text. 54
42 Average manual evaluation scores for the school material summarization scenario. 55
43 Example of NeMo summary containing an extractive chunk (in bold) with a demonstrative adjective missing a referent (highlighted in red). In the original text, the sentence containing the reference is highlighted in green. 56
44 Average manual evaluation scores for email thread summarization scenario. 57
45 Example of mLongT5 summary containing both verb forms in the third and in the second person. The use of the second person ("you") comes from extraction of a phrase from the original text (highlighted in bold both in text and summary) that has not been reformulated properly. 59
46 Example of mLongT5 summary confusing who wrote what to whom. 59
47 Example of NeMo summary containing an incorrect fact (highlighted in red, while the corresponding correct information in the original text is highlighted in green). 60

48 Example of NeMo summary not focusing on the result on the exchange. For readability reasons, the preprocessed email thread is reported in cell *Original thread*, even though the summary summarizes the corresponding unprocessed thread. 61

49 Example of summary that contains information not given in the text (highlighted in red). 74
50 Example of a PubMed article in which a sentence is missing (highlighted in red). 75
51 An unprocessed email from the Enron dataset. 76
52 The same email as in Table 51, but preprocessed. 77

53 NeMo prompts per scenario and per language. 78

54 Example of LexRank summary with incorrect sentence segmentation. The sentence highlighted in red is incomplete, because the dot after "B.1.1.7." in the original text has been interpreted as a full stop signaling the end of a sentence. The corresponding sentence in the original article is highlighted in green. 80
55 Example of mLongT5 summary with a pronoun with unclear referent (highlighted in blue). 80
56 Example of mLongT5 summary containing a conjunction used in a misleading way (highlighted in blue). 81
57 Example of NeMo summary containing incorrect information (highlighted in red). The correct information is highlighted in green in the original text. 81
58 Example of minimal information addition in a NeMo summary. The original text only talks of "Hachette", whereas NeMo, in its summary, uses the complete name of the company, "Hachette Book Group". All mentions of the company name are highlighted through bold text. 82
59 Example of NeMo summary containing additional information (highlighted in red): the information on Francesco Schettino, although correct, is not contained in the original text. For space reasons, the original article is not reported completely, but it can be found in GitLab, or at the link in field *Source*. 83
60 Example of NeMo summary containing a hallucination (highlighted in red): the website mentioned in the summary is not mentioned in the original text, and it does not exist either. For space reasons, the original article is not reported completely, but it can be found in GitLab, or at the link in field *Source*. 84
61 Example of NeMo summary containing too much *interpretation* (highlighted in red): we do not know if the recipe is "a popular choice", we only know that the author of the article chose to try it because they found it interesting (highlighted in green in the original text). 85
62 NeMo summary with numbered sentences. 86

63 This NeMo summary only reports information from the first part of the article, and does not mention what most of the article (highlighted in blue) is about. For space reasons, the original article (which can be found in GitLab) is not reported completely; the part not appearing here would also be highlighted in blue. 87
64 Example of NeMo summary containing a factual mistake (“75”, highlighted in red – the correct number, 72, is highlighted in green in the original text. Moreover, this summary contains a hallucination (sentence highlighted in red). 88
65 Example of NeMo summary with improvable information coverage. The summary talks about a speech that Macron held on work ethics, but it does not mention in which occasion he held it. This missing information (highlighted in green in the original text) is not strictly necessary, but it would still help contextualize the speech. 89
66 Example of LexRank summary lacking syntactic and semantic connection between the sentences. 90
67 Example of NeMo summary only focusing on the results of the study, and not introducing the topic or the research question. 91
68 Example of NeMo summary focusing on the quantitative results of a paper (highlighted in blue both in the summary and in the original abstract) more than on the qualitative ones (green), even though the latter play a bigger role in the original text. Also notice that the summary is in English even though the original paper is in French. 92
69 Example of LexRank summary of an original text with two ordered lists. The lists are only partially extracted: in the first case, the text of 3 out of 4 list items is (partially) extracted, but only the first item with its number. It the second case, only the first item number is extracted. List content is highlighted in blue in the summary, and the corresponding sentences are highlighted in blue in the original text. 93
70 Example of NeMo summary with insufficient information coverage, which retrieves most information from the first half of the text. In the original text, sentences from which information was retrieved are highlighted in blue. 94
71 Example of LexRank summary containing a misleading header and sentences written by different people. Sentences used in the summary are highlighted in bold in the original email thread. For space reasons, only a part of the original email thread is reported here, and some line breaks have been removed. The whole thread has already been reported in Table 51. . 95

List of Figures

1 Website containing an introduction which is separated from the rest of the text. 75

Chapter 1

Introduction

1.1 Motivation

In the last decades, the amount of textual content available on the Internet has largely increased. When searching for information on the Internet, one is often presented with an overwhelming number of results in textual form. It is hardly possible for a single user to manually process all this information. This results in the necessity to summarize large amounts of textual data in a fast and effective way. For this reason, Automatic Text Summarization (ATS), a sub-field of Natural Language Processing (NLP), has gained a lot of significance in recent times. Given one or more input documents, ATS systems summarize them in a shorter text which distills the key points from the original document(s), and which is produced for a particular user and a particular task [Maybury, 1995]. Based on the number of input texts, a differentiation between single-document and multi-document summarization can be made.

Multilingual text summarization focuses on systems that can process input and produce output in several languages.[1] Historically, most resources used in ATS have been in English, and data in other languages has been scarce. This poses problems both during training of multilingual summarization systems, and during evaluation of non-English automatically generated summaries, since ROUGE [Lin, 2004], the most common automatic evaluation metric, relies on the use of at least one reference summary. However, in the last years, some efforts have been made to create non-English ATS datasets, and some multilingual summarization models have been developed.

There are two main approaches to ATS: extractive and abstractive (a.k.a. generative). Extractive systems select some sentences from the input text and put them together to create a summary, without changing them. Abstractive systems, on the other hand, reformulate the content of the input text to generate a summary.

In recent years, the advent of Large Language Models (LLMs) has strongly influenced NLP. Therefore, the generation of summaries through LLMs can be added to the two above-mentioned ATS approaches. Research has shown that human evaluators prefer LLM-generated summaries to summaries generated through other methods [Pu et al., 2023]. Still, the use of LLMs is not entirely unproblematic: for example, their training

[1] In particular, given an input text in language A, a multilingual ATS system will produce a summary in language A. This differentiates it from cross-lingual ATS, in which a system, given an input text in language A, will produce an output text in another language B.

process has significant environmental implications [Hadi et al., 2023]. Therefore, it makes sense to reflect on whether the use of LLMs is always indispensable in text summarization tasks. This leads back to a fundamental question: for what tasks can ATS be used, and what requirements do these tasks have?

In his above-mentioned definition of a summary, Maybury considers the user(s) and task(s) for which the summary is created. However, little research has focused on these two aspects in the last years. Therefore, this thesis has the following objectives:

1. Define five multilingual ATS scenarios, each of them with a goal, some specific requirements, and one or more corresponding datasets.

2. Define some quantitative and qualitative evaluation metrics basing on the scenarios requirements.

3. Select three existing summarization methods (an extractive one, an abstractive one and a LLM) and evaluate their performance on the five scenarios.

4. Compare the performance of the three methods and draw some conclusions on which method(s) best suit which scenarios, and on whether computationally more expensive methods always yield better results.

1.2 Thesis structure

In this master's thesis, chapter 2 goes into more detail on different summarization approaches, on summary evaluation methods and on how to define summary requirements. Chapter 3 presents the five aforementioned scenarios. Chapter 4 describes the datasets and the summarization systems used in this work. Chapter 5 presents the evaluation metrics used in this work and the quantitative and qualitative evaluation results. Lastly, Chapter 6 recapitulates the contributions of this master's thesis, draws some conclusions and adds some final remarks.

Chapter 2

Background and related work

2.1 Summarization approaches

2.1.1 Extractive summarization

Extractive summarization was the first ATS approach to develop, with the first algorithms being developed in the late 1950s. Extractive algorithms rank the sentences in the original text by importance, then extract them to create a summary. There are several methods to rank sentences: statistical ones, graph-based ones, and semantic-based ones [Giarelis et al., 2023].

The Luhn algorithm [Luhn, 1958], developed in 1958, is an example of a statistical approach to sentence ranking. It considers word frequency of non-stopwords, and assumes words with higher frequency to be more important. Consequentially, sentences containing many high-frequency words are considered important.

In 2001, Gong and Liu [2001] applied Latent Semantic Analysis (LSA) to text summarization. The result is an example of a semantic-based algorithm. It first models the input document as a term-by-sentence matrix, in which each value represents the frequency of a word in a sentence. Then, it applies Singular Value Decomposition (SVD) to this matrix, which highlights the key semantic features of the input document, such as semantically related terms and sentences. This information is used to rank and extract sentences.

TextRank [Mihalcea and Tarau, 2004] and LexRank [Erkan and Radev, 2004] are two graph-based extractive algorithms from the early 2000s. They both represent the original text as a weighted graph, where each node is a sentence and edges between nodes represent connections between sentences. These connections are computed through cosine similarity of sentence vectors. The two algorithms differ in some specifics concerning graph representation, similarity computation and sentence ranking.

In comparison to abstractive summarization, extractive summarization is fast and less resource-consuming. Moreover, the sentences contained in the summaries are factually correct, and there is no risk of hallucinations. However, the sentences in extractive summaries often lack coherence, or the necessary context to make sense when read together. Moreover, extractive summaries cannot synthesize, compress or combine information to summarize important concepts.

In recent years, progress in deep learning models has significantly advanced text summarization. Some of these models have been applied to extractive summarization: for example, BertSum [Liu and Lapata, 2019] leverages transformer model BERT [Devlin

et al., 2019] to generate extractive summaries. However, in the last years, the focus of the ATS community has clearly shifted to abstractive summarization.

2.1.2 Abstractive summarization

Gupta and Gupta [2019] distinguish between three categories of abstractive summarization: structure-based (using structures such as graphs or trees), semantic-based, and deep-learning-based. The latter has become the predominant approach in the last years. In this context, transformer-based architectures [Vaswani et al., 2017] have proven to deliver the best results [Giarelis et al., 2023]. The transformer architecture implements the encoder-decoder framework and uses self-attention to weigh the importance of different parts of the input. This architecture is the foundation of most Pre-trained Language Models (PLMs), which are first trained on a large amount of data, and can then be fine-tuned on smaller, domain-specific datasets to specialize in a task (which in our case would be ATS). These models leverage transfer learning, which allows them to apply the general knowledge that they acquired in pre-training to domain-specific tasks [Wang et al., 2023]. Some examples of pre-trained language models that rely on the transformer architecture are T5 [Raffel et al., 2020], GPT [Radford et al., 2019] and BART [Lewis et al., 2020].

Compared to extractive approaches, abstractive ATS techniques can generate more fluent and compressed summaries that resemble human-written summaries [El-Kassas et al., 2021]. However, abstractive summaries might contain grammar mistakes, incorrect statements or hallucinations.

LLMs

Research has shown that scaling PLMs (e.g., the number of parameters they use, or the amount of data they are trained on) leads them to show surprising abilities in solving complex tasks [Zhao et al., 2023]. These scaled PLMs have been named Large Language Models. Unlike PLMs, LLMs can solve a wide range of tasks, do not require fine-tuning and offer a prompting interface that makes them easy to use for people outside the NLP community. The summaries produced by LLMs are abstractive.

LLM-generated summaries can be perceived as even better than human-generated summaries [Pu et al., 2023]. However, there are several ethical and environmental concerns linked to the use of LLMs. Hadi et al. [2023], for example, provide an overview of these concerns, as well as of the history and applications of LLMs.

2.2 Summary evaluation methods

Another important question in the field of ATS is how to evaluate automated summaries. It can be distinguished between intrinsic and extrinsic evaluation methods ([Galliers and Spärck Jones, 1993], [Lloret et al., 2018]). Intrinsic methods focus on the summary itself and mainly address three aspects: the readability of the summary, its information coverage, and its non-redundancy [Lloret et al., 2018]. Extrinsic methods, on the other hand, evaluate the utility of a summary with regard to the use case or task for which it was produced [Lloret et al., 2018].

Summary evaluation can be performed manually or automatically. Manual evaluation (a.k.a. human evaluation) is labor intensive and inherently subjective [Lloret et al., 2018].

Automatic evaluation, on the other hand, often does not have a high correlation with human judgment, as will be noted in the next subsection.

2.2.1 Automatic evaluation

Information coverage

In recent years, most papers evaluating ATS models have focused on intrinsic evaluation, and in particular on information coverage. In the vast majority of the cases, information coverage is computed with ROUGE [Lin, 2004], an automatic metric that focuses on n-gram overlap between the candidate summary and one (or, more rarely, several) reference summaries. In particular, ROUGE-1 is based on the number of overlapping unigrams, ROUGE-2 on the number of overlapping bigrams, and ROUGE-L on the longest common sequence of words between candidate and reference summary. One of ROUGE's problems is that it does not match alternative phrasings, so for example "a big car" will not be matched to "a large vehicle", even though the two phrases have very similar meanings [Tratz and Hovy, 2008]. Moreover, despite its popularity, several studies have shown that ROUGE is only weakly correlated to human judgment (i.a. [Liu and Liu, 2009], [Kryściński et al., 2019]).

Another automated metric that measures content overlap is BE (Basic Elements) [Lin et al., 2006]. Basic elements are small units of text consisting of a head (a major syntactic constituent such as a noun), a modifier and the relationship between the two. BEs in the candidate summary are matched to BEs in the reference summary. Several matching strategies exist: lexical identity, lemma identity, synonym identity, semantic generalization. The former are considered the most exact matches, and therefore they result in higher scores than the latter. In comparison to ROUGE, BE allows greater flexibility and granularity in matching, but it is also more complex and resource intensive.

Another approach to computing content overlap is by using contextual word embeddings. BERTScore [Zhang et al., 2019], for example, leverages contextual word embeddings from pre-trained language model BERT ([Devlin et al., 2019]). A contextual word embedding represents a word as a vector by also considering its surrounding words in the sentence in which it appears. This allows it to take into account the fact that words can have different meanings based on the context in which they are used. In BERTScore, content overlap is computed as the cosine similarity between each word vector in the candidate summary and each word vector in the reference summary.

Other intrinsic evaluation metrics

There are numerous other metrics that compute content overlap (e.g. BLEU [Papineni et al., 2002], or METEOR [Banerjee and Lavie, 2005]). However, efforts have also been made to develop automatic metrics that evaluate other intrinsic aspects. Grusky et al. [2018], for example, implemented a metric called extractive fragment density, which computes the average length of extractive fragments in the candidate summary. Therefore, this metric is useful to determine how abstractive or extractive a summary is.

Bommasani and Cardie [2020] introduced a metric to evaluate redundancy in summaries. This is done by computing ROUGE-L scores for every pair of distinct sentences in a summary, which allows it to identify if summary sentences have similar content (meaning that they are redundant). Moreover, Bommasani and Cardie also developed a metric to

evaluate semantic coherence within summaries. This metric leverages aforementioned language model BERT to predict the probability of each successive sentence in the summary, based on the previous sentence.

Goel et al. [2021] presented, among others, three metrics that focus on the location of summary content in the original text: position, dispersion and ordering. Position matches sentences from the summary with similar sentences from the original text, then computes the mean position of the matched sentences in the original text. It can be useful in identifying lead bias (meaning that the summary only contains information from the beginning of the input text) or other positional biases. Similarly, dispersion is "the degree to which summary sentences match content that is distributed broadly across the article versus concentrated in a particular region" [Goel et al., 2021, 19]. Finally, ordering evaluates whether content in the summary is ordered similarly as it is in the original text.

Some other metrics, such as QAFactEval [Fabbri et al., 2022], try to measure factual consistency in summaries. There are several approaches to this, one of which is using automatic question generation to formulate questions about the content of the input text and comparing the right answers to the content of the candidate summary. However, automatically evaluating factual correctness remains a complex, unsolved task [Gehrmann et al., 2023].

2.2.2 Manual evaluation

Like ROUGE, BE and BERTScore, some manual evaluation techniques, such as the pyramid method ([Nenkova and Passonneau, 2004], [Nenkova et al., 2007]), also focus on information coverage. The pyramid method requires several human written summaries, from which information bits called Summary Content Units (SCUs) are extracted and ranked in a pyramid model. SCUs appearing in several reference summaries are ranked higher; therefore, candidate summaries containing them are considered good, and vice-versa. On the one hand, the fact that several reference summaries are used improves the quality of the evaluation; on the other hand, generating these summaries also requires high effort.

The 2006 Document Understanding Conference (DUC 2006) [Dang, 2005] evaluates both readability and content of summaries. Readability is assessed through five criteria:

- Grammaticality, meaning that the summary "should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read" [Dang, 2005, 2].
- Non-redundancy, meaning that information should not be repeated unnecessarily.
- Referential clarity, meaning that the referent of pronouns or noun phrases should be easy to identity.
- Focus, meaning that "the summary should have a focus; sentences should only contain information that is related to the rest of the summary" [Dang, 2005, 4].
- Structure and coherence, meaning that the summary should "build from sentence to sentence to a coherent body of information about a topic" [Dang, 2005, 4].

Summary content in DUC 2006 is evaluated extrinsically, by judging if the information in the summary helps fulfill the information need defined in a query-oriented summarization

task. An example of query is the following: "Describe theories concerning the causes and effects of global warming and arguments against these theories" [Dang, 2005, 2].

Other studies use a list of explicitly formulated questions for manual evaluation. Lloret et al. [2013], for example, formulates three questions to evaluate summaries of scientific papers. The first question focuses on whether the summary provides relevant information; the second question evaluates whether the summary provides a general idea of the contents of the original paper; the third question judges whether the summary can be used as a substitute of the original abstract. The last two questions define respectively what Spärck Jones [1999] defines as style and use of a summary, as will be explained in the next section.

2.3 Constraining the task of ATS

As seen in the last sections, intrinsic evaluation is far more common than extrinsic evaluation. This is because, in order to perform the latter, a task or use case for the summaries has to be defined, as was done in DUC 2006. However, as ter Hoeve et al. [2020] observe, current ATS research often solely focuses on improving model performance on ROUGE, and does not define use cases or requirements for the summaries.

Kryściński et al. [2019] note that selecting what information is relevant for a summary is a complex task that depends on the expectations and knowledge of the target reader. They show that the usual application of ATS, in which a model is just given a text to summarize, leaves the task of ATS too underconstrained to be solved in a satisfactory way. Therefore, it can be assumed that constraining it by defining use cases or summary requirements would not only help to perform more complete evaluation, but also help to generate better summaries.

Research by Spärck Jones [1999] can help define summary constraints. In particular, Spärck Jones defines three classes of factors to take into account in the context of ATS: input factors, output factors and purpose factors.

- Input factors describe the text(s) that should be summarized, e.g. with regard to length, language, structure, genre.
- Output factors describe the desired summary, for example how it should be structured (should it have headings?) or what style it should have. Style can be informative, meaning that it reports information from the original text; indicative, meaning that it states the topic of the original text, but does not go into detail about the pieces of information it presents; critical, if the summary reviews the original text critically; aggregative, if it puts several input texts in relationship to another.
- Purpose factors, which according to Spärck Jones is the most important class of factors, establish what the summary should be used for. Among others, they describe the the target audience and its previous knowledge, as well as the use of the summary. Spärck Jones suggests five possible uses: retrieving input text, previewing the original text, substituting it, prompting to read it, or refreshing one's memory about its content.

Chapter 3

Scenarios

In this chapter, five scenarios for ATS are presented. Basing on the research presented in Section 2.3, for each scenario, the following requirements are defined:

- a goal,
- a style and a use (as introduced by Spärck Jones [1999] and described in Section 2.3),
- a target summary length,
- some evaluation criteria, specifying what aspects should be prioritized during the qualitative evaluation of the summaries.

3.1 Generation of introduction paragraphs for news articles

News summarization is the most common use of ATS in scientific literature, because, as Aumiller et al. [2023] observes, news resources are overrepresented in summarization datasets. These datasets use the introductory paragraph to a news article (which may be in form of regular text or of bullet points) as its reference summary. The length and information coverage of these reference summaries vary from newspaper to newspaper, and therefore from dataset to dataset. Still, independently of dataset differences, it is important to note that these reference summaries often do not summarize the whole content of the article that they preview. In many cases, they only mention one of the aspects covered in the article. So their information coverage is usually low, which means that they should not be seen as a substitute for the article, but rather as an introduction to it, or as a teaser to catch the reader's attention (for these reasons, the name of this scenario is not "news summarization"). For example, online newspapers *Süddeutsche Zeitung*[1] and *Le Monde*[2] (from which articles are contained in the MLSUM dataset [Scialom et al., 2020]) use these texts not only as introductory texts to their articles, but also as a caption when sharing their articles on their social media pages.

Therefore, the goal of this scenario can be defined as follows: given a news article, to provide a short text that conveys information about the topic of the article.

[1] `www.sueddeutsche.de`, last accessed: 3rd January 2025.
[2] `https://www.lemonde.fr/`, last accessed: 3rd January 2025.

The information coverage of the summaries to be produced for this scenario is flexible, meaning that both indicative and informative style can be accepted. Following Sparck Jones' categorization, the former means that the summary only states the topic of the article, while the latter means that it covers the most important information contained in the summary. Therefore, indicative style implies lower information coverage than informative style.

With reference to Sparck Jones's categorization, the most obvious use of the summaries in this scenario is to preview the original text. However, prompting to read the original text is also possible as a use. An example of a prompting summary would be: "Read the following article to find out more about topic X", where X is the topic of the article.

The target length of the summaries is also flexible: it can vary from a minimum of 20-25 words to a maximum length of 85-90 words (which roughly corresponds to a minimum of 1 and a maximum of 3 or 4 sentences).

When it comes to evaluation criteria, the form of the text (grammaticality, cohesion and coherence) is more important than its content. This is because, as already said, information coverage is flexible, as long as the summary gives an idea of the topic of the article. While it is important that the information from the article is not changed in the summary (i.e., the summary should be factually correct with regard to what is written in the article), it would be acceptable for the summary to contain some minor additional information, as long as this information is true. For example, if the article mentions a city without mentioning in which country it is located, it would be acceptable for the summary to add this information.

3.2 Abstract generation for scientific papers

The summarization of scientific articles is also a common ATS scenario. In this domain, the summarization of biomedical articles is particularly widespread, thanks to the existence of article retrieval systems such as PubMed, MEDLINE and BioMed central, which are used to create datasets [Luo et al., 2024].

In summarization datasets of scientific articles, the article's abstract is used as its reference summary. Therefore, the goal of this scenario is defined as abstract generation. Koopman [1997] claims that, in papers on computer architecture, an abstract should provide information on the motivation of the paper, its problem statement, its approach, its results, and the conclusions drawn from the results. However, particular types of papers, such as surveys, literature reviews, or biomedical case reports, do not always contain results or conclusions. Moreover, some papers from the humanities, as they can be seen in the TermITH dataset [ATILF et al., 2017], do not have a section on methods. Therefore, generalizing Koopman's list of requirements to all types of papers, we can expect an abstract generated by an ATS system to provide information on:

- the topic or motivation of the paper that it summarizes,
- the paper's central question or contribution,
- the paper's methods, if available,
- the paper's main findings and implications, if available.

Abstracts can be structured in paragraphs with headings ("Background", "Methods" etc.), or unstructured [Drury et al., 2023]. Therefore, both formats would be acceptable in a summary.

Style should be between informative and indicative. On the one hand, informativeness is required, since the summary should cover the main points from the paper, as listed above. On the other hand, considering the high compression ratio required in the summary, not all important information can be covered – in this sense, the summary will tend more towards indicative style.

Several uses are possible for this scenario: an abstract can be seen as a preview to an article, but also as a substitute of an article (if the reader only wants to roughly know what the article is about, without going into detail), or a way to refresh one's memory about the article's content.

In the corpus used by Lloret et al. [2013], which consists of 50 research articles, the average abstract length is 6 sentences. In the filtered version of the PubMed dataset [Cohan et al., 2018] which will be used in this paper (cf. Section 4.1.1), the average abstract is 8 sentences, or 175 words; in filtered TermITH (also presented in Section 4.1.1), it is 5 sentences, or 129 words. Therefore, for this scenario, the accepted target length is variable, and can range from approximately 115 to approximately 200 words.

For the evaluation of this scenario, both the form and the content of the summaries are important. As for the content, all the above-listed information should be contained in the summary, and the text should not be redundant.

3.3 Web snippet generation

On Search Engine Result Pages (SERPs), each result is usually presented as a triple consisting of the URL of the corresponding webpage, its title and a snippet of its content. The extraction or generation of a snippet can be seen as the summarization of the content of the page. The goal of such a summary should be to give a good overview of the content of the webpage, in order to help users determine if the result is relevant to their search.

As for news summaries, summary style can be both indicative or informative. Considering that web snippets are very short, it will probably tend more towards indicative. The use of this scenario can be to preview the content of the webpage, or to prompt the user to read it.

According to Google for Developers, snippets are "truncated in Google Search results as needed, typically to fit the device width".[3] Google Developers also provides some examples of good-quality snippets.[4] These snippets consist of 2 or 3 short sentences each, with an average of 123 characters per snippets. Therefore, this scenario has a target summary length of about 120 to 130 characters. Some self-conducted tests on Google's SERP have shown that, on desktop devices, snippets seem to be cut off after approximately 160 characters. For this reason, summaries generated for this scenario should by no means be longer than 160 characters.

As in the news scenario, the form of the text is more important than its content. As

[3] `https://developers.google.com/search/docs/appearance/snippet#meta-descriptions`, last accessed: 4th January 2025.

[4] `https://developers.google.com/search/docs/appearance/snippet#use-quality-descriptions`, last accessed: 4th January 2025.

long as the snippet conveys the topic of the webpage, information coverage is flexible. Moreover, as in the news scenario, in some cases it would be acceptable for the summary to contain some minor additions to the content of the original text. If the text is a salad recipe, for example, its snippet could mention that the recipe is delicious. However, it is important that this added information is not factually wrong ("delicious" expresses an opinion, so it is neither right nor wrong; "vegetarian", on the other hand, might be wrong).

3.4 School material summarization

The fourth scenario is about summarizing school material. The idea behind this scenario is to provide school students with summaries of texts that they have already read. The goal is to support the students when studying or when preparing for a test.

Given their purpose, the summaries should have a wide information coverage. Every concept presented in the source text should be included in the summary. Therefore, the style of the summaries should be informative. The summaries' use is to refresh the students' memory of the original text.

The target length of the summaries is variable, depending on the information density of the original text. If the source document repeats the same concepts several times, or if it contains several examples, a target length of 20% to 30% of the original text length will be sufficient. If the opposite is the case, target length can be up to 50% of the original text.

When it comes to evaluation, the summary's content is more important than its form. While the summary should be grammatical, cohesion and fluidity are not fundamental: as long as it does not hinder readability, summary sentences presenting different pieces of information do not need to be linked syntactically or semantically.

3.5 Email thread summarization

The summarization of email threads is not a common scenario in the ATS community, but it has been explored in some work, e.g. in Zhang et al. [2021].

In this work, only business emails will be considered, because this allows to formulate a more specific goal for the summaries. Business emails usually have a clear objective, such as to schedule a meeting or to receive some information. Therefore, the summary should provide information on what should be done, or what has already been done, and in order to achieve what objective.

Style should be indicative. The summary does not need to cover all the information contained in the email thread, but it does have to mention the result of the email exchange. For example, if the email thread is about scheduling a meeting and several time suggestions are made before reaching an agreement, the summary does not need to contain all of the time suggestions, but it should contain the final time for which the meeting was set.

The use of the summaries in this scenario is to refresh the user's memory on the result of the email exchange.

The summaries should be short and quick to read, so their target length can vary from 1 sentence to 3 short sentences.

As for the the school material summarization scenario, the focus during summary evaluation should be on the content of the summary, not on its form.

Chapter 4

Implementation

Each scenario, as described in Chapter 3, is associated to one or more datasets. These datasets are needed to test and evaluate the performance of three different summarization methods on the scenarios. This chapter presents the datasets in 4.1, and the three selected summarization methods in 4.2.

4.1 Datasets

The news scenario and the abstract generation scenario will be tested on a larger amount of data from several existing datasets, so that a quantitative evaluation of the results can be performed. The three remaining scenarios will be tested on very small datasets (two of which self-collected), because the collection of larger datasets for these scenarios would go beyond the scope of this master's thesis.

Table 1 provides an overview of the scenarios, the datasets with which they are associated, their language and their size. Table 2 compares the average article and summary length of the news summarization datasets used in this work (i.e., after preprocessing). and Table 3 does the same for the datasets of scientific articles. All datasets as they are used in this work are available in GitLab.[1]

4.1.1 Existing datasets

With the exception of the Enron Email Dataset, which will be handled in Section 4.1.1, all existing datasets used in this work were filtered and preprocessed as follows:

1. Duplicate texts and duplicate summaries were removed.
2. Texts under a certain length were removed.
3. Summaries under another certain length were removed.
4. Text-summary pairs under a certain compression ratio were removed.
5. In news datasets, all completely extractive summaries were removed. Text-summary pairs with a bigram overlap fraction over a certain value were also removed. This

[1] `https://gitlab.rz.uni-bamberg.de/minf/theses/ma-dal-cin`

allowed the removal of entries with summaries that were not completely extractive, but that were still using several "chunks of text" from the original text.

6. In news datasets, all line breaks were removed. This action was performed to make all news datasets consistent, since only some of them contained line breaks.

Steps 1 to 5 were performed using a dataset filtering Python library created by Aumiller et al. [2023].[2]

The choice to only keep dataset entries with abstractive summaries in the news datasets was met under the assumption these would be better reference summaries, because they synthesize content from several parts of the text, instead of just copying sentences.

Minimal text length, minimal summary length, minimal compression ratio and maximal bigram overlap fraction were set to different values depending on the language of the dataset. These values can be found in the dataset preprocessing notebooks in the aforementioned GitLab repository.

Some of the datasets required additional filtering or preprocessing steps. These are described in the next paragraphs, when relevant. The code used to filter each dataset can be found in GitLab as well.

After the datasets to be used for the news scenario and for the abstract generation scenario had been preprocessed, their size was reduced by selecting a number of random dataset entries and discarding the rest. Table 1 shows how many entries were selected per dataset. Only TermITH, having a small number of entries, was not reduced by size.

CNN/DailyMail

CNN/DailyMail [Nallapati et al., 2016] is one of the most popular English-language ATS datasets. It was originally created for question answering [Hermann et al., 2015], but was later adapted to text summarization by Nallapati et al. [2016]. As its name says, the articles are crawled from the websites of the *CNN*[3] and of the *Daily Mail*.[4] On its original website, each retrieved article was preceded by some bullet points summarizing its content. In the summarization dataset, the content of these bullet points is concatenated into a single string, which is used as the article's reference summary. For this reason, reference summaries in this dataset usually consist of several sentences (i.e., each former bullet point is a sentence) and are on average longer than in other news datasets (cf. Table 2).

As found on Hugging Face,[5] the dataset contained several non-breaking spaces. Since no other dataset used in this work contains non-breaking spaces, these were replaced with regular spaces.

One of the models that will be tested in this work, mLongT5 (cf. Section 4.2.2), was fine-tuned on a dataset that contains over thirty-nine thousands samples from the train split of CNN/DailyMail. For this reason, this thesis uses part of the test split of CNN/DailyMail, but none of its train split. During evaluation, the fact that mLongT5 was partially trained on CNN/DailyMail will be taken into account.

[2] Available at: `https://github.com/dennlinger/summaries`, last access: 6th January 2025.

[3] `https://www.cnn.com/`, last access: 6th January 2025.

[4] `https://www.dailymail.co.uk`, last access: 6th January 2025.

[5] `https://huggingface.co/datasets/abisee/cnn_dailymail`, last visited: 6th January 2025.

Table 1: Languages and datasets per scenario.

Scenario	Language	Dataset	Size
News article introduction generation	EN	Preprocessed subset of Newsroom [Grusky et al., 2018]	2500 texts
		Preprocessed subset of CNN/DailyMail [Nallapati et al., 2016]	2500 texts
	DE	Preprocessed subset of 20 Minuten [Rios et al., 2021]	2500 texts
		Preprocessed subset of MLSUM [Scialom et al., 2020]	2500 texts
	FR	Preprocessed subset of OrangeSum [Eddine et al., 2021]	2500 texts
		Preprocessed subset of MLSUM [Scialom et al., 2020]	2500 texts
Abstract generation	EN	Preprocessed subset of PudMed [Cohan et al., 2018]	1000 texts
	FR	Preprocessed subset of TermITH [ATILF et al., 2017]	779 texts
Web snippet generation	EN, DE, FR, IT	Self-collected parallel texts	10 texts per language
School material summarization	DE, FR	Self-collected texts	10 texts per language
Email thread summarization	EN	Enron Email Dataset [Klimt and Yang, 2004]	10 texts

Table 2: Average article and reference summary length per preprocessed dataset in the news scenario.

Preprocessed dataset	Avg. article length		Avg. summary length	
	Words	Sentences	Words	Sentences
CNN/DailyMail	719.49	36.26	52.07	3.92
Newsroom	914.78	45.05	26.88	1.36
20 Minuten	382.4	26.37	40.35	3.73
MLSUM-de	544.26	35.12	23.17	2.15
MLSUM-fr	666.84	39.67	29.79	1.89
OrangeSum	452.98	20.38	36.26	1.71

Table 3: Average article and summary length per preprocessed dataset in the paper generation scenario.

Preprocessed dataset	**Avg. article length**		**Avg. summary length**	
	Words	**Sentences**	**Words**	**Sentences**
PubMed	2559.53	95.24	174.78	8.33
TermITH	5160.71	250.91	128.92	5.16

Newsroom

Newsroom [Grusky et al., 2018] is an English-language summarization dataset containing news articles crawled from several online publishers. As reference summaries, it uses the summaries provided in the HTML metadata of the article's page.

The dataset, as downloaded from its official website,[6] contained several HTML escape strings. To make the data consistent with the other news datasets used in this work, these strings were replaced with their corresponding UNICODE characters.[7]

20 Minuten

20 Minuten [Rios et al., 2021] is a German-language dataset crawled from Swiss online newspaper *20 Minuten*.[8] This means that the dataset uses Swiss German spelling, the biggest difference to standard German being the use of a double S (*ss*) instead of the letter *ß*. As in the CNN/DailyMail dataset, the reference summaries consists of a concatenation of the sentences that appear at the beginning of every article in form of bullet points.

Because of its short, relatively easy summary sentences, the dataset was originally created for text simplification. However, it is also well suited for the task of text summarization.

In comparison to other datasets, the data is relatively clean: as far as a manual analysis of some random samples could show, the dataset does not contain any titles, subtitles, figure captions or other noise elements.

MLSUM

When MLSUM [Scialom et al., 2020] was released, it was one of the first large-scale multilingual summarization datasets containing several European languages. In this thesis, we are using its German and its French subsets, which from now on will be referred to as MLSUM-de and MLSUM-fr. MLSUM-de articles are retrieved from the website of German newspaper *Süddeutsche Zeitung*,[9] MLSUM-fr articles from the website of French

[6] `https://lil.nlp.cornell.edu/newsroom/`, last access: 7th January 2025.

[7] Another reason for replacing the aforementioned strings is that summaries produced by LexRank [Erkan and Radev, 2004], the extractive ATS systems that will be used in this work, would contain them too. This would create problems during evaluation, e.g. in tokenization, which is necessary to compute the length and compression ratio of the generated summaries.

[8] `https://www.20min.ch/`, last access: 6th January 2025.

[9] `www.sueddeutsche.de` add last access: 4th January 2025.

newspaper *Le Monde*.[10] Both subsets use the introductory paragraph to news articles as their reference summary.

It has already been observed (e.g. by Aumiller et al. [2023]) that MLSUM contains a large number of fully extractive summaries. However, this will not be a problem in this work, since these summaries will be removed through the fifth preprocessing step described in Section 4.1.1.

Similarly to what has already been said for CNN/DailyMail, the dataset on which mLongT5 was fine-tuned also contains almost 2.6 thousands texts from the train split of the French MLSUM subset. For this reason, only samples from the test split of French MLSUM will be used in this work.

OrangeSum

OrangeSum [Eddine et al., 2021] is a French-language summarization dataset. According to the authors, its summaries are not catchy, "but rather convey the essence of the article" [Eddine et al., 2021, 9374].

The dataset, as found in Hugging Face,[11] was often missing a space between the punctuation sign signaling the end of a sentence (usually a full stop, sometimes a question mark) and the first letter of the following sentence. This lack of spaces was very widespread, appearing several times in each dataset sample, and led the two sentences between which the space was missing to syntactically appear like a single sentence. This would have been problematic for LexRank [Erkan and Radev, 2004], one of the ATS systems that will be used in this work, which extracts sentences as they are from the input text. Therefore, during the preprocessing of the datasets, a space was added after every point or question mark directly followed by a capital letter.

Common problems in news datasets

After the news datasets were filtered and preprocessed as described in the previous paragraphs, 30 to 50 random samples per dataset were manually analyzed. This helped identify some common problems, which can be separated into data cleanliness problems and semantical problems. As for the first category, following issues were found:

1. In Newsroom, OrangeSum, MLSUM-de and particularly often in MLSUM-fr, several news articles contain paragraph titles, as can be seen in Table 4. Being titles, these phrases are not followed by full stops. Therefore, from a syntactical point of view, they appear to be part of the sentence that follows them. This is especially problematic for LexRank, which extracts sentences from the input text as they are.

2. Quite often in MLSUM-fr, and less often in Newsroom, articles contain image descriptions, which do not end with a full stop (which is problematic for the same reason already mentioned for paragraph titles). Texts from CNN/DailyMail also contain image descriptions, which appear very often in the subset from *Daily Mail*. These image descriptions do end with a full stop, but they are still problematic, because, from a semantical point of view, they do not belong to the body of the article,

[10] `https://www.lemonde.fr/`, last accessed: 3rd January 2025.
[11] `https://huggingface.co/datasets/EdinburghNLP/orange_sum`, last access: 7th January 2025.

and interrupt its flow. Moreover, in CNN/DailyMail, they often repeat information that also appears in the body of the article, making the text redundant.

3. MLSUM-fr often contains sentences that encourage the reader to read a related article, such as "Also read: *Title of another article*"; less often, it also contains the sentence "Article réservé à nos abonnés", meaning that the article is only for subscribers. An example can be seen in Table 5. These sentences do not have any full stop at the end, and interrupt the flow of the article.

4. Some articles from Newsroom, OrangeSum and MLSUM-fr contain the author's name at the beginning, with no full stop afterwards. This is also the case in many CNN/DailyMail articles; in these cases, however, the name is followed by a full stop. CNN/DailyMail articles also contain the online newspaper's name at the beginning (often), or publishing date and time of the article (less often).

Other work looking at noise in scraped datasets (e.g. Kryściński et al. [2019]) found similar problems. Since it would be challenging to identify the above listed problems manually, the noise that results from them has to be kept in the data.

Table 4: Snippet of an article containing a subtitle (highlighted in red).

Text ID	mlsum-fr-1448 (in preprocessed dataset in GitLab)
Text	[...] Par mesure de précaution, les CDC recommandent dans l'immédiat de ne pas utiliser de cigarettes électroniques, quelles qu'elles soient. Symptômes d'une pneumonie lipidique Les problèmes respiratoires sont d'autant plus choquants qu'ils apparaissent subitement, chez des patients souvent jeunes et sans problème de santé. [...]
Source	https://www.lemonde.fr/international/article/2019/09/07/aux-etats-unis-cinq-morts-liees-a-la-cigarette-electronique_5507496_3210.html

Table 5: Snippet of an article containing a subscribers-only hint, followed by an invitation to read another article (both highlighted in red).

Text ID	mlsum-fr-1328 (in preprocessed dataset in GitLab)
Text	[...] Le dirigeant libéral, qui brigue un deuxième mandat, a rencontré la gouverneure générale Julie Payette pour lui demander de dissoudre la chambre basse du Parlement, conformément à son rôle de représentante de la reine Elizabeth II, chef de l'Etat. Article réservé à nos abonnés Lire aussi La « diplomatie du selfie » de Justin Trudeau « J'ai rencontré son excellence, la gouverneure générale, qui a accédé à ma demande de dissoudre le Parlement », a déclaré M. Trudeau à la presse. [...]
Source	https://www.lemonde.fr/international/article/2019/09/11/canada-trudeau-annonce-la-dissolution-des-communes-donnant-un-coup-d-envoi-des-legislatives_5509276_3210.html

As for semantical problems appearing in the data, the following points were observed:

1. In Newsroom, the first sentence(s) of an article are often missing. This was the case in approximately in 1/6 of the manually analyzed articles. An example can be seen

Table 6: Snippet of an article in which the first paragraph is missing. Please follow the link to the source to compare it to the original article.

Text ID	newsroom-1855 (in preprocessed dataset in GitLab)
Text	Similar collaborations between private companies and nonprofits will pose tricky questions under a policy intended to end earmarks to profit-making firms, which Obey helped shepherd through the House Democratic caucus last week. [...]
Source	https://web.archive.org/web/20100315191d_/http://www.washingtonpost.com/wp-dyn/content/article/2010/03/14/AR2010031402305.html

in Table 6. This means that important information is not in the article, and that the context of the article can hardly be reconstructed. Similarly, but less frequently, the end of long articles is missing in the Newsroom dataset. Articles missing their end part were also found in MLSUM-de, but only rarely. Both in the case of a missing article beginning and that of a missing end, it is possible that the reference summary associated to the truncated article only partially makes sense, since it might contain some information that is missing in the article.

2. In all datasets, it sometimes happens that reference summaries contain information that is not contained in the article to which they are associated, even if the article is not truncated. This can be seen in the Appendix (Table 49).

3. In all datasets (but most frequently in CNN/DailyMail), some reference summaries miss the necessary context to make sense on their own – an example can be see in Table 7. Often, the missing context or information can be found in the article's title (which, in the datasets, is neither part of the summary, nor part of the article). Therefore, in these cases, it would make sense to merge an article's title and its summary. However, these cases are a minority: in all the other cases, adding the article's title to the beginning of the corresponding summary would result in new issues, such as redundant summaries, or summaries lacking fluency, since most titles are noun phrases and not independent sentences.

In these three cases, the reference summary is not an appropriate summary of its corresponding article. This is problematic for evaluation metrics that compare the content of a candidate summary to that of a reference summary, such as ROUGE. It is also problematic for model training: if a reference summary contains information that is not included in the corresponding article, a model may be trained on writing summaries containing information that is not in the original text, leading to hallucinations.

The three above-listed issues do not appear often enough in datasets to be considered a significant problem, with the already-mentioned exception of article beginnings missing in Newsroom. Moreover, there is no easy way to automatically identify and remove dataset entries containing these issues. Therefore, they are still present in the datasets used in this work.

Table 7: Example of summary that does not have enough context without the article's title.

Text ID	mlsum-de-2099 (in preprocessed dataset in GitLab)
Title	Bahnrad-Olympiasiegerin Kristina Vogel ist querschnittsgelähmt
Summary	Die 27-Jährige war im Juni beim Training mit einem anderen Fahrer zusammengestoßen und gestürzt. Nun spricht sie erstmals über den Unfall.
Source	https://www.sueddeutsche.de/sport/kristina-vogel-querschnittsgelaehmt-1.4120963

TermITH

TermITH [ATILF et al., 2017] is a French-language dataset of scientific papers. It contains articles from several domains: archaeology, chemistry, linguistics, psychology, and information and communication sciences. Even though the dataset was not created for ATS, it is well suited for the abstract generation scenario, since every dataset entry contains a scientific paper and its abstract.

In the original dataset, every paper consists of a separate XML file in which each token is an XML element. Therefore, before performing the preprocessing steps mentioned in 4.1.1, these tokens had to be merged into single strings (one string per article or abstract). Details on how this was done can be found in GitLab.

After the dataset was preprocessed, a manual analysis of some random dataset samples showed that, rarely, some entries contains some mistakes which seem to result from the use of OCR. For instance: missing spaces, points instead of commas, curly brackets instead of round brackets, Greek letters not being recognized correctly. These errors are difficult to identify automatically, therefore they could not be removed from the dataset.

PubMed

PubMed [Cohan et al., 2018] is an English-language summarization dataset containing scientific papers from the biomedical field, crawled from the homonymous online database *PubMed*.[12] As can be seen in Table 3, its papers are on average shorter that in TermITH, but their abstracts are longer.

Like TermITH, before it could be processed following the steps described in 4.1.1, PubMed needed some additional preprocessing. The code used to perform these additional preprocessing steps can be found in GitLab.

After the dataset was preprocessed, a manual analysis of some random dataset samples highlighted the following aspects:

1. The data contains some inconsistencies: in some dataset entries, table titles and table descriptions are included in the article text, while in some other entries, they are not.

2. In some articles, one or a few sentences are missing. This can be observed by comparing articles from the PubMed dataset with the corresponding original papers in the *PubMed* database, as in Table 50 in the Appendix.

[12] `https://pubmed.ncbi.nlm.nih.gov/`, last access: 10th January 2025.

3. Special characters like Greek letters or mathematical symbols, which are sometimes contained in the original papers, are missing in the dataset. In some cases, the absence of these symbols makes it hard to reconstruct the sense of the article.[13]

4. The dataset contains no capital letters.

Problems 1 to 3 are challenging to identify automatically, hence they cannot be solved on a large scale. Point 4 is not a problem in this work, since the PubMed dataset is not used for model training.

Enron Email Dataset

The Enron Email Dataset [Klimt and Yang, 2004] is an email dataset made public during the legal investigation on the Enron corporation. In this work, only 10 manually selected email threads from the dataset are used, since an analysis performed on larger data would go beyond the scope of this work. These 10 email threads are used in two variants.

For the first variant, emails are left as they are. This means that emails inside a thread are in descending chronological order, that they contain metadata, and, sometimes, indentation. This small dataset of unprocessed email is assumed to be particularly challenging to handle for LexRank, since these elements make the data noisy.

For the second variant, email threads were preprocessed as follows:

1. All metadata (e.g. "Message-ID", "Date", "From", "To") were removed.

2. A consistent separation sign was added between the single emails of a thread. Already existing separation signs (which were inconsistent) were deleted.

3. Emails in each thread were ordered in ascending chronological order.

4. Automatically generated email footers were removed.[14]

5. Special characters used to indent forwarded emails, such as ">", were removed.

Given the very small amount of data, these steps were performed manually. Table 51 in the Appendix shows an unprocessed email thread. Table 52 shows the same email thread, but preprocessed. This preprocessing operation has a significant impact of text length, making the average length of an email thread go from 418 to 227 words.

4.1.2 Self-collected datasets

No adequate datasets were found for the school material summarization scenario and the web snippet generation scenario. Therefore, the texts to be used in these two scenarios have been collected manually from several websites.

While some work on the summarization of educational material does exist (e.g. Shimada et al. [2017], or Yang et al. [2013]), no datasets in French or in German, the two

[13] Some examples, which cannot be reported here for length reasons, are dataset entries with IDs pubmed-736 and pubmed-740. They can be found in GitLab.

[14] Some examples of such footers were "This e-mail is the property of Enron Corp. and/or its relevant affiliate and may contain confidential and privileged material for the sole use of the intended recipient (s). [...]" and "Do You Yahoo!? Get email alerts & NEW webcam video instant messaging with Yahoo! Messenger".

languages needed for this scenario, are available. As for the web snippet generation scenario, it is the only one for which it was decided to use parallel texts; therefore, the challenge is to find such texts in all four required languages (English, German, French and Italian). A manual analysis of corpora like ParaCrawl [Bañón et al., 2020] or CCAligned [El-Kishky et al., 2020] showed that, even though these datasets do contain parallel sentences crawled from multilingual websites, these sentences often do not form longer texts, and are not always available in all four languages required in this work. For this reason, it was deemed easier to self-collect texts, rather than to search for adequate data in pre-existing datasets.

Text collection

Most of the 20 texts for the school material scenario (of which 10 are in German and 10 in French) have been collected from websites that are explicitly meant to provide educational material to school students. The few remaining texts have been collected from websites presenting knowledge to a broader, non-specific audience. The collected texts cover several school subjects: history, literature, philosophy, economic and social sciences, technology, geography, biology, and physics. They do not contain any technical terms, and they are suitable for school students aged 16 to 18. The average length of the texts is 688.8 words (44.6 sentences) in the German subset, and 634.4 words (35.5 sentences) in the French subset.

The parallel texts to be used in the web snippet generation scenario have been collected from a set of multilingual websites on different topics, among which, for example, one providing touristic information, one on public health and vaccinations, and one presenting instructions on how to write a eulogy. Depending on the language, the average length of the texts varies between 697 and 823 words, or 41.4 and 49.6 sentences.

Text preprocessing

The self-collected texts have been processed as follows:

- Line breaks were kept in the texts, since they create structure.
- Since they usually do not end with a full stop, which makes them problematic for LexRank, titles and subtitles were removed. This applies to all elements that semantically correspond to a title, even if they are not inside an <h> element in the HTML code of the website. For every removed title, an additional line break was added, in order to maintain the section structure.
- Tables and table captions, as well as figures, figure captions, ALT texts, and footnotes, were removed.
- Introductory paragraphs that were clearly separated from the rest of the text (if existing) were removed. An example of such a paragraph can be seen at in Figure 1 in the Appendix.
- Sections at the end of websites, whose content clearly did not belong to the main text on the website, were removed too. An example of such a section would be a *Related links* section.

- In texts for the school material scenario, exercise questions to the students (e.g. to assess text comprehension) were removed too.
- In order for all texts in the school material dataset to have comparable lengths, some of the school texts were truncated by removing semantically independent sections.[15]
- List item markers were included in the datasets, because they help identify a list as such and create structure among the list elements. Since they belong to the (CSS) style of a website, the symbols used as list item markers can differ from website to website. For consistency, in the two self-collected datasets created for this thesis, the same two list item markers (one for unordered lists, one for ordered lists) were used in all texts.[16]

4.2 Models and algorithms

4.2.1 LexRank

LexRank [Erkan and Radev, 2004] is a graph-based, extractive text summarization algorithm. It is based on the idea that the more important sentences in a text are those which are similar to many other sentences in the text. In order to compute similarity between sentences, these first have to be represented as vectors. Each sentence is represented as an N-dimensional vector, where N is the number of all possible words in the document language, and the value of each dimension in the vector is directly proportional to the number of occurrences of the corresponding word in the sentence. Then, the similarity between every two sentences is computed as the cosine similarity between their two vectors. Sentences and their similarities can be represented as a weighted graph in which sentences are nodes and their similarities are edges between them. In order to only keep meaningful similarities, a threshold is applied. At this point, the overall centrality of a sentence is computed through eigenvector centrality; relationships to high-scoring sentences contribute more to the score of the sentence in question that relationship to low-scoring sentences. Finally, the highest-ranking sentences are put together to create the summary of the document.

LexRank requires to specify the number of sentences that should be included in a summary. This was done case by case, depending on the requirements of the scenario, the language of the data and the length of the reference summaries (if available).

- For the news scenario, manual evaluation of LexRank summaries of different lengths showed that 2 sentences is the ideal summary length for English and French data, 3 sentences for German data.

[15] In GitLab, the truncated texts have following IDs: school-de-4, school-de-5, school-de-6, school-fr-2, school-fr-4, school-fr-5, school-fr-6, school-fr-7, school-fr-10.

[16] For unordered lists, bullet points (•) were used. They were chosen because they are the most commonly used list item marker in the collected texts, and also the most unambiguous, since other markers, such as hyphens, en dashes and em dashes, have some other specific uses in the English language (source: `https://www.merriam-webster.com/grammar/em-dash-en-dash-how-to-use`, last access: 10th January 2025). For ordered lists, cardinal numbers followed by points were used as item markers, since this was the most common way to signalize ordered list items in the collected texts.

- For the abstract generation scenario, 6-sentence summaries proved to be ideal for PubMed, while longer summaries were lacking coherence and not adding any important information. For TermITH, which has longer sentences than PubMed and shorter reference summaries, summary length was set to 5 sentences.
- For the snippet generation scenario, which has a strict requirement for very short summaries, summary length was set to 1 sentence for all languages.
- For school material summarization, summary length was computed dynamically, according to the length requirements of the scenario: the number of sentences to be included in a summary was set to 30% of the number of sentences in the original text.
- For email thread summarization, summary length was set to 3 sentences, since shorter LexRank summaries proved to be too short to provide the necessary context.

The LexRank implementation used in this work is the one provided by Python library sumy.[17] The execution scripts can be found in GitLab.

4.2.2 mLongT5

Pre-trained language model mLongT5 [Uthus et al., 2023] bases on LongT5 [Guo et al., 2022], which in turn bases on T5 [Raffel et al., 2020].

T5, short for Text-to-Text Transfer Transformer, is a text-to-text pre-trained language model, which means that it converts all text-based NLP problems to a format where both the input and the output are in the form of text. As Xue et al. [2021] observed, this approach is natural for tasks like ATS, but less usual for other tasks, such as text classification, where the output would usually be a class index. As its name indicates, T5 bases on encoder-encoder-decoder transformer architecture [Vaswani et al., 2017]; moreover, it is scalable, offering pre-trained sizes from 60 million to 11 billion parameters. T5 was pre-trained on C4, a large-scale dataset of web-crawled data, introduced in the same paper as the model itself. The pre-training was based on the span-corruption objective, which consists of feeding the model data in which consecutive spans of tokens are replaced with a mask token, and training it to reconstruct these sequences.

In LongT5 the attention mechanism of the original T5 encoder was modified to handle long inputs. Like T5, LongT5 was pre-trained on the C4 dataset. mLongT5 builds upon LongT5's architecture, but is pre-trained on the mC4 dataset, a multilingual version of C4 introduced by Xue et al. [2021]. Therefore, mLongT5 is suitable to handle long multilingual input, which makes it suitable for our scenarios, one of which (the abstract generation scenario) implies the summarization of long texts. Other multilingual models, such as mT5 [Xue et al., 2021], on the other hand, are not suitable to process inputs longer than 512 tokens, since memory consumption grows quadratically when input length increases [Guo et al., 2022].

This work uses mLongT5 fine-tuned on the sumstew dataset,[18] as it is available on Hugging Face.[19] Sumstew consists of a variable number of samples from several existing summarization datasets. An overview of the datasets contained in sumstew is given in Table 8.

[17] `https://pypi.org/project/sumy/`, last access: 11th January 2025.
[18] `https://huggingface.co/datasets/Joemgu/sumstew`, last access: 11th January 2025.
[19] `https://huggingface.co/Joemgu/mlong-t5-large-sumstew`, last access: 11th January 2025.

Most of them are news datasets, but lay summarization of scientific articles is also present (in PLOS and eLife, [Goldsack et al., 2022]), as well as patents (BIGPATENT, [Sharma et al., 2019]), legislative bills (BillSum, [Kornilova and Eidelman, 2019]), government reports (GovReport, [Huang et al., 2021]), narrative literature (BookSum, [Kryściński et al., 2022]), and encyclopedia articles (Klexikon, [Aumiller and Gertz, 2022]). With 175,692 rows in the train split, English is by far the most represented language in sumstew. Italian has 5,525 rows in the train split, French 2,843 rows, German 1,901, Spanish 1,260. It should be noted that Italian, French and Spanish are only represented in news datasets, while German is almost completely represented in an encyclopedia dataset.

In this work, fine-tuned mLongT5 was executed in a Docker container. Docker files and execution scripts can be found in GitLab. The code was run on a machine with two NVIDIA RTX 2080 Ti GPUs with 11 GB of memory each.

Table 8: Composition of the train split of the sumstew dataset.

Dataset name	**Language**	**Number of rows in sumstew**
Multi-news [Fabbri et al., 2019]	EN	43,084
CNN/DailyMail [Nallapati et al., 2016]	EN	39,243
BIGPATENT [Sharma et al., 2019]	EN	34,589
PLOS [Goldsack et al., 2022]	EN	20,977
BillSum [Kornilova and Eidelman, 2019]	EN	15,036
GovReport [Huang et al., 2021]	EN	10,811
BookSum [Kryściński et al., 2022]	EN	7,994
Fanpage [Landro et al., 2022]	IT	5,525
MLSUM-fr [Scialom et al., 2020]	FR	2,592
eLife [Goldsack et al., 2022]	EN	2,533
Klexikon [Aumiller and Gertz, 2022]	DE	1,896
XL-Sum-en [Hasan et al., 2021]	EN	1,343
XL-Sum-es [Hasan et al., 2021]	ES	1,161
XL-Sum-fr [Hasan et al., 2021]	FR	251
MLSUM-es [Scialom et al., 2020]	ES	99
DialogSum [Chen et al., 2021]	EN	82
MLSUM-de [Scialom et al., 2020]	DE	5

4.2.3 Mistral NeMo

Mistral NeMo [Mistral, 2024] is a LLM based on the transformer architecture. It has 12 billion (12B) parameters, 40 transformer layers and a context window of up to 128k tokens, meaning that, when generating a response, it can consider up to 128 thousands tokens at once. Given its number of parameters, it is a relatively small model – GPT-3 [Brown et al., 2020], for example, has 175B parameters. However, some self-conducted experiments showed that NeMo suits well the scenarios presented in this work.

According to the blog post in which its release is announced, NeMo is "designed for global, multilingual applications".[20] It is trained both to generate code and to converse in natural languages.

[20] `https://mistral.ai/news/mistral-nemo/`, last access: 11th January 2025.

In this work, the instruction-tuned version of NeMo[21] was run in a Docker container through Ollama,[22] an open-source tool designed to set up and execute LLMs locally. The code was run on the same machine described in 4.2.2. In this case too, the execution scripts can be found in GitLab. The prompts used for each scenario and each language can be found in Table 53 in the Appendix. It is important to note that prompts were kept unspecific: the type of input text (e.g. news article, or scientific paper) was not mentioned, and no information on the purpose of the summary, or on its desired tone, was provided. This was done in order to keep the comparison with LexRank and mLongT5 fair, since these two methods cannot be given any additional information on the type of input text or on what the summary will be used for.

[21] Available of Hugging face at `https://mistral.ai/news/mistral-nemo/` (last access: 11th January 2025), or on Ollama at `https://ollama.com/library/mistral-nemo` (last access: 11th January 2025).

[22] Available for download at `https://ollama.com/`, last access: 25th January 2025.

Chapter 5

Evaluation

In this chapter, Section 5.1 presents the evaluation metrics used in this work, while Sections 5.2 and 5.3 respectively present quantitative and qualitative evaluation results.

5.1 Evaluation metrics

5.1.1 Automatic quantitative evaluation

Content overlap

As an automatic metric to measure content overlap between reference summary and candidate summary, ROUGE [Lin, 2004] was chosen. Even though the metric has several known problems, some of which were described in Section 2.2.1, it still is by far the most used automatic metric in ATS, which makes it useful for comparability. BERTScore [Zhang et al., 2019], which was also described in Section 2.2.1, was also taken into consideration and used to evaluate some samples of summaries. However, its result did not correlate with the human evaluation results of the same summaries. Moreover, BERTScore's computation times are significantly higher than ROUGE's. For these reasons, ROUGE was chosen over BERTScore.

Length metrics

Three automatic metrics used in this work are related to summary length: word count, sentence count and compression ratio. These will be used to proof whether the generated summaries respect the length constraints specified in the scenario definitions.

Word count and sentence count are computed through spaCy,[1] a Python NLP library offering, among others, language-specific tokenization and parsing. Since LexRank uses another system for sentence segmentation, there are some discrepancies between the target summary length specified when running the LexRank algorithm, and the summary length computed during automatic evaluation. In Table 9, for example, CNN/DailyMail LexRank summaries have an average sentence count of 2.13 sentences, even though the specified target length was always 2 sentences.

[1] `https://spacy.io/`, last access: 28th January 2025.

Compression ratio is computed through its formula from Bommasani and Cardie [2020]. Given a text T and its candidate summary S:

$$CompessionRatio(T, S) = 1 - \frac{|S|}{|T|},$$

where $|T|$ and $|S|$ are respectively T's and S's lengths in words.

Extractivity

Extractivity measures how extractive (or abstractive) a summary is. The formula used in this work was implemented by Grusky et al. [2018], who call it density. Given a text T, its candidate summary S and the set of extractive fragments used in the summary $F(T, S)$, extractivity is computed as follows:

$$Extractivity(T, S) = \frac{1}{|S|} \sum_{f \in F(T,S)} |f|^2,$$

where $|f|$ is the length of an extractive fragment f. Therefore, the longer the extractive fragments in a summary are, the higher its extractivity score is. In this work, as in Grusky et al. [2018],

- if $Extractivity(T, S) \leq 1.5$, the summary is considered abstractive;
- if $1.5 < Extractivity(T, S) < 8.2$, the summary is considered mixed, meaning that it contains both extracted segments and abstractive parts;
- if $Extractivity(T, S) \geq 8.2$, the summary is considered extractive.

Truncated summaries

This metric consists of a boolean value indicating whether a summary ends with an incomplete sentence. It was implemented because it was observed that defining a maximal length for mLongT5 sometimes led summaries to be cut off in the middle of a sentence. It has a trivial implementation: a summary is marked as truncated if it does not end with one of the following (sequences of) characters: point, question mark, exclamation mark, point followed by quotes, question mark followed by quotes, exclamation mark followed by quotes, point followed by a closing bracket.

This metric was computed for summaries generated by mLongT5 or by NeMo. Then, manual evaluation was conducted on the summaries that had automatically been identified as truncated, in order to correct the result, if necessary. This was done because it was observed that there are cases of summaries that are not truncated even if they do not end with one of the above-mentioned characters (an example would be a summary consisting of bullet points, since these do not need a point at the end).

For summaries generated by LexRank, the boolean value was always set to false, since LexRank does not generate new sentences, and therefore, these sentences cannot be incomplete. The only exception would be if sentences are also incomplete in the original text, in which case it would be unfair to penalize LexRank.

5.1.2 Qualitative evaluation

Five questions, Q1 to Q5, were developed for qualitative evaluation. The first three questions will be presented in the following subsections, while Q4 and Q5 do not need to be described here, since they are respectively about the summary's style and use, which have already been introduced in Section 2.3 and defined for every scenario in Chapter 3.

For each questions, scores from 1 to 5 can be assigned. If 1 is assigned, the summary does not fulfill the requirements formulated in the question at all; if 2 is assigned, the summary is poor with regard to the question requirements; 3, 4 and 5 respectively mean that the summary is acceptable, good and very good with regard to the question requirements. When assigning scores to the summaries, the evaluation criteria described in Chapter 3 (e.g. the importance given to the summary's form and, in opposition, to its content) were taken into account.

The scores for the five manual evaluation metrics were computed independently from each other. For example, if a summary contained factual mistakes (evaluated in Q3), this did not have an impact on the score for the summary's style (evaluated in Q4). Moreover, since Q1–Q5 focus on different, but equally important aspects of a summary, scores cannot compensate each other: a summary is considered qualitatively good only if all of its Q1–Q5 scores are good.

In addition to Q1–Q5, two more metrics were computed manually and are also described in the following subsections: lead bias and information dispersion. These are inspired by the position and the dispersion metrics proposed by Goel et al. [2021] and described in Section 2.2.1.

Qualitative evaluation was performed manually. For three datasets from the news scenario (one per language) and for both datasets from the abstract generation scenario, qualitative evaluation was performed on 1% of the generated summaries. Summaries to evaluate were extracted randomly. For the remaining scenarios, which only contain a small number of texts, qualitative evaluation was performed on every summary.

Q1: grammar and lexicon

This metric is partially bases on the grammaticality criterion used in DUC 2006 [Dang, 2005] and described in Section 2.2.2. However, Q1 in this work also considers spelling and lexical choices, which are not explicitly mentioned in the evaluation criteria of DUC 2006.

For LexRank summaries, Q1 scores were always set to 5, since LexRank does not write new sentences but only extracts sentences from the original text, which is assumed to be grammatically and lexically correct. There are indeed some rare cases of original texts containing grammatical mistakes, or, more often, containing noise elements such as paragraph titles not followed by full stops (cf. Section 4.1.1). This is a data cleanliness problem, and not a problem linked to the LexRank algorithm; therefore, even if grammatically incorrect or noisy sentences appeared in LexRank summaries, no points were subtracted from the Q1 score. The only exception to this rule was made in the scenario with unprocessed emails threads. Said scenario was designed with the intention to evaluate how different ATS systems manage texts containing metadata and other noise elements, and to compare the results with those from the preprocessed email dataset. Therefore, for LexRank too, points were subtracted if summaries contained noise elements.

Q2: structure and coherence

This metric is also inspired by the homonymous metric from DUC 2006, but it extends it. Its goal is to evaluate whether the text is well structured, and whether the sentences in a summary make sense both on their own and in combination with the other sentences. In order to evaluate the latter, both cohesion (i.e., whether the sentences are linked on a syntactical level) and coherence (i.e., whether there is a conceptual link between the pieces of information presented in different sentences) are taken into account.

This metric also includes three other evaluation criteria from DUC 2006: referential clarity, focus and non-redundancy (cf. Section 2.2.2). Referential clarity is relevant because a text containing pronouns or phrases without a referent cannot be coherent. Focus is important because, in order for a text to be coherent, its sentences should present information that is related to the rest of the text. Finally, non-redundancy is relevant in the context of structure: a well-structured summary should not be redundant or repetitive.

Q3: factual correctness

Factual correctness means that the information presented in a summary should be in line with the content of the original text. In this work, we distinguish between three types of factual incorrectness:

- If facts from the original text are reported incorrectly, we talk about factual mistakes. Incorrectly reported names, dates, numbers or quotes also fall into this category.
- If a summary presents information that is not contained in the original text, but that is not factually incorrect in the greater scheme of things, we talk about information addition. In most cases, this phenomenon is unwanted, because it might mislead the reader about the content of the original text. However, in some scenarios, minor information additions can be acceptable: two examples have been provided at the end of Sections 3.1 and 3.3.
- If a summary presents information that is not contained in the original text, and that is factually incorrect, we talk about hallucination.

Like Q1 scores, Q3 scores too were always set to 5 for LexRank summaries: since LexRank extracts sentences from the original text without changing them, factual mistakes cannot happen. However, there are cases in which structure and coherence problems in LexRank summaries result in misleading text content – an example is shown in the section on manual evaluation results (Table 19). In these cases, points were subtracted for Q2, but not for Q3.

Lead bias

This metric, which consists of a boolean value, indicates whether a summary only contains information from the very beginning of the original text. Lead bias, a common phenomenon in news writing, means that the most important information is given at the beginning of the article. Since most ATS datasets are news datasets, ATS models are often trained with data containing lead bias. As a consequence, they might also tend to leverage this bias in texts in which the most important information is not at the beginning, leading

to bad-quality summaries. This is a well-known phenomenon in the field of ATS (cf. i.a. Jung et al. [2019] or Zhu et al. [2021]).

In order to compute whether a summary leverages lead bias, summary sentences have to be matched to original text sentences with the same (or similar) content. For abstractive summaries, this is challenging to do automatically, since there is no 1-to-1 correspondence between summary sentences and original text sentences.[2] For this reason, summaries were manually evaluated for a lead bias. Lead bias was set to true if all summary sentences only contained information presented at the beginning of the original text.

Information dispersion

This metric (suggested by Goel et al. [2021]) measures whether a summary only contains information from a particular part of the original text, or whether it extracts content that is broadly distributed in the original text. In this work, it can be assigned three values: low, medium, or high. Like the lead bias metric, it was computed manually.

This metric might be used to predict if a summary has good information coverage: it can be expected that summaries with high information dispersion also have higher information coverage.

5.2 Quantitative evaluation results

5.2.1 News scenario

ROUGE scores

For summaries from the same datasets, ROUGE F-measure scores are quite similar across the three different summarization methods, with a light trend of LexRank having slightly lower ROUGE-1 scores than mLongT5 and NeMo (cf. Table 9). What is more striking is that different datasets of the same language have noticeably different ROUGE scores: all three summarization methods achieve significantly better results on CNN/DailyMail than on Newsroom, better results on 20 Minuten than on MLSUM-de, and slightly but consistently better results on OrangeSum than on MLSUM-fr. However, these apparent quality differences could not be confirmed in manual evaluation (cf. Section 5.3.1): for each summarization system, manual evaluation scores tend to be similar in the two datasets of the same language.

Therefore, differences in ROUGE scores are probably related to the ROUGE metric itself, and not to the quality of the summaries. In particular, by comparing Table 9 with Table 2 from Chapter 4, it can be observed that ROUGE scores tend to be better when reference summaries are longer: in CNN/DailyMail, reference summaries are nearly twice as long as in Newsroom; in 20 Minuten, they are significantly longer than in MLSUM-de, and in OrangeSum, they are slightly longer than in MLSUM-fr. Moreover, by comparing candidate summary lengths from Table 9 with reference summary lengths from Table 2, it can be observed that, for all news datasets, candidate summaries have a higher word count than reference summaries. This means that candidate summaries probably cover

[2] Python library Sentence Transformers (`https://huggingface.co/sentence-transformers`, last access: 29th January 2025) was tested for this purpose. It offers several methods for similarity computation: cosine similarity, dot product, negative euclidean distance, negative manhattan distance. However, none of these methods produced results that corresponded to human judgment.

Table 9: Average quantitative evaluation scores for news scenario. R-1, R-2 and R-L stand for ROUGE-1, ROUGE-2 and ROUGE-L F-measure scores. The abbreviations in the remaining columns stand for compression ratio, word count, sentence count, extractivity, and truncated summaries.

Dataset	Method	R-1	R-2	R-L	Com. r.	Word c.	Sent. c.	Extr.	Trunc.
CNN/ DailyMail	LexRank	31.76	9.87	19.84	0.91	50.46	2.13	33.46	0/2500
	mLongT5	37.48	10.03	21.1	0.88	73.31	4.73	2.04	8/2500
	Nemo	36.1	11.08	21.76	0.89	65.99	2.97	2.24	0/2500
Newsroom	LexRank	19.05	3.05	12.76	0.91	54.79	2.28	38.71	0/2500
	mLongT5	20.32	3.59	12.26	0.87	91.07	4.28	1.9	126/2500
	Nemo	21.18	4.55	13.79	0.89	81.59	3.88	1.78	0/2500
20 Minuten	LexRank	24.81	6.98	15.78	0.85	49.53	3.26	28.12	0/2500
	mLongT5	27.18	4.65	14.99	0.8	68.49	6.17	1.23	4/2500
	Nemo	29.11	9.18	18.93	0.82	60.82	4.06	2.48	6/2500
MLSUM-de	LexRank	16.76	3.75	11.29	0.88	52.75	3.47	29.97	0/2500
	mLongT5	19.1	3.01	11.24	0.85	70.93	6.26	1.26	30/2500
	Nemo	20.79	5.93	13.98	0.87	62.24	4.01	2.91	7/2500
OrangeSum	LexRank	27.08	7.86	16.52	0.85	61.4	2.41	43.51	0/2500
	mLongT5	31.2	7.88	17.36	0.85	59.8	3.33	1.78	12/2500
	Nemo	29.57	10.33	17.69	0.79	88.24	3.98	3.56	43/2500
MLSUM-fr	LexRank	22.72	6.08	14.55	0.88	66.5	3.6	45.64	0/2500
	mLongT5	27.37	6.82	16.02	0.89	62.4	3.37	1.61	10/2500
	Nemo	24.29	7.72	15.26	0.84	93.32	4.36	2.73	48/2500

Table 10: Length distribution (in words) of summaries in news scenario, per dataset and per method.

Dataset	Method	Number of summaries per length category (in words)				
		< 25 w.	25-50 w.	51-75 w.	75-100 w.	> 100 w.
CNN/DailyMail	LexRank	115	1298	865	182	40
	mLongT5	0	10	1636	810	44
	Nemo	9	582	1410	401	98
Newsroom	LexRank	213	1099	865	232	91
	mLongT5	0	30	603	1208	659
	Nemo	9	482	1168	485	356
20 Minuten	LexRank	57	1380	921	126	16
	mLongT5	0	18	2063	409	10
	Nemo	62	972	922	358	186
MLSUM-de	LexRank	58	1195	988	223	36
	mLongT5	1	21	1793	656	29
	Nemo	30	1031	990	243	206
OrangeSum	LexRank	74	831	973	465	157
	mLongT5	1	239	2198	62	0
	Nemo	2	127	813	858	700
MLSUM-fr	LexRank	77	672	946	541	264
	mLongT5	0	140	2201	159	0
	Nemo	5	154	739	846	756

more information than reference summaries. Therefore, having longer reference summaries probably increases the chance that the information covered in a candidate summary is also mentioned in the corresponding reference summary, which leads to higher lexical overlap and therefore to a higher ROUGE score.

When observing Newsroom ROUGE scores, it also has to be kept in mind that they might be skewed by the fact that some of the articles in the dataset are missing their initial sentence(s) or, more rarely, their last paragraphs, as mentioned in Section 4.1.1. This means that ATS systems have less information to produce good summaries, which might reflect on ROUGE scores.

Summary length and truncated summaries

In Chapter 3, it was defined that summary length for the news scenario can vary from a minimum of 20-25 words to a maximum of 85-90 words.

LexRank summaries always respects the length requirements. For almost every dataset, they are shorter than mLongT5's and NeMo's summaries, and therefore they have higher compression ratio.

On average, mLongT5 summaries respect the length requirements for every dataset except for Newsroom, for which they are too long. Newsroom is the dataset with the longest articles by far (cf. Table 2), so it is not surprising that mLongT5 summaries for Newsroom texts are longer than for other datasets. Moreover, 126 out of 2500 mLongT5 Newsroom summaries were truncated by the model itself, probably because they exceeded the provided maximum length parameter.

It is also interesting to notice that mLongT5-generated Newsroom summaries, although being longer, have a slightly lower sentence count than mLongT5-generated CNN/Daily-Mail summaries, indicating that sentences are, on average, longer in the former than in the latter. In the two French datasets, word and sentence count for mLongT5 summaries are similar. The same applies to the two German datasets, where sentences in mLongT5 summaries tend to be very short (cf. Table 9: $68.49/6.17 = 11.1$ words per sentence in 20 Minuten, and $70.93/6.26 = 11.33$ words per sentence in MLSUM-de). This last aspect will be deepened in manual evaluation (cf. Section 5.3.1).

NeMo summaries exceed the length requirement for MLSUM-fr, and almost do for OrangeSum. Unlike mLongT5, NeMo was not given a maximum length parameter, since target summary length was defined through natural language in the summarization prompt (all prompts used for NeMo are reported in Table 53 in the Appendix). Still, NeMo automatically truncated some of its summaries (43 for OrangeSum, and 48 for MLSUM-fr, cf. Table 9). In slightly more than half of the cases, this happened when summaries exceeded the length requirement specified in the prompt; however, cases of short truncated summaries (less than 50 words) were also observed.

Table 10 shows length distribution for the news summaries produced by the three systems. It can be observed that, even in cases where average summary length respects the scenario requirements, there are a considerable number of summaries that do not. NeMo's summaries of OrangeSum texts, for example, have an average length of 88.24 words, which is still acceptable, but Table 10 shows that 700 of these summaries are even longer than 100 words. Moreover, 356 NeMo-generated Newsroom summaries are also longer than 100 words.

Several NeMo summaries also seem to have been truncated after one or more words

in the wrong language had been generated. Truncated NeMo summaries containing foreign languages at the end were mostly observed in the French datasets, but a few cases were found in German datasets too. In particular, of 46 OrangeSum NeMo summaries automatically identified as truncated, 29 were completely in French, while 17 contained a foreign language at the end. Figures are similar for MLSUM-fr.

Extractivity

Since LexRank is an extractive algorithm, all its summaries have high extractivity scores (cf. Table 9).

For all datasets, mLongT5's summaries have relatively low average extractivity values. For the two German datasets, values are even under 1.5, which makes the summaries fall into category "abstractive" according to Grusky et al. [2018] (cf. 5.1.1). However, manual evaluation has shown that mLongT5 summaries often contain spelling, grammar and factual mistakes, as well as minor reformulations that are still very close to the original text and that cannot be considered proof of good abstraction capabilities. Both phenomena are shown in Table 11. Through these mistakes and reformulations, extractive fragments in the summaries appear to be shorter, which leads to a lower extractivity score. However, if grammar, spelling and factual mistakes are corrected, the extractivity score increases. For the aforementioned summary, for example, correcting the mistakes highlighted in red makes extractivity go from 1.92 to 3.68.

Table 11: Example of mLongT5 summary containing several spelling, grammar and factual mistakes, as well as minor reformulations. The former are highlighted in red (while the corresponding correct information in the original text is highlighted in green), the latter in blue (both in summary and in original text).

Text ID	newsroom-460 (in GitLab)
Summary	If you're a flight attendant, it's hard to get into an airline's lounge. And if you're one of those passengers who can't enjoy flying as usual, the New York Times has some good news for you. The president of the Animal Welfare International Fund tells the Times that he was on a flight to Portugal when his plane decelered so suddenly that everything that wasn't strapped up went flying through air. A man next to him lost his Scotch glass and the plane got pitch black. One pilot came out of cockpit with flashlight. He figured this could't be good-he moved over to the windows and looked at the engine for a long time, reports the Times. "You could almost see them thinking, 'Well, we might turn around, But there's nothing in Gambia. We'd be stuck where, so we possibly as well try again and make Portugal.' Apparently
Article	[...] Fred O'Regan, president of the International Fund for Animal Welfare, with Quincy, a golden retriever puppy. [...] But about 10 minutes into the flight, the plane decelerated so suddenly that anything that wasn't strapped down went flying through the air. The man next to me lost his glass of Scotch. And the plane went pitch black. One of the pilots came out of the cockpit with a flashlight. I figured this couldn't be good. He moved over to one of the windows and examined the engine for what seemed like a long time. You could almost see him thinking, "Well, we could turn around, but there's nothing in Guinea-Bissau. We'd be stuck there, so we might as well try and make Portugal." Apparently, he concluded that we could make it to Portugal. [...]

All NeMo summaries fall into category "mixed" ($1.5 <$ extractivity score < 8.2), mean-

ing that they contain both extracted phrases and abstractive parts. English NeMo summaries tend to be slightly more abstractive than French and German summaries.

Table 12: Average quantitative evaluation scores for abstract generation scenario.

Dataset	Method	R-1	R-2	R-L	Com. r.	Word c.	Sent. c.	Extr.	Trunc.
PubMed	LexRank	43.3	15.68	21.99	0.9	218.82	6.7	61.83	0/1000
	mLongT5	38.52	9.27	18.87	0.93	140.67	6.68	1.97	56/1000
	Nemo	38.48	11.3	18.86	0.9	205.4	13.68	2.54	0/1000
TermITH	LexRank	34.23	9.74	16.46	0.95	208.17	9.04	67.21	0/779
	mLongT5	36.8	8.32	17.09	0.97	107.56	5.36	1.88	14/779
	Nemo	30.61	8.41	14.55	0.93	262.6	12.15	7.43	174/779

5.2.2 Abstract generation scenario

For PubMed, LexRank has higher ROUGE scores than mLongT5 and NeMo, which does not correspond to what was found during manual evaluation (cf. Section 5.3.2 and Table 34). The same applies to the differences between the TermITH ROUGE scores of the three systems.

For all three summarization systems, ROUGE scores for English data (PubMed) are better than for French data (TermITH) (cf. Table 12). This is in line with the results of manual evaluation (cf. Table 34).

Accepted summary length, as defined in Section 3.2, can vary from approximately 115 to approximately 200 words. For TermITH, the average length of mLongT5 summaries is slightly lower than 115 words, which is a sign that these summaries might have insufficient information coverage. LexRank summaries and NeMo PubMed summaries, on the other hand, are slightly too long with regard to the scenario requirements. Moreover, NeMo TermITH summaries are significantly too long, with an average length of 262.6 words and 174 truncated summaries (of which almost 75% are longer than 200 words). This confirms NeMo's tendency to generate overly long summaries in French, which was already observed in the news scenario (cf. Table 10). Moreover, when analyzing the aforementioned 174 truncated summaries manually, several summaries partially or completely written in English or in other foreign languages were found, even though the original texts were in French, being part of TermITH.

Based on their average extractivity scores (cf. Table 12) mLongT5's summaries appear to be more abstractive than NeMo's summaries. However, what was observed in 5.2.1 also applies to the current scenario: mLongT5-generated summaries would have higher extractivity scores if they did not contain grammar of factual mistakes. As in the news scenario, NeMo's English summaries are more abstractive than its French summaries.

5.2.3 Web snippet generation scenario

As mentioned in the scenario definition (cf. Section 3.3), summaries generated for this scenario should by no means be longer than 160 characters, since this is approximately the length after which snippets are cut off on SERPs. To simulate this use case, all generated summaries longer than 160 characters were truncated and terminated by three suspension

Table 13: Average quantitative evaluation scores for web snippet generation scenario. Compression ratio, word count, sentence count and extractivity were computed **before** truncating summaries longer than 160 characters. The last table column (*> 160 char.*) shows how many summaries exceeded the length limitation and therefore had to be truncated.

Dataset	Method	Comp. ratio	Word c.	Sentence c.	Extr.	> 160 char.
EN	LexRank	0.96	26.8	1.0	22.94	2/10
	mLongT5	0.89	64.4	3.9	1.98	10/10
	Nemo	0.93	41.2	2.7	1.62	8/10
DE	LexRank	0.97	13.9	1.1	17.2	1/10
	mLongT5	0.88	66.1	5.8	1.54	10/10
	Nemo	0.89	69.3	4.4	3.1	10/10
FR	LexRank	0.96	25.6	1.0	26.8	2/10
	mLongT5	0.92	51.8	2.8	1.71	9/10
	Nemo	0.89	83.1	4.1	2.93	10/10
IT	LexRank	0.96	28.5	1.1	23.69	2/10
	mLongT5	0.9	62.8	3.5	1.66	10/10
	Nemo	0.89	89.8	4.9	4.96	10/10

points. Compression ratio, word count, sentence count and extractivity were computed before this operation, in order to have data on which summarization systems were able to fulfill the length requirements of the scenario. Qualitative evaluation (discussed in Section 5.3.3), on the other hand, was conducted on the truncated texts, since in a real-world use of the summaries as web snippets only their first 160 characters would be visible.

As can be seen in Table 13, LexRank was the only summarization system that mostly produced summaries that respect the length requirements of the scenario. mLongT5 and NeMo, on the contrary, generated summaries that were almost always longer than 160 characters. As already observed in the news and in the abstract generation scenario, NeMo tends to generate very long summaries in French, with an average summary length of 83.1 words or 4.1 sentences, which clearly exceeds the target length of two short sentences specified in the prompt (cf. Table 53 in the Appendix). Changing to a prompt that asked to summarize the text in one sentence, or in 25 words maximum, also had little effect: results were, on average, not shorter than with the two-sentence-prompt.

The same length problem also applies to Italian NeMo summaries (with an average summary length of 89.8 words). In English, on the other hand, the length of NeMo summaries is 41.2 words, which might be a sign that NeMo has a better ability to condense content in English than in other languages. This was also observed in manual evaluation (cf. Section 5.3.3). Moreover, English NeMo summaries have a significantly lower extractivity value than in other languages.

5.2.4 School material summarization scenario

The length requirement for this scenario, as defined in Section 3.4, is that summary length should be between 20% and 50% of the length of the original text. This corresponds to a compression ratio between 0.80 and 0.50. As can be seen in Table 14, the compression ratio for mLongT5 summaries is higher than desired both for the German and the French

Table 14: Average quantitative evaluation scores for school material summarization scenario.

Dataset	Method	Compr. ratio	Word c.	Sentence c.	Extractivity	Truncated
DE	LexRank	0.66	234.1	15.0	33.23	0/10
	mLongT5	0.85	101.8	9.8	1.25	0/10
	Nemo	0.72	185.5	11.2	4.34	0/10
FR	LexRank	0.64	227.8	12.1	37.2	0/10
	mLongT5	0.87	82.8	4.8	1.18	0/10
	Nemo	0.64	227.2	10.3	2.96	3/10

Table 15: Average quantitative evaluation scores for email thread summarization scenario.

Dataset	Method	Compr. ratio	Word c.	Sentence c.	Extractivity	Truncated
Preprocessed	LexRank	0.77	48.6	3.3	12.2	0/10
	mLongT5	0.75	50.6	3.0	1.9	1/10
	Nemo	0.83	36.0	1.3	1.35	0/10
Unprocessed	LexRank	0.82	75.0	3.9	11.24	0/10
	mLongT5	0.85	58.8	2.9	1.83	1/10
	Nemo	0.93	28.1	1.0	1.34	0/10

dataset. Therefore, it is likely that mLongT5-generated summaries for this scenario will have insufficient information coverage.

For LexRank and NeMo, average summary lengths are within the desired range. However, three NeMo summaries for the French data were truncated by the model itself (even if they were not excessively long).

In this scenario, German NeMo summaries are more extractive than in the news scenario (cf. Table 9) and in the web snippet generation scenario (cf. Table 13). They are also more extractive than French NeMo summaries in this same scenario.

For mLongT5, what has been written about extractivity in Section 5.2.1 also applies to this case.

5.2.5 Email thread summarization scenario

Summary length is in the acceptable range for all summarization systems. For the unprocessed scenario, LexRank summaries appear to have higher word and sentence counts because they sometimes contain metadata and noise.

For mLongT5, extractivity values are in the usual range. For NeMo, they are slightly lower than in the other scenarios, meaning that summaries are more abstractive. This might be linked to the fact that this scenario requires more reformulation effort than the other ones, since email threads are written in the first and second person, but will probably be summarized in the third person.

Table 16: Average manual evaluation scores for news scenario.

Dataset	Method	Q1	Q2	Q3	Q4	Q5	Lead bias	Information dispersion (low, medium, high)
Newsroom (25 samples)	LexRank	5.0	3.04	5.0	3.0	2.88	3/25	8, 17, 0
	mLongT5	3.32	3.36	2.48	3.6	3.4	2/25	5, 15, 5
	Nemo	5.0	4.96	4.64	5.0	4.8	0/25	0, 8, 17
CNN/DailyMail (3 samples)	LexRank	5.0	2.33	5.0	3.67	2.33	0/3	0, 3, 0
	mLongT5	3.33	2.67	2.67	4.0	3.67	0/3	0, 3, 0
	Nemo	5.0	5.0	5.0	5.0	5.0	0/3	0, 2, 1
20 Minuten (25 samples)	LexRank	5.0	2.76	5.0	3.44	3.24	2/25	9, 15, 1
	mLongT5	3.16	2.76	2.16	3.28	2.76	0/25	0, 22, 3
	Nemo	4.64	4.88	4.72	4.96	4.76	1/25	1, 15, 9
MLSUM-de (3 samples)	LexRank	5.0	2.0	5.0	3.33	3.0	0/3	1, 2, 0
	mLongT5	3.0	2.67	1.33	3.0	2.0	0/3	0, 3, 0
	Nemo	4.33	5.0	4.0	5.0	5.0	1/3	1, 2, 0
OrangeSum (25 samples)	LexRank	5.0	2.48	5.0	3.0	2.64	5/25	8, 17, 0
	mLongT5	3.36	3.48	2.64	3.76	3.6	1/25	1, 23, 1
	Nemo	4.8	4.64	4.68	4.96	4.52	0/25	0, 11, 14
MLSUM-fr (3 samples)	LexRank	5.0	2.0	5.0	3.33	2.33	0/3	1, 2, 0
	mLongT5	2.67	3.0	3.67	4.67	4.67	1/3	1, 2, 0
	Nemo	5.0	5.0	4.33	5.0	5.0	0/3	0, 2, 1

5.3 Qualitative evaluation results

5.3.1 News scenario

25 texts each were randomly extracted from Newsroom (for English data), 20 Minuten (for German data) and OrangeSum (for French data). Every summary produced for these texts (one generated by LexRank, one by mLongT5, one by NeMo) was manually evaluated. Moreover, three texts each were randomly extracted from the three remaining news datasets (CNN/DailyMail, MLSUM-de, MLSUM-fr). For these texts too, the summaries produced by the three ATS systems were manually evaluated, in order to find out if the tendencies observed in the first group of datasets are confirmed in this second group of datasets. So, for each language, manual evaluation was performed on a total of 84 summaries of 28 texts. The average scores for each ATS system and each dataset can be seen in Table 16.

Of the 25 random samples extracted from Newsroom, five contained articles missing the first paragraph(s), which is a known problem already discussed in Section 4.1.1.[3] It would be difficult to perform a fair qualitative evaluation of the generated summaries for these samples, and it would also jeopardize comparability with qualitative evaluation results of other datasets. For these reasons, said samples were not evaluated; instead, they were replaced by new random samples.

[3]In GitLab, the IDs of these five samples are: newsroom-1155, newsroom-1774, newsroom-1855, newsroom-2370, newsroom-2498.

LexRank results

As already mentioned in 5.1.2, Q1 and Q3 always score 5 for LexRank, since the algorithm does not create new sentences, meaning that its summaries can neither be grammatically nor factually incorrect, if we consider the original text to be grammatical and factually correct. For this reason, LexRank's Q1 and Q3 scores will not be discussed in this or in the following paragraphs. Moreover, since the LexRank algorithm is language-independent (with the exception of the part responsible for sentence segmentation), and since its scores are similar across the three scenario languages, its results are discussed here without language distinctions.

Depending on the dataset, the average Q2 scores for LexRank-generated news summaries are between 2 and 3, meaning that, on average, the summaries have insufficient structure or coherence – only for Newsroom summaries, the average Q2 score is slightly over 3 (3.04, as can be seen in Table 16). By far the biggest problems concerning structure and coherence is lack of context. Many summaries (in all languages and datasets) contain pronouns or other elements without a referent. One example can be seen in the summary in Table 17, where it is unknown to what pronoun "they" refers. Moreover, the two summary sentences are not linked (neither syntactically nor semantically), which is another problems that was observed several times, although not as often as the previously mentioned one. Because of these issues, the summary was assigned a Q2 score of 2.

Table 17: Example of LexRank summary containing a pronoun without a referent (highlighted in red) and unrelated sentences.

Text ID	newsroom-1838 (in GitLab)
Summary	And then they announced their lawsuit on March 16, 1970, inspiring the headline "Newshens Sue Newsweek for Equal Rights" in the New York Daily News, which went out of its way to note that most of the plaintiffs were young, "and most of them pretty." In a March 18, 2010 cover story "Are We There Yet" written by three Newsweek women, the authors said 39 percent of the people on the masthead were women, up from 25% in 1970.

Coherence problems also arise when quotes are only partially included in the summaries, meaning that either the opening or the closing quotation marks are missing (Table 18).

Table 18: Example of LexRank summary containing an incomplete quote, where quotation marks (highlighted in red) are closed, even though they have never been opened.

Text ID	orangesum-173 (in GitLab)
Summary	Moi, j'ai été victime de crachats, de coups d'épaule, d'insultes. On m'insulte de menteuse, on me dit que je vais le payer (...) Je suis plus choquée par ce regain de violence à mon encontre que par sa libération en fait".

Incorrect sentence segmentation also led to point subtraction in Q2, even though it was only observed rarely. Incorrect sentence segmentation happens when LexRank interprets

a point as a full stop signaling the end of a sentence, even though this is not its meaning in the original text. In Table 54 in the Appendix, for example, the point after the name of covid variant "B.1.1.7." is interpreted as a full stop, even if it is not. However, cases like this are rare: the LexRank implementation used in this work leverages language-dependent lists of abbreviations or other words containing dots, and some more abbreviations were added manually in the execution scripts that can be found in GitLab. Therefore, in most cases, sentence segmentation is correct even if a sentence contains a dot.

As already mentioned in Section 5.1.2, structure and coherence problems can also result in factual correctness problems, which happens several times in the analyzed samples. An example of this phenomenon can be seen in Table 19: the first sentence in the summary mentions airline fares, and the second one presents the prices to move between several places. Therefore, the reader could assume that the prices in the second sentence are airline prices. However, in the original text, the two sentences are not related, and the second one is about bus prices. Since this is a major problem, and since the summary also lacks context, it was assigned a Q2 score of 1.

Table 19: Example of LexRank summary in which the juxtaposition of two sentences that are not close to each other in the original text (where they are highlighted in bold) results in factual correctness problems.

Text ID	newsroom-1517 (in GitLab)
Summary	Fares vary considerably according to the time of year, and whether or not the airlines are offering discounted deals. Rodney Bay to Castries costs EC$2.25 (55p), and Soufrière to Castries EC$8 (£1.85).
Article	[...] Two airlines fly non-stop from the UK to St Lucia's Hewanorra International Airport, in the far south of the island: British Airways (0844 493 0787; www.ba.com) and Virgin Atlantic (0844 209 7777; www.virgin-atlantic.com), both from Gatwick. **Fares vary considerably according to the time of year, and whether or not the airlines are offering discounted deals.** You can also book flights through agents such as Trailfinders (020 7368 1200; www.trailfinders.com), and online through www.expedia.co.uk. Inter-island flights - for example, with Liat (www.liat.com) - arrive at the much smaller George FL Charles Airport just north of Castries. [...] Buses: these come in the form of minibuses, and can get you to pretty much anywhere on the island. Services along the main road between the northern corner of the island and Castries are frequent. **Rodney Bay to Castries costs EC$2.25 (55p), and Soufrière to Castries EC$8 (£1.85).**

Depending on the dataset, the average Q4 score for LexRank summaries is 3 or slightly above it, meaning that the style of the summaries barely satisfies the scenario requirements. Average Q5 scores are between 2 and 3 (with the exception of German data, where they are slightly above 3), meaning that the summaries are, in most cases, not adequate to preview the original articles. It is not surprising that the average Q4 scores are higher than the average Q5 scores: since the information coverage for the news scenario is flexible (cf. Section 3.1), in order to fulfill Q4 it is enough for a summary to convey what the topic of the original text is. Q5, on the other side, requires the summary to be suitable for either previewing the article or prompting to read it, which is not automatically fulfilled as soon as the summary conveys the topic of the text.

The main problem with LexRank summaries with low Q4 or Q5 scores is that they do not have a sufficient information coverage. An example can be seen in Table 20: the

summary does not convey what the original text is about, and it could not be used to preview it. Therefore, both the Q4 and the Q5 score assigned to this summary are 1. The same applies to the summary already shown in Table 19. The summary already shown in Table 17, on the other hand, manages to convey the topic of the article (which is the discrimination of women in the field of journalism), although some information is missing, and is also more adequate to preview the original article. Therefore, it was given a 3 both for Q4 and Q5.

Table 20: Example of LexRank summary with insufficient information coverage. The original text is about the work of sculpture and painter Thomas Schütte and his exhibitions in London, but the summary does not allow this to be understood.

Text ID	newsroom-1349 (in GitLab)
Summary	Schütte knows this. “You don’t make art,” Schütte observed when we spoke last week.

As can be seen in Table 16, LexRank summaries contain a lead bias more often than mLongT5 or NeMo summaries. Of 10 total LexRank summaries with lead bias, 8 have good Q1–Q5 scores, which shows that, at least in some cases, selecting the first sentences of a news article is a good strategy to fulfill the requirements of the news scenario. There are also high-scoring summaries that are not biased towards the lead of the article, but summaries that are have shown to be more likely to achieve good scores than summaries that are not. This is probably due to the fact that, as already mentioned in 5.1.2, news articles often present the most important information at their beginning. An example of a summary that leverages lead bias and achieves very good scores is shown in Table 21.

Table 21: Example of LexRank summary leveraging lead bias and scoring a 5 in each of the Q1–Q5 metrics.

Text ID	newsroom-1905 (in GitLab)
Summary	After the wacky and bungled second film, the state of the Insidious franchise – jack-in-the-box spookshows from the Saw brigade – looked parlous. Insidious 3 takes everything that was broken and fixes it.

The presence of these high-scoring summaries also sheds light on another aspect: the scores in Table 16 are only average values, but the set of LexRank results contains both very good and very poor summaries.

Since LexRank-generated summaries are short (summary length was set to 2 sentences for English and French data and to 3 sentences for German data), they usually have medium or low information dispersion.

mLongT5 results on English data

Grammar mistakes or wrong lexical choices are present in almost every mLongT5-generated English news summary, but they usually do not hinder comprehension. An example can been seen in Table 22, where “There will have” is used instead of “There will be”, and the

verb “honored” is used incorrectly, since a committee can award prizes and honor people, but it cannot honor prizes. Since these mistakes are not minimal, but the text is still well understandable, the summary was assigned a Q1 score of 3.

Table 22: mLongT5 summary containing a grammar mistakes and a wrong lexical choice (both highlighted in red).

Text ID	newsroom-131 (in GitLab)
Summary	Roger Kornberg, the 59-year-old Stanford University professor who won a Nobel Prize for research into how genes are transferred from one molecule to another, is making more antibiotics for tuberculosis. “There will have specific cures,” he said at a press conference Wednesday, per Fox News. Kornberg has been working on yeast cells since 2010 and produced detailed pictures of what scientists called transcription in organisms that include humans and other animals, the Swedish Academy of Science says. The prize was honored by the Nobel committee earlier this week.

As for structure and coherence, mLongT5 scores are only slightly higher than LexRank scores. Like LexRank summaries, mLongT5 summaries often lack context: in the summary in Table 23, for example, the first summary sentence lacks a temporal complement, and in the third sentence, it is not clear which study the summary refers to, since no particular study has been mentioned before. Despite these issues, the summary is overall coherent; therefore, it was assigned a Q2 score of 3.

Table 23: mLongT5 summary containing redundancy (“British Britons”) and lacking some context in the parts highlighted in blue. Please note that grammar mistakes are ignored here.

Text ID	cnndm-1463 (in GitLab)
Summary	Research shows a quarter of British Britons will come from minority background. The rise is due to baby booms among Pakistani and Bangladeshi immigrants. Campaigners are fearing this sudden increase would put pressure on NHS, schools and housing. It could also worsen Britain’s quality of life as a whole, according to the study. Professors predicted that by midpoint of 25th century ethnic minorites would make up 14.3% of total population.

Moreover, several mLongT5-generated summaries contain redundancy or repetitions (Table 23), or unclear pronoun references (Table 55 in the Appendix). More rarely, a wrong or misleading use of connectors creates coherence problems in the summaries: in Table 56 in the Appendix, for example, conjunction “and” links two clauses that are in opposition to each other, which makes the sentence appear contradictory. The use of a conjunction expressing contrast, such as “but”, would be more appropriate.

Factual mistakes occur in every single summary, and Q3 scores for Newsroom and CNN/DailyMail (cf. Table 16) show that, on average, factual correctness is poor. In particular, many summaries contain wrong numbers (e.g. years or percentages); names are also misspelled quite frequently, and literal quotes are often misquoted. Hallucinations are also frequent. Table 24 shows a summary containing a hallucination (Swift’s age is not mentioned in the original text), as well as a partially wrong quote and a misreported fact

(it was Madonna who sang Ghost Town, not Swift). The quote also highlights mLongT5's coherence issues once again, since its text, as it is reported in the summary, is not sensible. Since the factual problems identified in this summary are not fundamental, and the principal information from the original text is reported correctly, this summary was assigned a Q3 score of 3.

Table 24: mLongT5 summary containing a general factual mistake, a partially wrong quote and a hallucination. Mistakes are highlighted in red, the corresponding correct information in the original text in green. Please note that grammar mistakes and coherence problems are ignored here.

Text ID	newsroom-1664 (in GitLab)
Summary	Taylor Swift won the music awards at iHeartRadio on Sunday, and she also assisted Madonna with her guitar strumming. The 20-year-old pop star was nominated for artist of his year as well as song of that year for hit Shake It Off in Los Angeles, reports Billboard. "More than anything other than you know, we've got closer and closer to each year, just not further apart," she said while singing an understated version of Ghost Town.
Article	Taylor Swift cleaned up at the iHeartRadio music awards on Sunday, winning artist of the year and song of the year for Shake It Off, and even assisted Madonna by strumming her guitar onstage while the pop legend sang a new song. "More than anything in the world, I just hope that any of the fans watching know how much I adore you ... we've gotten closer and closer with each year, not further apart," Swift said at the Shrine Auditorium in Los Angeles. "Like, you make me so happy." The pop star also won best lyrics for her other hit, Blank Space. Madonna performed Ghost Town in understated style while Swift played guitar next to her. [...]

Style and use are better than in LexRank summaries, with average scores between 3.6 and 4.0 for Q4, and around 3.5 for Q5 (cf. Table 16). mLongT5 leverages lead bias slightly less than LexRank, and information dispersion in mLongT5-generated summaries is mostly medium.

mLongT5 results on German and French data

The average Q1–Q5 scores for German mLongT5-generated news summaries are slightly lower than for their English counterparts. In particular, as for Q1, incorrect lexical choices are significantly more frequent than in English.

As for Q2, German summaries have the same problems as the English ones: lack of context, wrong use of connectors, contradictions, redundancy, insensate sentences. The last two occur more frequently than in English. Moreover, as already mentioned in Section 5.2.1, German summaries tend to have very short sentences. Sentences are often overly simple too, in a writing style that sometimes appears childish (or written for children), as can be observed in Table 25. This might result from the fact that most German data in sumstew, the dataset on which mLongT5 was fine-tuned (cf. Section 4.2.2), is from Klexikon (Aumiller and Gertz [2022]), which contains German Wikipedia articles and, as their summaries, corresponding articles from children's encyclopedia "Klexikon" (Schulte and van Dijk [2015]).

As for Q3, factual mistakes and hallucinations are more frequent in German than in

Table 25: Example of mLongT5 German summary containing several short, simple sentences. Grammar mistakes are ignored.

Text ID	20min-242 (in GitLab)
Summary	Der Schweizer Fussballverband sucht einen neuen Trainer. Nach dem abgang von Peter Petkovic ist der Verein auf der Suche nach einem Nachfolger. Urs Fischer, der seit drei Jahren an die Seitenlinie steht, hat viel Erfolg. Er arbeitete bei Union Berlin und wurde zuletzt vorzeitig verlängert. Doch viele Fussball-Leser wünschen sich auch den deutschen Coach Jogi "Löw". Das wäre eine grosse Sensation. Die Nati würde aber nicht gewinnen. Es gibt noch einige andere Kandidaten.

English. With regards to Q4 and Q5, several German summaries miss the point of the corresponding original articles.

The average Q1–Q5 scores for French mLongT5 summaries of OrangeSum articles are very similar to mLongT5's Newsroom scores, and the main issues identified in the summaries are also similar. For the three randomly extracted summaries of MLSUM-fr articles, factual correctness, style and use are significantly above mLongT5-average (cf. Table 16). This is probably just due to the small, non-representative sample size, but it might also be linked to the fact that mLongT5 was trained on a dataset that also contains MLSUM-fr texts, as mentioned in Section 4.2.2.

NeMo results on English data

In NeMo summaries of English-language news articles, grammar and lexicon (measured in Q1) are always correct. Structure and coherence (Q2) are also very good, with only one case in which two adverbs are used in a misleading way, resulting in a slightly incoherent text that was assigned a Q2 score of 4 (Table 26).

Table 26: NeMo summary with misleading use of adverbs "however" and "still" (highlighted in red). Their use probably aims to express an opposition to what is written in the first summary sentence. However, since the second summary sentence already mentions some of the difficulties of the Colorado Avalanche, the use of the red-highlighted adverbs in the third sentence is misleading.

Text ID	newsroom-292 (in GitLab)
Summary	The Colorado Avalanche, after years of postseason struggles, is now guaranteed to make the playoffs under Patrick Roy's leadership. Despite ranking among the league's top offenses with stars like Duchene and Parenteau, their defense allows many shots per game (32.6), putting pressure on their goaltenders. However, historical data suggests they might still falter in the first round due to lack of puck possession and reliance on one-goal games.

Scores for factual correctness (Q3) are also good. However, some rare summaries contain imprecision or minor mistakes, as shown in Table 57 in the Appendix, where a year reported in a summary is incorrect. Moreover, in a few cases, NeMo adds new information to the summaries. In some cases, these are only minor additions that can

be considered acceptable. In Table 58 in the Appendix, for example, the original text only talks of "Hachette", whereas NeMo, in its summary, uses the complete name of the company, "Hachette Book Group". In other cases, however, these additions are too strong of an intervention: despite being correct, the additional information is misleading with regard to the content of the original text. The most extreme example of NeMo adding new information to a summary can be seen in Table 59 in the Appendix: in this case, two sentences in the summary contain information about a person that is not even mentioned in the original text.

In one case, the additional information in a NeMo-generated summary is incorrect, making it a hallucination: this can be seen in Table 60 in the appendix, where the summaries mentions a non-existing website. Since the rest of the summary is factually correct, it was assigned a Q3 score of 4.

Similarly to adding new information to the summaries, in some rare cases, NeMo also adds too much *interpretation*, as can be seen in Table 61 in the Appendix: the summary states that a recipe is "a popular choice", but the original text only reports that the author tried the recipe after reading it in a reader's message, which does not imply that the recipe is popular. In this case too, the summary was assigned a Q3 score of 4.

For Q4, NeMo scored 5 in all analyzed samples, meaning that the information coverage of the summaries always fulfilled the requirements of the news scenario. As for Q5, some generated summaries were too long for their designated use (cf. Section 5.2.1 and Table 10).

None of the analyzed summaries contains a lead bias, and all summaries have a medium or high information dispersion.

NeMo results on German data

German NeMo news summaries have overall good grammar, but contain some rare grammar mistakes. In the summary in Table 27, for example, a wrong verb form is used (indicative "waren", while conditional "wären" would be correct, since the sentence is about a hypothesis), and a wrong article declination as well ("Der Wechsel" instead of "Den Wechsel"). One lexical mistake was also observed: in the summary in Table 28, the word "Langzeitzugriff" is used improperly.

Table 27: Example of NeMo summary containing some grammar mistakes (highlighted in red) and an adverb that is not semantically coherent to the rest of the text (highlighted in blue).

Text ID	20min-2377 (in GitLab)
Summary	Christian Eriksen hatte nach seinem Wechsel zu Inter Mailand Schwierigkeiten, ein Haus zu finden und musste während des Lockdowns auf dem Trainingsgelände leben. Er hatte zwar bei Mannschaftskollegen gefragt, aber die beiden Wochen waren ihm zu lang gewesen. Er war also nicht alleine auf dem Gelände, sondern zusammen mit einem Koch und fünf Staff-Mitgliedern in Quarantäne. Der Wechsel nach Mailand hat Eriksen jedoch nicht bereut, trotz der schwierigen Startphase wegen der Corona-Epidemie.

Structure and coherence are also generally good, and only rarely show some minor problems. One of these problems is showed in Table 27: the use of adverb "also", which in German expresses a consequence, is not coherent with the preceding sentences, in which no fact is stated from which the consequence introduced by "also" could be drawn. Moreover, the summary is partially lacking context: at the beginning of the second article sentence, it is not clear what Eriksen has asked his team mates (even though it can be assumed).

It was also observed that, in more than half of the summaries, sentences are separated by line breaks (this can be observed both in Table 27 and in Table 28). This is unnecessary and can be confusing for the reader, since a line break usually suggests a change of topic, but there is no change of topic in the summaries. It is possible that this separation of the sentences is caused by the prompt that was given to NeMo, which, translated from German, asks to summarize the texts in "maximum three short sentences" (cf. Table 53 in the Appendix). In one case (see Table 62 in the Appendix), the summary sentences are even numbered, which is inadequate, since the summary is not an ordered list.

Table 28: Example of NeMo summary containing a word used improperly (highlighted in red).

Text ID	20min-1018 (in GitLab)
Summary	Social Media-Plattformen erfordern vertikales Scrollen auf Smartphones, was zu einer großen Daumenbewegung führt. Ilk hat berechnet, dass die jährliche Distanz etwa zwei Marathons entspricht. Langzeitzugriff kann zu Verletzungen wie dem "Scroller Daumen" führen.

Most summaries are factually correct, but some contain some mistakes, as can be seen in the summary in Table 29: according to the summary, two celebrities have just announced that they are dating, but according to the article, they announced it some time ago. The tendency to add new information to the summaries, which was observed in NeMo's results for the English news datasets, is less strong in the German NeMo-generated summaries that were manually analyzed.

Table 29: Example of NeMo summary containing a factual mistake (highlighted in red, while the corresponding correct information in the original text is highlighted in green).

Text ID	20min-1213 (in GitLab)
Summary	Luca Hänni (25) und Christina Luft (30), ein oficiais Paar seit Anfang Juli, sind ans Meer gereist und verbringen ihre ersten gemeinsamen Strandferien. Sie teilen zahlreiche Fotos und Videos auf Instagram und geben ihrer Beziehung öffentlich bekannt. Der genaue Ort ihrer Reise bleibt jedoch geheim.
Article	Seit Montag unterhalten sie uns mit zahlreichen Fotos und Videos, die neidisch machen: Luca Hänni (25) und Christina Luft (30) sind ans Meer gereist und halten ihre Strandferien auf Instagram fest. Es sind die ersten gemeinsamen Ferien, die die zwei offiziell als Paar verbringen. Erst Anfang Juli brachten sie auf den Tisch, was Wochen zuvor schon längst gemunkelt wurde. "Ja, wir sind zusammen", sagten der Berner Sänger und die deutsche Tänzerin. [...]

As for Q4, the information coverage of the summaries is generally in line with the scenario requirements. Only in one case (Table 63 in Appendix), the summary partially misses the point: most of the original text is about presenting possible future trainers for the Swiss national football team, but the summary, which has a lead bias, does not mention that, and only focuses on the content at the beginning of the article.

Like the English NeMo news summaries, German summaries are sometimes too long for the news scenario, which led to penalization in the Q5 scores.

In the analyzed German news summaries, lead bias was used more that in the English NeMo summaries, and info dispersion was slightly but consistently lower.

As already mentioned in Section 5.2.1, when analyzing truncated summaries, it was noticed that some of these summaries contain some words or sentences in English at the end. There are six of such cases in NeMo-generated summaries of MLSUM-de articles, and three in NeMo summaries of 20 Minuten articles – an example can be seen in Table 30. It cannot be excluded that some non-truncated summaries also contain foreign languages, but these were not searched for automatically. Of the 28 manually evaluated NeMo summaries of news text, none contained words or sentences in English (or in any other languages that are not German).

Table 30: Example of NeMo summary containing a sentence in English at the end (highlighted in red).

Text ID	mlsum-de-1540 (in GitLab)
Summary	Tony Martin musste die Tour de France nach einem Sturz auf der sechsten Etappe aufgeben und verlor sein Gelbes Trikot. Martin suffers key fracture; future uncertain

NeMo results on French data

In French NeMo news summaries, grammar mistakes are more rare than in German, leading to an average Q1 score of 4.8 for OrangeSum (cf. Table 16).

Structure and coherence are usually good, with some minor problems in a few summaries. The summary in Table 31, for example, is moderately repetitive: the last sentence presents some facts that have already been stated in the first sentence, and adds little new information. Moreover, two summaries contain sentences that are too long to be easily read and understood (Table 32). Two summaries also lack some context: the second sentence of the summary in Table 32, for example, mentions a video that is referred to as “the video”, even if the previous sentence does not mention this video. Therefore, the reader does not know what the content of the video is, or in which circumstances it was recorded.

The summaries are usually factually correct; factual mistakes, hallucinations and excessive interpretations are approximately as frequent as in the English NeMo-generated news summaries. Table 64 in the Appendix shows an example of a summary containing both a factual mistake and a hallucination.

Like in the German NeMo summaries, only 1 out of 28 French NeMo summaries has improvable information coverage, and therefore it does not have the highest Q4 score

Table 31: Example of moderately repetitive NeMo summary. The information that is presented several times is highlighted in blue.

Text ID	orangesum-1216 (in GitLab)
Summary	Quatre jeunes hommes ont été mis en examen pour proxénétisme en bande organisée, accusés d'avoir prostitué au moins quatre jeunes filles mineures à Rouen. Certains des accusés sont eux-mêmes mineurs et l'un d'eux était déjà détenu pour des faits similaires. Les suspects âgés de 16 à 22 ans sont soupçonnés d'avoir séquestré et forcé plusieurs jeunes filles à se prostituer pendant plusieurs mois, avec jusqu'à 15 relations sexuelles tarifées par jour à 100 euros la passe.

Table 32: Example of NeMo summary containing a long, not easily readable sentence (highlighted in blue) and using a determinative article (highlighted in red) to refer to something that has not been mention in the summary before, resulting in a lack of context.

Text ID	orangesum-1216 (in GitLab)
Summary	Arthur, un malvoyant âgé de 25 ans, a été expulsé avec son chien guide du Monoprix de la Blancarde à Marseille pour des raisons d'hygiène selon le directeur du magasin qui refuse de changer d'avis malgré les protestations de clients et l'application de la loi autorisant l'accès aux lieux publics pour les chiens guides. La vidéo devenue virale a suscité des excuses publiques de Monoprix et des sanctions contre le responsable.

(Table 65 in Appendix).

Manual evaluation showed that French NeMo-generated summaries sometimes are too long for the news scenario. This is the case more often than in the English and the German NeMo summaries, as it was also already observed in the automatic evaluation results (cf. Section 5.2.1).

Lead bias is never used in the analyzed French summaries, and information dispersion is similar as in the English NeMo summaries.

The manual evaluation conducted on truncated summaries (cf. Section 5.2.1) showed that, significantly more often than the German ones, French NeMo-generated summaries contain foreign languages at their end (mainly English, but also other languages, such as Mandarin and Russian). Of 46 out of 2500 OrangeSum summaries that were automatically identified as truncated, 17 contained a foreign language at their end. For MLSUM-fr, it was 23 out of 56 truncated summaries. Sometimes, the content in a foreign language stretches over one or more sentences, similarly to what has already been shown for German in Table 30. Some other times, it is only one word, and the rest seems to have been cut off (Table 33). It is interesting to observe that, often, the text in a foreign language starts with the English word "despite". Moreover, as already said for German summaries, even though foreign languages were only found at the end of summaries, it cannot be excluded that some other summaries also contain foreign languages positioned in other locations.

Table 33: Example of NeMo summary containing a word in English at the end (highlighted in red).

Text ID	orangesum-59 (in GitLab)
Summary	Jennifer Aniston a révélé que les acteurs de Friends n'étaient pas fans de la chanson du générique ni de la scène où ils dansent dans une fontaine. La scène a été tournée à 4 heures du matin pendant l'hiver, ce qui n'a pas dû aider les comédiens à apprécier l'expérience. despite

Table 34: Average manual evaluation scores for abstract generation scenario.

Dataset	**Method**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Lead bias**	**Information dispersion (low, medium, high)**
PubMed (10 samples)	LexRank	5.0	1.7	5.0	2.3	2.5	0/10	1, 5, 4
	mLongT5	3.4	3.4	2.8	3.6	3.8	0/10	2, 6, 2
	Nemo	5.0	4.7	4.8	3.7	4.4	0/10	2, 6, 2
TermITH (8 samples)	LexRank	5.0	1.0	5.0	1.75	2.0	0/8	1, 6, 1
	mLongT5	2.88	3.38	2.62	2.5	2.5	1/8	1, 6, 1
	Nemo	3.75	4.5	4.38	2.5	2.62	0/8	3, 4, 1

5.3.2 Abstract generation scenario

LexRank results

In the abstract generation scenario, LexRank Q2 scores are very poor (cf. Table 34), and significantly lower than in the news scenario. This is not surprising, since in the abstract generation scenario original texts are longer and compression ratio in the summaries is higher, meaning that only a small percentage of sentences are selected, so it is less likely for the resulting text to be well-structured and coherent.

In the English data, the main structure and coherence problems are lack of context and lack of syntactic and semantic connection between the sentences. The former can sometimes results in contradictions, while the latter, of which an example is shown in Table 66 in the Appendix, implies that summaries are just a heap of apparently unrelated information. Redundancy is also frequent in English summaries. An example of several of these phenomena can be seen in Table 35: the two sentences highlighted in red appear contradictory, since the first one states that only a small percentage of questionnaire respondents often uses mosquito nets, while the second one states that mosquito net usage among the respondents is high. The lacking context is that "Ninety-one point five percent (65) of the respondents indicated that they always used mosquito nets", as is written in the original paper immediately before the first one of the two extracted sentences highlighted in red. The aforementioned summary is also redundant, since some information (the fact that living close to mosquito breeding sites is a risk factor) is given twice.

In the French data, for which every analyzed summary received a Q2 score of 1, lack of context and lack of connection between the sentences are more extreme, and wrong sentence segmentation was also observed (an example of this phenomenon has already

Table 35: Example of LexRank summary lacking some context and containing redundancy (highlighted in blue) and an apparent contradiction (highlighted in red).

Text ID	pubmed-375 (in GitLab)
Summary	the questionnaire had four thematic sections; first part focused on sociodemographic characteristics such as gender, age, marital status, education level, farming practice, household income, ethnicity, respondent's relationship to household head, employment status, and occupation and the second part included aspects on how the community or individuals got exposed to mosquito bites due to late night activities, location of homesteads in relation to animal shelters and mosquito breeding sites, visits made to other areas outside tubu village in the last 8 months, and history of malaria episodes in the last 8 months. more than half of the respondents who possessed mosquito nets had not experienced any malaria attack indicating an association between previous malaria episode and possession of mosquito nets (table 3). only 1.4% (1) mentioned that they used mosquito nets more often whilst 5.6% (4) sometimes used them and 1.4% (1) never used mosquito nets. reported mosquito net use in tubu village was very high regardless of whether it is treated or not. the location of homesteads relative to mosquito breeding habitats was one of the major risk factors for malaria transmission. limited use of malaria protective measures such as insecticide treated nets, house structure (traditional or modern), and close location of homesteads in relation to breeding sites exposed individuals to mosquito bites.

been shown for the news scenario, cf. Table 54 in the Appendix). In this context, it has to be noted that some papers in TermITH are particularly long and have diversified content,[4] which makes it very challenging to summarize them by extracting a few sentences, without being able to abstract and condense content.

For both English and French summaries, Q5 scores are slightly higher than Q4 scores, since the abstract generation scenario, as defined in Section 3.2, allows multiple uses (either previewing the article, or substituting it, or refreshing one's memory about it). Therefore, use requirements are easier to (partially) fulfill. Q4, on the other hand, has stricter information coverage requirements, and is therefore harder to fulfill. As in the news scenario, the main reason why LexRank has poor Q4 and Q5 scores is insufficient information coverage. Insufficient information coverage was also observed when information dispersion was high, and at the same time, some of the summaries with sufficient information coverage have low information dispersion.

mLongT5 results

The Q1 score for mLongT5-generated PubMed summaries is 3.4 (cf. Table 34), meaning that grammar and lexicon are relatively good; this score is very similar to the Q1 score of English mLongT5 news summaries. In addition to regular grammar mistakes (e.g. wrong verb conjurations, or incorrect use of prepositions), mLongT5 often misspells technical terms, which are frequent in scientific papers.

For TermITH summaries, mLongT5 scores on average 2.88 on Q1, which is half a point less than it scores on OrangeSum news data. French paper summaries also contain misspelled technical terms.

[4]For example paper with IDs termith-93-archeologie or termith-340-communication in GitLab.

Both for English and French summaries, average Q2 scores are between decent and good, i.e., 3.4 and 3.38 respectively. This means that they are significantly higher than LexRank's Q2 scores (cf. Table 34). For most datasets used in the news scenario, on the contrary, structure and coherence in mLongT5 summaries were only marginally better than in LexRank summaries.

mLongT5's structure and coherence problems are very similar to those already observed in the news scenario: both in English and in French, summaries often contain redundancy, and, less often, insensate sentences, wrong use of connectors and lack of coherence. Two French summaries also lack context, and two English summaries contain long, complicated sentences that are hard to follow.

Factual correctness scores for both languages are similar to those observed for mLongT5 in the news scenario, and common issues are also similar.

For PubMed, Q4 and Q5 scores are on average quite good. Only 2 out of 10 analyzed summaries have generally insufficient information coverage, while in most cases only some specific information needed for Q4 is missing (e.g., the methods used in the paper).

For TermITH, both Q4 and Q5 average to 2.5, which means that style and use are not satisfactory with regard to the scenario requirements. These mLongT5 scores are more than one point lower than the corresponding scores for PubMed and for OrangeSum. Summaries often miss the point(s) of the paper, or have insufficient information coverage. This might be due to the nature of the papers in TermITH, which tend to be long and complex, as already mentioned.

For both languages, information dispersion is mostly medium. As already observed for LexRank paper summaries, high information dispersion and good information coverage seem to be independent from each other.

NeMo results

English grammar and lexicon are very good, as in the news scenario. In French summaries, grammar and lexicon are also good, but the average Q1 score (3.75, cf. Table 34) is lowered by the fact that 2 out of 8 summaries are completely written in English instead of French. Another French summary uses an English word in the middle of a sentence. The problem of NeMo using English or other languages in French summaries for the abstract generation scenario has already been observed in Section 5.2.2.

Structure and coherence are also good. Some minor problems are that two English and three French summaries contain some repetitions; moreover, one English summary contains a wrong use of a connector (the same type of mistake already shown in Table 27).

Most summaries are factually correct. Sporadic, small factual mistakes are found in both languages. One English summary also contains too much interpretation (this phenomenon, which was already observed in the news scenario, is shown in Table 61 in the Appendix). One other English summary contains a hallucination (a made up reference), and one French summary contains three pieces of information that are not in the original paper, even though they are correct.

Average style scores are very similar to the corresponding mLongT5 scores (cf. Table 34), meaning that, on average, style is satisfactory for PubMed, but not for TermITH. For both languages, it was observed that NeMo, in its summaries, tends to focus on the final part of the original paper. Therefore, its summaries always contain the results of the study, but it almost never mention its methods. In three cases in PubMed, the summaries

are so focused on specific results that the topic is not introduced, or the central question of the paper is not clear (see Table 67 in the Appendix).

In TermITH, it was observed that NeMo tends to focus on the final part of the paper also in the case of surveys or historical reviews, which do not have a results section.[5] Moreover, if a study contains both a qualitative and a quantitative analysis (or a survey), NeMo tends to focus on the quantitative results, even if the qualitative results play a more important role in the original text. An example is shown in Table 68 in the Appendix.

Another style-related observation is that, when mentioning the paper's contributions, some NeMo summaries of TermITH papers use impersonal verb forms, which do not make it clear that the mentioned facts or insights are a result of the study itself. An example can be seen in Table 36: formulations "peut être décrite" ("can be described") and "permet de déterminer" ("allows to determine") should be replaced by "we describe" and "we determine", in order for the reader to understand that those are the contributions of the paper, and not already-established knowledge.

Table 36: Example of NeMo summary using impersonal forms (highlighted in red). Corresponding personal forms used in the original abstract of the article are highlighted in green.

Text ID	termith-241-chimie (in GitLab)
Summary	Les émulsions sont des systèmes instables qui finissent par se décomposer, principalement par coalescence, c'est-à-dire la fusion des gouttelettes. La coalescence nécessite l'ouverture d'un trou entre deux gouttelettes adjacentes, ce qui est plus probable si les gouttelettes sont plus grandes. La croissance homogène des gouttelettes peut être décrite par un modèle de champ moyen à un paramètre, qui permet de déterminer l'énergie d'activation de la coalescence. Cependant, il existe également des régimes de croissance intermédiaires qui restent encore mal compris et pourraient être liés à plusieurs temps caractéristiques de la coalescence.
Original abstract	La destruction par coalescence des émulsions concentrées monodisperses peut se développer selon différents scénarios. Nous montrons que l'un de ces scénarios correspond à une croissance très homogène des gouttelettes sans accroissement de la polydispersité. Ce type de croissance est bien décrit par une équation type champ moyen et, par comparaison avec les expériences, permet de déduire la fréquence d'ouverture de film par unité de surface ω(T). De l'évolution de ω avec T, nous déduisons l'énergie d'activation du processus de coalescence pour une émulsion d'hexadécane dans l'eau stabilisée par un tensioactif non ionique.

The average Q5 score for TermITH is 2.6, which is very close to mLongT5's score of 2.5. Especially when it comes to the already-mentioned long and complex TermITH papers, NeMo summaries focus on some specific points, but fail to provide a good overview of the papers. PubMed summaries, on the other hand, have an average Q5 score of 4.4. Considering that they mostly focus on the paper's results, these summaries are suitable to refresh one's memory of the content of the original article, but less suitable for the other two uses defined for this scenario (previewing or substituting the original text).

[5]Example of such papers have following IDs in GitLab: termith-340-communication, termith-710-linguistique.

Table 37: Average manual evaluation scores for web snippet generation scenario. Before performing manual evaluation, all summaries longer than 160 characters were truncated. This is important to keep in mind when looking at data about lead bias.

Dataset	Method	Q1	Q2	Q3	Q4	Q5	Lead bias	Information dispersion (low, medium, high)
DE	LexRank	5.0	3.3	5.0	2.4	2.5	1/10	10, 0, 0
	mLongT5	4.4	3.9	2.4	3.5	3.1	0/10	6, 4, 0
	Nemo	5.0	5.0	4.9	4.5	4.3	4/10	5, 4, 1
EN	LexRank	5.0	3.8	5.0	3.3	3.0	1/10	10, 0, 0
	mLongT5	4.1	4.1	3.1	4.2	4.3	5/10	6, 3, 1
	Nemo	4.9	5.0	4.9	4.8	4.8	3/10	3, 6, 1
FR	LexRank	5.0	3.5	5.0	3.0	2.7	1/10	10, 0, 0
	mLongT5	4.7	4.5	3.4	4.4	4.4	4/10	5, 5, 0
	Nemo	4.9	5.0	5.0	4.3	4.4	6/10	6, 3, 1
IT	LexRank	5.0	3.9	5.0	3.1	3.1	2/10	10, 0, 0
	mLongT5	4.3	4.1	3.2	3.8	3.8	3/10	5, 5, 0
	Nemo	4.8	4.9	5.0	4.1	4.3	5/10	6, 4, 0

5.3.3 Web snippet generation scenario

As already mentioned in Section 5.2.3, since this scenario has a strict length requirement, all generated summaries longer than 160 characters were truncated and terminated by three suspension points. Manual evaluation was conducted on the truncated results.

LexRank results

Like in the news scenario, some LexRank-generated summaries are very good, while some others are very poor. Two examples are shown respectively in Table 38 and Table 39. Moreover, as in the news scenario, it was observed that summaries extracting one of the first sentences from the original text are more likely to achieve good scores for Q2, Q4 and Q5.

Table 38: Example of LexRank summary scoring 5 in Q1–Q5. The summary consists of the first sentence of the original text.

Text ID	web-en-9 (in GitLab)
Summary	Between 6–9 June 2024, millions of Europeans will participate in shaping the future of European democracy on the occasion of the European elections.
Source	https://elections.europa.eu/en/why-vote/ (last access: 19th January 2025)

Average Q2 scores (cf. Table 37) are relatively good across all languages, the main structure and coherence problem of the summaries being lack of context.

Average Q4 scores are decent in English, French and Italian, and poor in German. Average Q5 scores are slightly lower than Q4 scores. In general, one sentence is not

Table 39: Example of LexRank summary scoring 1 in Q4 and Q5. The website to summarize (available at the link in field *Source*) is a guide on how to write an eulogy for a dead grandparent.

Text ID	web-de-10 (in GitLab)
Summary	Frage auch andere Menschen nach ihren Erinnerungen.
Source	https://de.wikihow.com/Eine-Trauerrede-f%C3%BCr-ein-verstorbenes-Gro%C3%9Felternteil-schreiben (last access: 19th January 2025)

always enough to convey what the text is about (cf. Table 39), and therefore, information coverage is not always sufficient to fulfill style and use requirements. This is particularly true if the original text cover several (sub)topics.[6]

mLongT5 results

In general, mLongT5 achieves good results on this scenario, with factual correctness standing out as the lowest average score in each language, but still being acceptable in 3 out of 4 languages.

The main problems with regards to Q2 are similar to those identified in the two previously analyzed scenarios: redundant or repetitive summaries, lack of context, contradictions, insensate sentences.

German summaries have consistently worse Q3 scores than summaries in other languages. In particular, they contain more hallucination than the other summaries. These hallucinations often consist of definitions or general, encyclopedia-like sentences, as can be seen in the first summary sentence in Table 40. Similarly to what was observed for short and overly simple sentences in Section 5.3.1, this behavior might result from the fact mLongT5 was fine-tuned on a dataset containing German Wikipedia article and, as their summaries, corresponding articles from a children's encyclopedia. This would also explain the tendency to redundancy (at least in the eyes of an adult reader), often observed in German summaries, and shown in the second summary sentence in Table 40.

Table 40: The first sentence in this German mLongT5 summary is a hallucination giving a (wrong) definition of what the EU is. The second sentence is redundant and written in a childish style. Moreover, the summary completely fails to convey the topic of the article, which is how the EU manages its finances.

Text ID	web-de-3 (in GitLab)
Summary	Die Europäische Union (EU) ist eine Weltgemeinschaft. Sie besteht aus 27 Ländern, die Mitglied sind in der Europäischen Union. Im Jahr 1950 wurde die EU...
Source	https://www.eca.europa.eu/de/the-eu-finances (last access: 19th January 2025)

[6]This applied to at least 3 out of 10 texts. In GitLab, these texts have the following IDs: web-1, web-8, web-10.

The average Q4 and Q5 scores are over 4 for English and French summaries, and over 3 for German and Italian summaries, meaning that, in most cases, summaries successfully convey the topic of the original text and could be used to preview it. In the few cases in which a summary does not fulfill the style and use requirements, the reason is either that hallucinations make the summary appear to be on another topic (which happens in two German cases, one of which in Table 40), or that the summary focuses on non-central aspects of the article (which happens once in Italian).

NeMo results

In all languages, NeMo summaries have very good Q1, Q2 and Q3 scores. However, the problem of foreign languages in non-English summaries was identified in this scenario as well: one French and one Italian summary contain an English word each.

Average Q4 and Q5 average scores are also good, but not as good as Q1–Q3 scores. Three summaries in total (of which one in French and two in Italian) focus on specific, non-central aspects of the text, but fail to convey what its broader topic is. In general, it was observed that, in English, NeMo has better abstraction and concision capabilities than in other languages (hence the higher Q4 and Q5 scores, cf. Table 37), especially when the original text covers several (sub)topics. In the case reported in Table 41, for example, the original text presents several touristic attractions in the city of Brno. The English summary is the only one that first mentions the name of the city and then provides a short list of several of its points of interest. The German, French and Italian summaries, on the other hand, start with a description of the first touristic attraction, which might lead the reader to think that the text is about that specific sightseeing spot, rather than about the city.

Table 41: NeMo summaries of the same text in different languages. The English summary stands out as the only one that truly gives an overview of the content of the original text.

Text ID	web-en-1 / web-de-1 / web-fr-1 / web-it-1 (in GitLab)
Summary (EN)	Brno boasts modern architectural gems like Villa Tugendhat, a UNESCO site, and historical attractions such as Špilberk Castle. It offers unique experiences incl...
Summary (DE)	Die Villa Tugendhat in Brünn ist ein bedeutendes Beispiel funktionalistischer Architektur und steht seit 2001 auf der UNESCO-Weltkulturerbeliste. Die neogotisch...
Summary (FR)	La villa Tugendhat, œuvre unique de Ludwig Mies van der Rohe, est un joyau de l'architecture moderne inscrit au patrimoine mondial de l'UNESCO, tandis que la ca...
Summary (IT)	La Villa Tugendhat, progettata da Ludwig Mies van der Rohe, è un gioiello dell'architettura moderna e un sito UNESCO; la Cattedrale dei SS. Pietro e Paolo a Brn...

Table 42: Average manual evaluation scores for the school material summarization scenario.

Dataset	Method	Q1	Q2	Q3	Q4	Q5	Lead bias	Information dispersion (low, medium, high)
DE	LexRank	5.0	2.0	5.0	2.2	2.4	0/10	0, 3, 7
	mLongT5	3.8	2.3	1.4	1.1	1.3	0/10	1, 9, 0
	Nemo	4.9	4.7	4.7	3.6	4.2	0/10	0, 3, 7
FR	LexRank	5.0	2.2	5.0	2.4	2.6	0/10	0, 7, 3
	mLongT5	3.0	1.8	2.0	1.1	1.2	0/10	0, 10, 0
	Nemo	4.1	4.7	4.9	4.1	4.3	0/10	1, 2, 7

5.3.4 School material summarization scenario

LexRank results

The low Q2 scores (cf. Table 42) are linked to problems already observed in other scenarios: both in German and in French, summaries often lack context and cohesion, and are sometimes repetitive. As already observed in the news scenario, in two cases structure problems in the summary lead to apparent factual mistakes. Incorrect sentence segmentation was also observed.

Several texts in this scenario contain lists, both ordered and unordered. Both proved to be problematic for LexRank. One example can be seen in Table 69 in the Appendix: for the first ordered list in the original text, the text of 3 out of 4 list items is (partially) extracted, but only the first item is extracted with its number. For the second ordered list in the original text, only the first list item number is extracted, but no actual content from the list.

Q4 and Q5 scores are also poor, since LexRank summaries often have insufficient information coverage. This is linked to the extractive nature of the summaries in combination with the scenario requirements. Since this scenario requires a very high information coverage, a good summary would have to retrieve information from (almost) every sentence and compress it into fewer sentences. LexRank summaries, being extractive, cannot do this. High information dispersion (which 7 out of 10 German summaries have) is not enough to fulfill the high information coverage requirement.

mLongT5 results

Q1 scores are decent (3.8 for German texts and 3.0 for French texts, cf. Table 42) and in line with the corresponding mLongT5 scores in other scenarios.

Q2 scores are poor, with summaries containing problems already observed in other scenarios: lack of context (e.g., pronouns without a referent), redundancy, insensate sentences, lack of cohesion, wrong use of connectors.

In German summaries, factual correctness is very poor, with an average score of 1.4. Summaries are often factually incorrect (which also applies to numbers and work titles) and often contain hallucinations. In particular, like in the web snippet generation scenario, summaries tend to provide definitions or include general, encyclopedia-like sentences with

information that is not in the original text. French summaries have better, but still poor factual correctness, and also contain wrong facts and some hallucinations.

As already mentioned in Section 5.2.4, summaries are too short to have a sufficient information coverage for the school material summarization scenario. Moreover, given the frequent hallucinations, summary content does not always overlap with the content of the original text. Therefore, Q4 and Q5 scores are very poor (for both languages, they are more than 1 point lower than for LexRank, cf. Table 42).

NeMo results

In German summaries, grammar and lexicon are very good, with only one minor lexical imprecision. In French, they are also good, but a few grammar mistakes are present. The fact that 3 out of 10 French summaries contain foreign languages (one Russian word in the middle of the text in one case, and one English sentence at the end in the other two cases) also contributes to lowering the Q1 score.

Table 43: Example of NeMo summary containing an extractive chunk (in bold) with a demonstrative adjective missing a referent (highlighted in red). In the original text, the sentence containing the reference is highlighted in green.

Text ID	school-de-9 (in GitLab)
Summary	Der Mensch hat seit jeher versucht, die grundlegenden Bausteine der Welt um uns herum zu verstehen und hat sich dabei Vorstellungen von diesen Bestandteilen gemacht, die in den Naturwissenschaften als Modelle bezeichnet werden. Der griechische Philosoph Leukipp war einer der Ersten, der sich mit dem Konzept der Atome beschäftigte und erkannte, dass es eine Grenze des Teilbaren geben müsse. Er und sein Schüler Demokrit gaben **diesen kleinsten Teilchen** den Namen „atomos“ – das Unteilbare. Obwohl die Idee der Atome zunächst nicht durchgesetzt werden konnte und erst Jahrhunderte später von John Dalton wieder aufgenommen wurde, bildet sie heute eine fundamentale Basis der Chemie und Physik.
Original text	[...] Der griechische Philosoph Leukipp war einer der Ersten, der sich mit dem Thema Atome beschäftigte. Er erkannte rund 450 Jahre vor Christus, dass es eine Grenze des Teilbaren geben muss: „Wenn man einen Apfel in immer kleinere Stücke teilt, so würden die Teile irgendwann unendlich klein sein. Sie bestünden aus Nichts. Wenn man dann wieder den Apfel zusammensetzen wollte, so müssten diese Teilchen aus Nichts plötzlich ein winziges Stück Apfel ergeben.“ Dieses ergab für Leukipp keinen Sinn. Seine Lösung für dieses Problem war eine Grenze des Teilbaren: Es musste kleinste Teilchen geben, die sich nicht weiter teilen ließen. Leukipp und sein Schüler Demokrit gaben **diesen kleinsten Teilchen** den griechischen Namen „atomos“ - das Unteilbare. [...]

Q2 scores are also good. The following minor issues were observed: in French, two summaries are slightly redundant, and three are truncated. In German, one summary lacks a logical link between two sentences, and in another one a referent is missing. This latter case, which has often been observed in LexRank and mLongT5 summaries, is registered for the first time in a summary generated by NeMo. The lack of a referent usually results from the use of extractive chunks that are not adapted to their new context in the summary (this is shown in Table 43). As observed in Section 5.2.4, German NeMo summaries in

Table 44: Average manual evaluation scores for email thread summarization scenario.

Dataset	Method	Q1	Q2	Q3	Q4	Q5	Lead bias	Information dispersion (low, medium, high)
Preprocessed	LexRank	5.0	2.9	5.0	2.1	2.6	1/10	1, 9, 0
	mLongT5	4.0	2.9	2.4	2.1	2.4	0/10	1, 8, 1
	Nemo	5.0	4.7	4.0	4.1	4.7	0/10	0, 4, 6
Unprocessed	LexRank	4.2	1.5	5.0	1.8	2.2	0/10	1, 7, 2
	mLongT5	4.0	2.4	1.8	1.1	1.1	0/10	0, 9, 1
	Nemo	4.9	4.8	4.8	4.2	4.5	0/10	1, 2, 7

this scenario have a higher extractivity score than in the news and in the web snippet scenario, which might explain why the lack of a referent was observed here.

Factual correctness is, on average, very good too. Some factual imprecision was found in a few summaries, and the addition of new information in one.

As for style, German summaries are sometimes missing some information, especially in texts that are more discursive and less about reporting events or data (see Table 70 in the Appendix for an example). Moreover, in 4 out of 10 cases, German summaries contain much information from the first half of the original text and little information from its second half. An example of this behavior is also shown in aforementioned Table 70. As for French, sometimes information coverage is also below the desired level, but less frequently than in German.

As in the abstract generation scenario, use has less strict requirements than style, so average Q5 scores are higher in both languages. Higher information coverage would improve not only Q4, but also Q5 scores.

5.3.5 Email thread summarization scenario

LexRank results

As expected, LexRank produces better results if the input texts are preprocessed, but its results are poor for both datasets (preprocessed and unprocessed). The most significant difference can be seen when comparing the average Q2 scores (2.9 for preprocessed texts, 1.5 for unprocessed texts, cf. Table 44), but the difference also shows for Q4 (2.1 vs 1.8) and Q5 (2.6 vs 2.2).

For both datasets, the biggest Q2 problem is lack of context. In particular, summaries often contain sentences written by different people involved in the email exchange, but they provide no information about which sentences where written by whom. To mitigate this problem and highlight the fact that summary sentences might not belong together, when joining several summary sentences extracted by LexRank in single texts, these sentences were separated by line breaks (instead of spaces, as in the other scenarios).

Some summaries of unprocessed email threads contain email headers. These include, among others, information about sender, receiver, subject and date of the email, so they should be useful in providing information about the authors of the summary sentences. However, headers appearing in the summaries do not always belong to the email from

which the following summary sentence is taken. Therefore, email headers do not help in understanding which person involved in the email exchange wrote what to whom, but rather create confusion. An example can be seen in Table 71 in the Appendix. Moreover, since the emails in the unprocessed data are in inverse chronological order, this is also the case in the summaries, which makes them hard to make sense of. This can also be seen in Table 71.

As for Q4 and Q5, LexRank results show the same problem already mentioned for the other scenarios, i.e. lack of meaningful information to fulfill the style and use requirements. This shows the most in the summaries of unprocessed emails, since they contain more noise and less useful information, but it also applies to the summaries of preprocessed threads.

mLongT5 results

mLongT5 also performs poorly in this scenario (both in its preprecessed and in its unprocessed version), with the only exception being its good Q1 average scores. Like LexRank, mLongT5 produces better results when the input texts are preprocessed, especially with regard to style (the average Q4 scores being 2.1 for preprocessed text, vs. 1.1 for unprocessed texts – cf. Table 44) and use (2.4 vs. 1.1), but also to factual correctness (2.4 vs 1.8). For preprocessed input texts, LexRank and mLongT5 have almost identical average scores for structure and coherence, style, and use, while for unprocessed texts, mLongT5 has better Q2 but worse Q4 and Q5 scores.

A coherence problem typical of mLongT5-generated summaries for this scenario is a confusion between different verb and pronoun forms: on one hand, mLongT5 reformulates some parts of the email threads, hence using the third person. On the other hand, mLongT5 also extracts chunks of the original text without adapting them to the context of the summary, which means that summaries also contain verbs and pronouns in the first and second person, as shown in Table 45.

As for factual correctness, one problem concerning almost half of the summaries is that mLongT5 often confuses what was said by whom in the email exchange, i.e. a summary suggests that A said X and B answered Y, whereas the reality is that B said X and A answered Y. The summary in Table 46, for example, suggests that John Lavorato emailed Alex Pritchard looking for trading opportunities at Enron. However, by reading the original email exchange, it becomes clear that it is Alex Pritchard who emailed John Lavorato with the above-mentioned intention. Similarly, mLongT5 also confuses the roles of people: in the summary shown in Table 45, for example, Brian is presented as an application developer, whereas the original email thread shows that it is Fangming who developed the application.

In addition to these scenario-specific problems, mLongT5 summaries have the same problems that have already been mentioned for other scenarios: concerning Q2, some redundant or senseless bits of text, and concerning Q3, misspelled names, wrong numbers and dates, some hallucinations.

As for Q4 an Q5, the summaries often miss the point of the email thread, and they fail to report what the result of the email exchange is (i.e., what has to be done or what has been done, and in order to achieve what goal).

Table 45: Example of mLongT5 summary containing both verb forms in the third and in the second person. The use of the second person ("you") comes from extraction of a phrase from the original text (highlighted in bold both in text and summary) that has not been reformulated properly.

Text ID	emails-prepr-2 (in GitLab)
Summary	Brian Hoskins, Enron's messageboard application developer, is out of town on Friday. He will not be available to show the demo until Wednesday **if you cannot see it today**. John and Brian have been added to the NT group so they can take part in the demo.
Original thread	Hi, Brian: Let me know if you have time for the messageboard application demo today. I am planning to take vacation on Friday and coming Monday. I will not be available to show you the demo until next Wednesday **if you can't see it today.** Let me know your schedule. Thanks, Fangming ——— Fangming, I am out of town until tomorrow so will not be able to see the demo today. Wednesday afternoon looks good for me if you'd like to do it then. John, does this work for you? Brian [...]

Table 46: Example of mLongT5 summary confusing who wrote what to whom.

Text ID	emails-prepr-1
Summary	John Lavorato, 3726 Divisadero. St. San Francisco, CA (94023), emailed Alex Pritchart at Yahoo.com in an attempt to find out about trading opportunities in the energy industry and more specifically at Enron's market. He has ten years of experience in the trade business, primarily trading listed options for futures and stock as well as risk management control for traders in a corporate environment.
Original thread	Dear John, My friend, Bill McIlwain, suggested that I write you in an effort to find out more about trading opportunities in the energy field and more specifically at Enron. I have ten years experience in the trading business primarily trading listed options on futures and stocks. I have traded both proprietary positions and individual accounts as well as providing risk management control for other traders in a corporate environment. [...] Thank you in advance for your consideration. Best Regards, Alex Pritchartt ===== Alex Pritchartt 3725 Divisadero St. San Francisco, CA 94123 [...]

NeMo results

Overall, NeMo performs well on this scenario. It achieves similar results on preprocessed and unprocessed data, with the only exception of factual correctness being consistently better on the latter.

Especially for preprocessed data, it stands out that NeMo summaries contain more incorrect facts than in other scenarios. These tend to appear when information is nuanced, as shown in Table 47: in the original email thread, Mandy wrote that the badge will "most likely" be ready by Monday morning, while in the summary, this information is presented as if it were a certainty.

Table 47: Example of NeMo summary containing an incorrect fact (highlighted in red, while the corresponding correct information in the original text is highlighted in green).

Text ID	emails-prepr-4 (in GitLab)
Summary	Mandy requested Ina to fill out a Badge Access Form, which she did; Mandy confirmed the badge would be ready by Monday morning for pickup at the 5th floor reception.
Original thread	« File: Badge Access Form.doc » Ina, Pleae fill out and return to me at ECS 05848. You can e-mail this to me if this is easier. Thanks! Mandy --- « File: Badge Access Form.doc » I filled out all of the information that I had on him. Will he be able to have his badge by Monday morning and where will he go to pick it up. Ina --- Ina, We can most likely have this by Monday morning and he can pick this up at the 5th floor reception. If he has any problems he can call me. Thanks! Mandy [...]

Moreover, in one summary of a preprocessed email thread, NeMo also confuses who wrote what in an email exchange (this phenomenon, as mentioned in Section 5.3.5, was often observed in mLongT5 summaries).

As for Q4 scores, summaries do not always report the result of the exchange. Summary in Table 48, for example, does not mention for what date and time the presentation was scheduled. An opposite example is the aforementioned summary in Table 47: the result of the exchange is that the badge will probably be ready by Monday morning and can be picked up the the 5th floor reception, and it is mentioned in the summary.

5.4 Discussion

5.4.1 Results per scenario

Based on the results of manual evaluation, NeMo performs better than the other methods in all scenarios, with consistently good results in every scenario except for the French abstract generation scenario. However, even if NeMo always outperforms LexRank and mLongT5, performance differences between the three methods also depend on the scenario and on the dataset. In particular:

Table 48: Example of NeMo summary not focusing on the result on the exchange. For readability reasons, the preprocessed email thread is reported in cell *Original thread*, even though the summary summarizes the corresponding unprocessed thread.

Text ID	emails-unpr-2 (in GitLab)
Summary	This email thread discusses scheduling a demonstration of a new tool for two recipients, John Arnold and Brian Hoskins, with Fangming Zhu coordinating.
Original thread	[...] Let's plan on meeting between 3 and 3:30pm on Wednesday. I'll call you on Wednesday. Brian [...] --- Oops, I hadn't gotten to this one yet. Can we do it any later? --- 4pm? Brian [...] --- yes

- In the news scenario, LexRank has acceptable scores (between 3 and 4) for German data, with the exception of Q2. Moreover, mLongT5 has acceptable scores for Newsroom and OrangeSum, with the exception of Q3.
- In the abstract generation scenario, mLongT5 has acceptable scores for PubMed, with the exception of Q3. This is also the scenario on which NeMo performs the worst, since information coverage requirements are quite strict and NeMo does not always fulfill them.
- In the web snippet generation scenario, mLongT5 has acceptable or good (> 4) scores for all languages, with the only exception of the Q3 score for German. Moreover, LexRank has acceptable scores for English and Italian data.
- In the school material and the email thread summarization scenarios, on the other hand, NeMo has by far the best scores, while other methods perform poorly. However, NeMo's style scores for these two scenarios are slightly lower than in the news and in the web snippet scenario (or considerably lower, in the case of German school material summaries).

5.4.2 Results per ATS system

LexRank

In general, structure and coherence scores in LexRank summaries are poor (between 2 and 3), or very poor (< 2). The main problems are lack of context and lack of syntactic or semantic connection between the sentences; in longer summaries (i.e., in the abstract generation scenario), repetitions are often an issue too. It was observed that structure and coherence tend to be worse when summaries are longer: in these cases, the set of selected sentences does not build a coherent text, but rather presents several pieces of information that appear unrelated (or even contradictory) because of lacking syntactic or semantic

connections. This is most evident in the abstract generation scenario, which has very long original texts, therefore increasing the chance that the sentences selected for the summary contain information on different topics. The web snippet generation scenario, on the other hand, has the best LexRank Q2 scores, since summaries are one-sentenced, which means that the problem of unrelated sentences does not arise.

Style and use scores vary depending on the scenario, but are on average between poor and decent. The biggest problem in this context is insufficient information coverage. This is most visible in the school material summarization scenario, which requires high information coverage: in order to fulfill this requirement, a summarization system would have to retrieve information from (almost) every sentence and compress it into fewer sentences, which is impossible for an extractive algorithm. Consequentially, LexRank's style and use scores are better in scenarios with flexible information coverage, i.e. in the news and in the web snippet scenario. Since summaries for these two scenarios are also short, they are the two scenarios in which LexRank generally shows its best results.

mLongT5

Grammar and lexicon are generally acceptable. Structure and coherence are acceptable in approximately half of the cases, and poor in the other half, the most frequent problems being redundancy, lack of context, incorrect use of connectors, and insensate sentences.

Factual correctness scores vary depending on the scenario, but are mostly poor. German summaries tend to contain more hallucinations, which leads to worse Q3 scores than in the other languages; this might depend on the German dataset on which mLongT5 was fine-tuned, as explained in Section 5.3.3.

The web snippet generation scenario stands out as the one in which mLongT5 achieves its best Q1–Q3 scores. This might depend on the fact that summaries in this scenario are very short (they were truncated after 160 characters), so they leave little space for linguistic or factual mistakes.

Style and use scores also vary depending on the scenario: they go from very poor in the school scenario (because of insufficient information coverage) to decent or good in English and French news datasets, English abstract generation scenario and web snippet generation scenario.

In general, it was observed that mLongT5 has the best results on the web snippet generation scenario, but no significant difference between its performance across several languages was observed, with the already-mentioned exception of German Q3 scores being below average.

Moreover, it was noticed that mLongT5-generated summaries are less abstractive than they appear: in several cases, they only slightly reformulate sentences from the original text, which sometimes results in grammatical mistakes (cf. Section 5.2.1).

NeMo

One of the main problems identified in NeMo summaries is that NeMo does not respect the length requirements specified in the prompts, especially in French and in Italian. Moreover, NeMo often truncates French summaries, mostly when those exceed the specified target length.

In non-English NeMo summaries, the use of foreign languages was also observed; it is more frequent in French summaries, but it was also noticed in German and Italian. Other than this, grammar and lexicon are good or very good in all NeMo summaries, with only isolated mistakes. The same applies to structure and coherence, with some relatively rare cases of redundancy or repetitions, especially in longer summaries (i.e., in the abstract generation and in the school material summarization scenario), and some isolated cases of incorrect use of connectors.

Factual correctness is also good in all scenarios, with isolated factual mistakes (more frequent in the preprocessed email thread scenario). Moreover, some summaries contain new information, hallucinations or excessive interpretation. This has to be kept in mind, especially if the use of LLMs continues to increase among people outside of the NLP community, who may perceive the results as trustworthy because of their good grammar, lexicon and cohesion.

Style is good in shorter summaries (especially in the news and web snippet scenarios) and worse, but mostly still acceptable, in longer summaries, which sometimes have insufficient information coverage (especially in the abstract generation and German school material scenario). Use is good in all scenarios except for French abstract generation. While in this work prompts were kept generic for the reasons explained in Section 4.2.3, tests have shown that using more specific prompts (e.g., prompts describing the use case of the summary) can help improve style and use.

It was also observed that NeMo produces better results in English than in other languages: this is particularly evident in the web snippet generation scenario, where English summaries are better at condensing information and at focusing on the general topic of the original text and not on irrelevant details.

5.4.3 Correlation between ROUGE scores and manual evaluation

In the news scenario, there is only little correlation between ROUGE scores and manual evaluation results. For Newsroom and for German news summaries (i.e., for datasets 20 Minuten and MLSUM-de), NeMo has better ROUGE scores than the other two methods, which is in line with manual evaluation results. However, for the remaining news datasets, ROUGE scores are quite similar across the three different summarization methods, and there is no method that clearly has better scores for all three metrics (ROUGE-1, ROUGE-2 and ROUGE-L). Manual evaluation scores, on the contrary, clearly shows that NeMo performs the best on the scenario.

In the abstract generation scenario, correlation between ROUGE scores and manual evaluation results is even lower. For PubMed, LexRank has the best ROUGE scores, but the worst Q2, Q4 and Q5 scores in manual evaluation. For TermITH, LexRank and mLongT5 share the best ROUGE scores, while the best Q4 and Q5 scores are shared by NeMo and mLongT5. What does correlate to manual evaluation is the fact that all methods perform better on PubMed than on TermITH.

Moreover, as it was discussed in Section 5.2.1, ROUGE scores seem to be dependent on some qualities of the reference summaries, such as their length. Therefore, ROUGE scores of summaries of different datasets can hardly be compared, at least if there is only one reference summary per text.

5.4.4 Usefulness of other metrics

Q1–Q5 proved useful in evaluating different aspects of the summaries. In several scenarios, Q4 and Q5 average scores are similar, since both metrics depend on information coverage. However, the two metrics have different focuses, and this is shown by the fact that their scores are not identical.

The metric focusing on lead bias was useful too: it showed that, in the news scenario, LexRank leverages lead bias more often than mLongT5 and NeMo, and that in most cases this leads to good results. It also showed that, in the other scenarios, none of the three ATS systems is particularly biased towards the beginning of the article. Instead, in the majority of the analyzed summaries, information dispersion is medium, meaning that content from different parts of the original text is retrieved, even if not every part is covered.

The metric on information dispersion did not provide any additional information on the quality of the summaries. In particular, against expectations, the metric was not useful to predict if summaries have sufficient information coverage.

Chapter 6

Conclusions

In this work, five multilingual ATS scenarios were defined. Each one was provided with requirements, both intrinsic (e.g. target summary length) and extrinsic (style and use of the summary, as in Spärck Jones [1999]). The idea behind this, presented by ter Hoeve et al. [2020], is that reflecting on the purpose of the summaries and defining corresponding summary constraints is useful to evaluate ATS results, i.e. to determine if a generated summary is a good summary for its intended use case. Therefore, based on the requirements of the scenarios, quantitative and qualitative evaluation metrics were defined as well.

Each scenario was also associated with one or more datasets. Existing datasets were filtered and manually analyzed before being used to evaluate the scenarios. This showed that almost all datasets have data cleanliness problems, such as paragraph titles or image descriptions included in the texts. Content problems were also observed: for example, English news summarization dataset Newsroom [Grusky et al., 2018] often contains incomplete articles; moreover, all news summarization datasets contain both reference summaries presenting information that is not mentioned in the corresponding article, and reference summaries that only make sense if read in combination with the article's title.

After the datasets were filtered, they were used to evaluate the performance of extractive algorithm LexRank [Erkan and Radev, 2004], pre-trained language model mLongT5 [Uthus et al., 2023] and large language model Mistral NeMo [Mistral, 2024] on the five scenarios. The summaries produced by these ATS systems were evaluated both automatically and manually. LexRank often failed at writing well-structured and coherent summaries; in minor extend, mLongT5 also did. Additionally, mLongT5 often reported factually incorrect statements or hallucinations in summaries. In some scenarios, both systems also produced summaries with insufficient information coverage. Problems linked to factually incorrect content, hallucinations and insufficient information coverage were also observed in NeMo-generated summaries, but only rarely. Additionally, NeMo sometimes did not respect length requirements, and it sometimes used words or sentences in the wrong languages in its summaries (both of these problems mostly appeared in French summaries).

Despite these problems, NeMo outperformed the other two systems in every scenario, and had good results for almost every dataset. However, performance differences between the systems varied based on the scenario and on the dataset. For some news and websites datasets, LexRank and mLongT5 obtained acceptable or good scores on most evaluation

metrics, and mLongT5 also did for one scientific paper dataset. Therefore, for some scenarios, using LexRank, which needs infinitesimally less resources than a language model, might be a good-enough, cost-effective choice. However, some LexRank summaries would probably need to be postprocessed to make them fully adequate for their use case, especially if a coherent text is expected. Using NeMo as a summarization tool, on the other hand, would be the safer choice to obtain good results. Nevertheless, NeMo would have to be used with caution: the fact that it mostly writes grammatically correct and coherent summaries (often undistinguishable from human-written summaries) should not lead to think that these summaries cannot contain any factual mistakes or hallucinations.

It was also observed that correlation between ROUGE scores and manual evaluation was very weak, and that reference summaries of different datasets are not always comparable. Different news summarization datasets, for example, have different average reference summary lengths, which might lead to different information coverages in reference summaries, and therefore to different ROUGE scores when these reference summaries are used to evaluate candidate summaries.

In conclusion, this thesis has shown that defining the purpose and the constraints of summaries is helpful for summary evaluation, and has highlighted some problems related to summarization datasets and evaluation metrics. Future research related to this work could focus on developing automatic evaluation metrics that correlate with human judgment better than ROUGE, or on creating high-quality summarization datasets in domains other than the news domain.

Bibliography

ATILF, INIST, LINA, LIDILEM, INRIA NGE, and INRIA Saclay (2017). Termith (terminology and indexation of full scientific papers in humanities and social sciences). ORTOLANG (Open Resources and TOols for LANGuage), available at: `www.ortolang.fr`, last access: 28th January 2025.

Aumiller, D., Fan, J., and Gertz, M. (2023). On the state of german (abstractive) text summarization. *arXiv preprint arXiv:2301.07095*.

Aumiller, D. and Gertz, M. (2022). Klexikon: A german dataset for joint summarization and simplification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2693–2701.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M., Kamran, A., Kirefu, F., Koehn, P., et al. (2020). Paracrawl: Web-scale acquisition of parallel corpora. Association for Computational Linguistics (ACL).

Bommasani, R. and Cardie, C. (2020). Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chen, Y., Liu, Y., Chen, L., and Zhang, Y. (2021). Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.

Dang, H. T. (2005). Overview of duc 2005 (draft). In *Proceedings of Document Understanding Conferences*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Drury, A., Pape, E., Dowling, M., Miguel, S., Fernández-Ortega, P., Papadopoulou, C., and Kotronoulas, G. (2023). How to write a comprehensive and informative research abstract. In *Seminars in Oncology Nursing*, volume 39, page 151395. Elsevier.

Eddine, M. K., Tixier, A., and Vazirgiannis, M. (2021). Barthez: a skilled pretrained french sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390. Association for Computational Linguistics.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). CCAligned: A massive collection of cross-lingual web-document pairs. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. 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.

Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

Fabbri, A. R., Wu, C.-S., Liu, W., and Xiong, C. (2022). Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Galliers, J. R. and Spärck Jones, K. (1993). Evaluating natural language processing systems. Technical report, University of Cambridge, Computer Laboratory.

Gehrmann, S., Clark, E., and Sellam, T. (2023). Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Giarelis, N., Mastrokostas, C., and Karacapilidis, N. (2023). Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*, 13(13):7620.

Goel, K., Rajani, N. F., Vig, J., Taschdjian, Z., Bansal, M., and Ré, C. (2021). Robustness gym: Unifying the nlp evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55.

Goldsack, T., Zhang, Z., Lin, C., and Scarton, C. (2022). Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604.

Gong, Y. and Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics.

Guo, M., Ainslie, J., Uthus, D. C., Ontanon, S., Ni, J., Sung, Y.-H., and Yang, Y. (2022). Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736.

Gupta, S. and Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., Mirjalili, S., et al. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. (2021). Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Huang, L., Cao, S., Parulian, N., Ji, H., and Wang, L. (2021). Efficient attentions for long document summarization. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, pages 1419–1436. Association for Computational Linguistics (ACL).

Jung, T., Kang, D., Mentch, L., and Hovy, E. (2019). Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3324–3335.

Klimt, B. and Yang, Y. (2004). The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

Koopman, P. (1997). How to write an abstract.

Kornilova, A. and Eidelman, V. (2019). Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., and Radev, D. (2022). Booksum: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558.

Landro, N., Gallo, I., La Grassa, R., and Federici, E. (2022). Two new datasets for italian-language abstractive text summarization. *Information*, 13(5):228.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin, C.-Y., Zhou, L., and Fukumoto, J. (2006). Automated summarization evaluation with basic elements. In *Proceedings of the 5th international conference on language resources and evaluation*.

Liu, F. and Liu, Y. (2009). Exploring correlation between rouge and human evaluation on meeting summaries. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):187–196.

Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Lloret, E., Plaza, L., and Aker, A. (2018). The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52:101–148.

Lloret, E., Romá-Ferri, M. T., and Palomar, M. (2013). Compendium: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*, 88:164–175.

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

Luo, M., Xue, B., and Niu, B. (2024). A comprehensive survey for automatic text summarization: techniques, approaches and perspectives. *Neurocomputing*, page 128280.

Maybury, M. T. (1995). Generating summaries from event data. *Information Processing & Management*, 31(5):735–751.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Mistral, A. (2024). Mistral nemo. Available at: `https://mistral.ai/news/mistral-nemo/`, last access: 11th January 2025.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2).

Nenkova, A. and Passonneau, R. J. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pu, X., Gao, M., and Wan, X. (2023). Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Rios, A., Spring, N., Kew, T., Kostrzewa, M., Säuberli, A., Müller, M., and Ebling, S. (2021). A new dataset and efficient baselines for document-level text simplification in German. In Carenini, G., Cheung, J. C. K., Dong, Y., Liu, F., and Wang, L., editors, *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161. Association for Computational Linguistics.

Schulte, M. and van Dijk, Z. (2015). Free children's encyclopedia project.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2020). Mlsum: The multilingual summarization corpus. In *2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067. Association for Computational Linguistics.

Sharma, E., Li, C., and Wang, L. (2019). Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Shimada, A., Okubo, F., Yin, C., and Ogata, H. (2017). Automatic summarization of lecture slides for enhanced student previewy. *IEEE Transactions on Learning Technologies*, 11(2):165–178.

Spärck Jones, K. (1999). Automatic summarizing: factors and directions. *Advances in Automatic Text Summarization*.

ter Hoeve, M., Kiseleva, J., and de Rijke, M. (2020). What makes a good summary. *Reconsidering the Focus of Automatic Summarization. CoRR, abs/2012.07619*.

Tratz, S. and Hovy, E. (2008). Bewte: basic elements with transformations for evaluation. In *Proceedings of the 1st text analysis conference*.

Uthus, D., Ontanon, S., Ainslie, J., and Guo, M. (2023). mlongt5: A multilingual and efficient text-to-text transformer for longer sequences. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, H., Li, J., Wu, H., Hovy, E., and Sun, Y. (2023). Pre-trained language models and their applications. *Engineering*, 25:51–65.

Xue, L., Constant, N., Roberts, A., Mihir, K., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498. Association for Computational Linguistics (ACL).

Yang, G., Chen, N.-S., Sutinen, E., Anderson, T., Wen, D., et al. (2013). The effectiveness of automatic text summarization in mobile learning contexts. *Computers & Education*, 68:233–243.

Zhang, S., Celikyilmaz, A., Gao, J., and Bansal, M. (2021). Emailsum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhu, C., Yang, Z., Gmyr, R., Zeng, M., and Huang, X. (2021). Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1462–1471.

Appendix A

Datasets

Table 49: Example of summary that contains information not given in the text (highlighted in red).

Text ID	20min-1018 (in preprocessed dataset in GitLab)
Summary	Die Scroll-Bewegung ist aus unserem Alltag nicht mehr wegzudenken. Tatsächlich legt ein menschlicher Daumen im Schnitt pro Tag rund 229 Meter allein mit Scroll-Bewegungen zurück. Rechnet man dies aufs Jahr hoch, scrollt ein Daumen über die Distanz von zwei Marathons hinweg. Dies kann zu Entzündungen und Schmerzen führen.
Article	Social-Media-Websites, E-Mail-Konten und News-Outlets haben eines gemeinsam: Wenn man über ein Smartphone auf sie zugreift, muss man sich vertikal durch sie hindurch scrollen. Tatsächlich verbringen Menschen im Durchschnitt rund 49 Minuten pro Tag allein damit, durch Social-Media-Plattformen zu stöbern, wie "Dailymail" berichtet. Vor diesem Hintergrund hat das Marketingunternehmen Ilk nun ausgerechnet, wie gross die Distanz ist, die ein Daumen pro Jahr beim Scrollen ungefähr zurücklegt. Das Resultat: Es ist die Strecke von etwa zwei Marathons. "Die schiere Distanz, die wir mit unserem Daumen auf dem Smartphone zurücklegen, ist erschütternd", so der PR-Chef von Ilk. "Wir wollten damit aufzeigen, welche grosse Rolle Social Media in unserem Leben spielt, egal ob wir uns damit unterhalten wollen, Informationen suchen oder mit der Familie und mit Freunden in Kontakt bleiben". Wie auch beim Training für einen Marathon besteht beim Scrollen durch endlose Social-Media-Seiten eine Verletzungsgefahr. Denn die Scroll-Bewegung ist laut Eugene Y. Tsai, Spezialarzt beim Cedars-Sinai Hospital in Los Angeles, eine unnatürliche Geste. "Die Sehne im Daumen kann durch die repetitive Bewegung entzündet werden, da sie wiederholt gegen den Tunnel, in welchen sie eingebettet ist, gerieben wird", so der Arzt. Daher wurden in den vergangenen Jahren immer mehr Fälle des sogenannten "Scroller Daumens" oder "Texting Daumens" verzeichnet. Dabei verkrampft sich der Finger, entzündet sich und bleibt schliesslich in einer gebogenen Position stecken. Aber nicht nur der Daumen leidet von der wiederholten Scrolling-Bewegung. Während hauptsächlich die Benutzung von kleineren Smartphones zum "Scrolling Daumen" führt, sind grössere Handys aufgrund des schwereren Gewichts für Schäden am Handgelenk und den restlichen Fingern verantwortlich. So kann es beispielsweise bei Kindern und Jugendlichen, die täglich stundenlang Smartphones in der Hand halten, am kleinen Finger, auf welchem das Handy ruht, zu Deformationen kommen. Wer sich um seinen Daumen sorgt oder gar bereits erste Anzeichen eines "Scrolling Daumens" verspürt, kann eine Verschlimmerung aber noch abwenden. So wird geraten, ab und zu einen anderen Finger für die Scrolling-Bewegung zu benutzen. Ausserdem können viele Handys heutzutage über Stimmbefehle kontrolliert werden. So können längere Text-Nachrichten oder E-Mails auch einfach diktiert anstellt getippt werden.
Source	https://www.sueddeutsche.de/sport/kristina-vogel-querschnittsgelaehmt-1.4120963

Table 50: Example of a PubMed article in which a sentence is missing (highlighted in red).

Text ID	pubmed-1 (in preprocessed dataset in GitLab)
Text in PubMed dataset	[...] ten patients in this study (5 pd with anxiety; 5 pd without anxiety) were taking psychotropic drugs (i.e., benzodiazepine or selective serotonin reuptake inhibitor). patients were also excluded if they had other neurological disorders, psychiatric disorders other than affective disorders (such as anxiety), or if they reported a score greater than six on the depression subscale of the hospital anxiety and depression scale (hads). [...]
Original paper	[...] Ten patients in this study (5 PD with anxiety; 5 PD without anxiety) were taking psychotropic drugs (i.e., benzodiazepine or selective serotonin reuptake inhibitor). Patients with an MMSE of less than 24 were excluded. Patients were also excluded if they had other neurological disorders, psychiatric disorders other than affective disorders (such as anxiety), or if they reported a score greater than six on the depression subscale of the Hospital Anxiety and Depression Scale (HADS). [...]
Source	https://pmc.ncbi.nlm.nih.gov/articles/PMC5075302/pdf/NRI2016-6254092.pdf

Figure 1: Website containing an introduction which is separated from the rest of the text.

Source: Screenshot of https://www.eca.europa.eu/en/the-eu-finances, last access: 10th January 2025.

Table 51: An unprocessed email from the Enron dataset.

Text ID	emails-unpr-4 (in dataset in GitLab)
Article	Message-ID: <29017619.1075861029387.JavaMail.evans@thyme> Date: Fri, 8 Mar 2002 08:59:08 -0800 (PST) From: robert.badeer@enron.com To: ina.rangel@enron.com Subject: RE: Badge Access Mime-Version: 1.0 Content-Type: text/plain; charset=us-ascii Content-Transfer-Encoding: 7bit X-From: Badeer, Robert </O=ENRON/OU=NA/CN=RECIPIENTS/CN=RBADEER> X-To: Rangel, Ina </O=ENRON/OU=NA/CN=RECIPIENTS/CN=Irangel> X-cc: X-bcc: X-Folder: \Robert_Badeer_Mar2002_1\Badeer, Robert\Sent Items X-Origin: Badeer-R X-FileName: rbadeer (Non-Privileged).pst thanks Ina —–Original Message—– From: Rangel, Ina Sent: Thursday, March 07, 2002 12:56 PM To: Badeer, Robert Subject: FW: Badge Access When you get here on Monday morning, come to the 5th floor reception of the new building. If your badge is not there, then I will come and pick you up when you get here and bring you up. Your badge will be ready Monday for sure, whether it be morning or afternoon I am not sure of. -Ina —–Original Message—– From: Curless, Amanda Sent: Thursday, March 07, 2002 2:50 PM To: Rangel, Ina Subject: RE: Badge Access Ina, We can most likely have this by Monday morning and he can pick this up at the 5th floor reception. If he has any problems he can call me. Thanks! Mandy —–Original Message—– From: Rangel, Ina Sent: Thursday, March 07, 2002 2:39 PM To: Curless, Amanda Subject: RE: Badge Access << File: Badge Access Form.doc >> I filled out all of the information that I had on him. Will he be able to have his badge by Monday morning and where will he go to pick it up. Ina —–Original Message—– From: Curless, Amanda Sent: Thursday, March 07, 2002 2:00 PM To: Rangel, Ina Subject: Badge Access « File: Badge Access Form.doc » Ina, Pleae fill out and return to me at ECS 05848. You can e-mail this to me if this is easier. Thanks! Mandy

Table 52: The same email as in Table 51, but preprocessed.

Text ID	emails-prepr-4 (in dataset in GitLab)
Article	Ina, Pleae fill out and return to me at ECS 05848. You can e-mail this to me if this is easier. Thanks! Mandy ——— << File: Badge Access Form.doc >> I filled out all of the information that I had on him. Will he be able to have his badge by Monday morning and where will he go to pick it up. Ina ——— Ina, We can most likely have this by Monday morning and he can pick this up at the 5th floor reception. If he has any problems he can call me. Thanks! Mandy ——— When you get here on Monday morning, come to the 5th floor reception of the new building. If your badge is not there, then I will come and pick you up when you get here and bring you up. Your badge will be ready Monday for sure, whether it be morning or afternoon I am not sure of. -Ina ——— thanks Ina << File: Badge Access Form.doc >>

Appendix B

NeMo Prompts

As can be seen in Table 53, NeMo prompts are always written in the same language as the input text. Moreover, since it was observed than NeMo sometimes uses English or other foreign languages in summaries of non-English texts, prompts for non-English data always specify what the target language is.

For news, web snippets and email scenarios, summary length was also specified in the prompts. However, NeMo often does not respect the specified length requirements, as observed during summary evaluation (e.g. Section 5.2.1, 5.2.2, 5.2.3).

Table 53: NeMo prompts per scenario and per language.

Scenario	**Language**	**Prompt**
News	EN	Summarize the following text in maximum 3 sentences
	DE	Fassen Sie den folgenden Text in maximal drei kurzen Saetzen zusammen
	FR	Résumez le texte suivant en français, en 3 phrases maximum
Papers	EN	Summarize this text in no more than 8 sentences
	FR	Résumez le texte suivant en français
Web	EN	Summarize the following text in two short sentences
	DE	Fassen Sie den folgenden Text in zwei kurzen Sätzen auf Deutsch zusammen
	FR	Résumez le texte suivant en français en deux courtes phrases
	IT	Riassumi il testo seguente in due brevi frasi in italiano
School	DE	Fassen Sie den folgenden Text auf Deutsch zusammen
	FR	Résumez le texte suivant en français
Emails	EN	Summarize the following email thread in one or two sentences

Appendix C

Results

Table 54: Example of LexRank summary with incorrect sentence segmentation. The sentence highlighted in red is incomplete, because the dot after "B.1.1.7." in the original text has been interpreted as a full stop signaling the end of a sentence. The corresponding sentence in the original article is highlighted in green.

Text ID	20min-320 (in GitLab)
Summary	Sorgen machen den Experten weiterhin vor allem die neuen Varianten. hat kontinuierlich zugenommen", so Martin Ackermann, Präsident der Corona-Taskforce. Wie auch letzte Woche sehe die epidemiologische Lage "unsicher" aus.
Article	Nachdem die Corona-Zahlen in zahlreichen Ländern Europas wieder ansteigen (siehe Box), gab am Dienstag die wissenschaftliche Corona-Taskforce des Bundes eine Beurteilung der epidemiologischen Lage in der Schweiz ab. Sorgen machen den Experten weiterhin vor allem die neuen Varianten. "Die ansteckendere Variante B.1.1.7. hat kontinuierlich zugenommen", so Martin Ackermann, Präsident der Corona-Taskforce. Mittlerweile seien über 70 Prozent aller neuen Covid-Fälle Mutationen, ergänzte Masserey vom Bundesamt für Gesundheit. Und weiter: "Die aktuellen Zahlen sind stagnierend oder leicht steigend", so Masserey. Wie auch letzte Woche sehe die epidemiologische Lage "unsicher" aus. Das BAG meldete am Dienstag eine wieder etwas höhere Positivitätsrate von 5,7 Prozent und einen durchschnittlichen R-Wert der letzten sieben Tage von 1,08 - nur in vier Kantonen lag der R-Wert noch unter 1. Weitere Öffnungen stellte der Bundesrat zuletzt für den 22. März in Aussicht. Dafür hat er vier epidemiologische Richtwerte definiert, die er am 19. März unter die Lupe nehmen wird. Wie bereits letzte Woche erfüllt die Schweiz immer noch nur einen der vier vorgegebenen Werte, die Anzahl Covid-Patienten auf der Intensivstation. Fallzahlen, Positivitätsrate und R-Wert erfüllen die Anforderungen des Bundesrates noch nicht. Man müsse eine günstige epidemiologische Situation schaffen, damit Öffnungen gut möglich seien, sagte Rudolf Hauri, Präsident der Vereinigung der Kantonsärztinnen und Kantonsärzte. Umso wichtiger seien jetzt die geltenden Schutzmassnahmen. "Es ist unerlässlich, die Schutzmassnahmen weiterhin und auch noch lange ganz konsequent einzuhalten". Virginie Masserey gab einen kleinen Hinweis zu den nächsten Öffnungsschritten. "Die geringsten Risiken sind da, wo man Distanz wahren kann und wo Maske getragen werden kann". Daneben spielten weitere Faktoren wie etwa Anzahl Personen, Personendichte und Ort eine wichtige Rolle. Viel Hoffnung setzen die Experten des Bundes auf die Impfungen. Auch wenn die Auslieferungen bisher mit vielen Problemen verbunden waren: Bis Ende Juni sollen alle, die geimpft sein wollen, auch geimpft sein, versprach BAG-Vizedirektorin Nora Kronig erneut. Derzeit sei die Zahl der Geimpften allerdings noch zu klein, um einen grösseren Einfluss auf die Infektionszahlen zu haben. Martin Ackermann glaubt allerdings, dass das bald möglich sein könne. "Mit den Massentests können wir die Pandemie besser managen, mit der Impfung können wir sie beenden".

Table 55: Example of mLongT5 summary with a pronoun with unclear referent (highlighted in blue).

Text ID	newsroom-748 (in GitLab)
Summary	Donald Trump is the 47th president of United States, and he's a billionaire. But his predecessor, Barack Obama, has made it clear that he would spend first week in Oval Office "reversing everything that we've done" before election day. The New York Times reports that as soon as President Trump gets under the White House on January 21, many policy passions have been pursued by him. Among them are efforts to reduce climate change, introduce gun control, and build alliances.

Table 56: Example of mLongT5 summary containing a conjunction used in a misleading way (highlighted in blue).

Text ID	cnndm-2366 (in GitLab)
Summary	McDonald's is raising wages to at least 15% of its employees in the US, and it's not going to help workers get a voice in their workplace conditions, writes Martin Luther King. "We are at an turning point in our history," he writes on CNN. "We are at a third reconstruction." The company says it will raise wages up to $1 per hour more than local minimum wages, but that only affects about 10-15% of restaurants in the US.

Table 57: Example of NeMo summary containing incorrect information (highlighted in red). The correct information is highlighted in green in the original text.

Text ID	newsroom-1418 (in GitLab)
Summary	In pre-digital America of 1960, news articles metaphorically described student sit-in movements as spreading like wildfire across Southern black campuses. On Feb. 1, 1968, four black freshmen at North Carolina A&T State University initiated a non-violent protest at Woolworth's lunch counter in Greensboro, sparking immediate interest among students from other historically black colleges nearby. This event swiftly united traditional civil rights leaders with thousands of student activists who became shock troops for the broader movement within weeks. Despite initial disagreements, the alliance proved powerful, dismantling segregation within three years. Reflecting on this past, we can consider how today's media environment might impact such a movement differently than the more unified, fact-based reporting prevalent in the 1960s.
Article	IN the pre-digital America of 1960, "viral" was still a medical term. So it was written in countless news articles that the student sit-in movement had "spread like wildfire" on black campuses across the South. On the morning of Feb. 1, 50 years ago today, four black freshmen at North Carolina A&T State University seated themselves at the all-white lunch counter in a Woolworth's dime store in Greensboro. [...]

Table 58: Example of minimal information addition in a NeMo summary. The original text only talks of "Hachette", whereas NeMo, in its summary, uses the complete name of the company, "Hachette Book Group". All mentions of the company name are highlighted through bold text.

Text ID	newsroom-91 (in GitLab)
Summary	Amazon, in its pursuit of retail dominance, has sparked controversies with **Hachette Book Group** over e-book pricing and Disney due to contractual disagreements, leading to boycotts. Critics argue Amazon's power is hurting industries, with authors rallying against its tactics; meanwhile, competitors like Google and Barnes & Noble are aligning against Amazon.
Article	In its storied, revolutionary history, Amazon.com hasn't been hesitant to employ ruthless strategies in its quest to rule retail. The company's tactics have been so tough that they've inspired consumer boycotts from time to time. Amazon's latest skirmishes position the world's largest e-retailer in standoffs against **Hachette**, a book publisher being pressured to lower its prices, and Disney, which failed to reach some contractual agreements with Amazon, and which is being punished by Amazon's refusal to sell preorders of some of its movies. We thought it would be helpful—or at least a heckuvalot more fun—to explain more about the ongoing disputes using classic quotes from Disney films and Marvel Comics, which Disney also owns. "I just can't wait to be king." These words, sung by Simba in Disney's "The Lion King," sum up the ambitions of Jeff Bezos and Amazon: The goal is to be the undisputed king of selling us stuff. As soon as possible, naturally. From one-click ordering to Amazon Prime, and from it forays into everything from groceries to a phone that encourages users to shop more at Amazon, it's clear that Amazon wants to be the Everything Store—and to so thoroughly dominate the world of e-commerce that it essentially takes over the retail world. In any attempted coup, the grab for money and power can be ugly. Often, the subjects aren't happy with the policies and terms dictated by the new ruler, especially when they question the legitimacy of the king. In this case, Disney, **Hachette**, and others are the subjects that aren't happy with how the self-appointed new ruler is trying to push them around. "If your heart is in your dream, no request is too extreme." Jiminy Cricket said these words to Pinocchio, who dreamed of being a real boy. Amazon's dream is different—to be the real boss of retail. To make Jeff Bezos's wish come true, Amazon has been making some fairly extreme requests, including an insistence than **Hachette** cap its e-book prices at $9.99. Amazon is also using some extreme negotiating tactics in its standoff with Disney, notably making it difficult or impossible for customers to pre-order some of the company's highly anticipated movies, including "Maleficent," "Muppets Most Wanted," and "Captain America: The Winter Soldier." "I am but a humble servant of my people!" Amazon's justification for playing hardball with movie companies and book publishers is that it is merely fulfilling its mission to serve its customers best, by way of figuring out how to offer them the absolute lowest prices possible. "We will never give up our fight for reasonable e-book prices," Amazon said in a recent statement, regarding its ongoing dispute with **Hachette**. "We know making books more affordable is good for book culture." [...]

Table 59: Example of NeMo summary containing additional information (highlighted in red): the information on Francesco Schettino, although correct, is not contained in the original text. For space reasons, the original article is not reported completely, but it can be found in GitLab, or at the link in field *Source*.

Text ID	newsroom-2487 (in GitLab)
Summary	**Summary:** - The Costa Concordia cruise ship struck a reef off Italy's Giglio Island on January 13, causing it to capsize. - Passengers reported chaos during the evacuation: lifeboats weren't stabilized initially, there were no evacuation drills, and survivors weren't counted upon reaching land. - A total of 4,200 passengers and crew were aboard; about 600 are still unaccounted for as of January 15. - The captain, Francesco Schettino, is being questioned by prosecutors. He allegedly abandoned the ship while passengers were still on board. - About 30 people were injured, with at least two in grave condition. - Survivors are being housed in schools and hotels on Giglio Island and nearby Porto Santo Stefano, awaiting transport home. - The Concordia had a previous minor accident in Palermo in 2008 but no injuries were reported.
Article	PORTO SANTO STEFANO, Italy – Survivors who escaped a luxury cruise liner that ran aground and tipped over off Italy's coast recounted a chaotic and terrifying evacuation through the ship's upended hallways on Saturday, as divers searched the submerged part for any people still unaccounted for in the confusion. Three bodies were recovered from the sea after the Costa Concordia with 4,234 people aboard ran aground hundreds of yards off the tiny island of Giglio near the coast of Tuscany late Friday, tearing a 160-foot gash in its hull and sending in a rush of water. As authorities and port officials carefully matched names on the cruise ship's list of passengers and crew with those of survivors getting off ferries or other boats on the mainland, the number of the unaccounted for steadily dropped to roughly 40. Italian news agency ANSA says 2 survivors found in cabin late Saturday. Among the 4, 234 passengers the U.S. State Department estimates 126 to be U.S. citizens, Fox News confirms. At this time the American passengers have not been listed among those missing or injured. Passengers described a scene reminiscent of "Titanic", saying they escaped the ship by crawling along hallways, desperately trying to reach safety as the lights went out and plates and glasses crashed around them. Helicopters whisked some survivors to safety, others were rescued by private boats in the area, and witnesses said some people jumped from the ship into the dark, cold sea. At darkness, the diving operations, involving fire department divers and coast guard divers, was suspended for the night, coast guard officials said. While only a small section of the submerged area was inspected, no signs of any survivors or victims were immediately found, said Capt. Emilio Del Santos, of the port captain's office in Livorno. Helicopters and sea searches of the area were continuing. The ship was lying virtually flat off Giglio's coast, its starboard side submerged in the water and the huge gash showing clearly on its upturned hull. [...]
Source	https://web.archive.org/web/20120115030602id_/http://www.foxnews.com/world/2012/01/13/cruise-ship-reportedly-runs-aground-off-coast-italy-at-least-6-dead/

Table 60: Example of NeMo summary containing a hallucination (highlighted in red): the website mentioned in the summary is not mentioned in the original text, and it does not exist either. For space reasons, the original article is not reported completely, but it can be found in GitLab, or at the link in field *Source*.

Text ID	newsroom-1517 (in GitLab)
Summary	**Summary of St. Lucia Travel Information:** 1. **Package Holidays:** - Upmarket hotels: Carrier, Elegant Resorts, ITC Classics, Seasons. - Wider price range: British Airways, Kuoni, Trailfinders, Tropical Sky, Virgin Holidays. [...] 8. **Safety and Beach Access:** - Violent crimes can occur; be cautious on isolated beaches and at street parties. - All St. Lucia beaches are public but access may be restricted due to recent hotel developments. - It's illegal for anyone to wear camouflage clothing. **Source:** [Summary of information from summarropicbreeze.co.uk](http://summarropicbreeze.co.uk)
Article	For the scenery as much as to flop on the beach. Much of St Lucia is mountainous and covered in a blanket of thick rainforest, and the south-west corner has arguably the most arresting sight in the whole Caribbean: of the green and precipitous volcanic cones of the Petit Piton and Gros Piton rising out of a blue ocean. That is not to say that St Lucia doesn't have some lovely beaches - they are golden in the north, and mostly silvery in the volcanic south. It's just that they are rather outshone by the island's other natural assets. St Lucia is also as good a choice as anywhere in the Caribbean for secluded, upmarket, romantic places to stay. Your hotel room might have an outdoor garden shower, a private plunge pool, a hammock for two on its balcony, and a view of the thrusting Pitons. Nature and fertility will be all around you. Several hotels have featured prominently in the US and, more recently, UK versions of The Bachelor television series, and wherever you go, you trip over wedding ceremonies and loved-up honeymooning couples. That said, a week on St Lucia doesn't have to be all about candlelit meals for two every night. There are also good family-oriented hotels, and you can party the night away in Rodney Bay Village, the island's only resort. Though St Lucia was badly hit by Hurricane Tomas in late October 2010 and you can still see the effects of the storm in roadside landslips, from a visitors' point of view everything is back to normal. Peak season on St Lucia runs from mid-December to April. In these months, accommodation rates are generally significantly higher than at other times of the year, but the weather is usually at its best. Rates are cheaper in the summer and autumn months, when it's stickier and wetter - and note that St Lucia gets more rain than other less mountainous Caribbean islands such as Barbados or Antigua. Also bear in mind that the hurricane season runs from June to November, with September and October statistically the likeliest months for major storms. Taking into account prices and weather, May is a good month to visit - and the island's major annual jazz festival is taking place in 2013 (May 4-8). Two airlines fly non-stop from the UK to St Lucia's Hewanorra International Airport, in the far south of the island: British Airways (0844 493 0787; www.ba.com) and Virgin Atlantic (0844 209 7777; www.virgin-atlantic.com), both from Gatwick. Fares vary considerably according to the time of year, and whether or not the airlines are offering discounted deals. You can also book flights through agents such as Trailfinders (020 7368 1200; www.trailfinders.com), and online through www.expedia.co.uk. Inter-island flights - for example, with Liat (www.liat.com) - arrive at the much smaller George FL Charles Airport just north of Castries. Transfers from Hewanorra airport to hotels in the north of the island are long - around 90 minutes - and to the Soufrière area take around 40 minutes. [...]
Source	https://web.archive.org/web/20121210191d_/http://www.telegraph.co.uk/travel/destination/stlucia/49765/St-Lucia-travel-guide.html

Table 61: Example of NeMo summary containing too much *interpretation* (highlighted in red): we do not know if the recipe is "a popular choice", we only know that the author of the article chose to try it because they found it interesting (highlighted in green in the original text).

Text ID	newsroom-1032 (in GitLab)
Summary	Tom Faglon creates unique dark chocolate and pomegranate bark with ginger and sea salt for the holidays, which is both sophisticated and easy to make despite its short shelf life. The recipe involves scattering fresh pomegranate seeds, chewy candied ginger bits, and crunchy sea salt on top of bittersweet chocolate. It's a popular choice among homemade holiday gifts along with Elizabeth Choinski's spiced pecan pralines and Corey Balazowich's cranberry cordial.
Article	Dark Chocolate and Pomegranate Bark: Tom Faglon makes homemade gifts for the holidays. His gift to Melissa Clark: a dark chocolate and pomegranate bark with just a hint of ginger and sea salt. Chocolate bark is the workhorse of homemade holiday gifts: easy, crowd-pleasing, expected. But you've never seen chocolate bark like Tom Faglon's chocolate bark. Instead of nuts, or dried fruit, or chopped-up peppermint candy canes, he scatters fresh pomegranate seeds over the surface, where they glisten like rubies. Chewy bits of candied ginger are hidden inside the bittersweet chocolate, lying in wait to surprise you with their spiciness. Then, to make things even more interesting, Mr. Faglon strews the top with crunchy sea salt. His bark is as sweet and juicy as a box of chocolate-covered cherries, but a lot more sophisticated. Even better for the busiest time of the year, a pound of it takes less than 10 active minutes to make. I came across Mr. Faglon's recipe as I was poring over the nearly 150 messages that Dining section readers sent to us with their favorite sweet homemade holiday gifts. Recipes came in from all over the world. There was an unusual anise-oat crescent cookie from Rome. A recipe for chocolate-dipped, marzipan-covered apricots was sent from Paris. From closer to home, Brooklyn, came a farmers'-market-inspired fruit compote. So I put on my apron and got to work, testing my way through a dozen of the offerings. My methodology for choosing what to test operated on a gut level. If my mouth started to water as I read the recipe, I printed it out. There were plenty of variations on chocolate bark, but none as intriguing as Mr. Faglon's. Mr. Faglon, a retired phone company employee living in Somerset, N.J., and his wife, Diane, blog about food, antique collecting and the adventures of their 10 cats. He started making chocolate bark many years ago, using dried cherries, almonds, pistachios and the like. He came up with this wonderful innovation a few weeks ago when he was trying to think up something new to do with chocolate bark, and he happened to have a container of pomegranate seeds in need of a home. (I can relate to this method of recipe developing.) The only downside to Mr. Faglon's recipe is that because the pomegranate seeds contain a lot of moisture, the bark isn't something you want to keep around for very long. It will start to weep after a day or so. This said, it's easy enough to make in the morning and give out that afternoon. And it will still taste good days later even if it starts to look a little tear-stained. A more stable homemade sweet that is just as tasty, if not quite as strikingly pretty, is the spiced pecan praline recipe sent in by Elizabeth Choinski of Oxford, Miss. [...]

Table 62: NeMo summary with numbered sentences.

Text ID	20min-746 (in GitLab)
Summary	1. Etwa 2500 Menschen feierten eine illegale Silvesterparty in Frankreich. 2. Bei der Auflösung der Party kam es zu gewaltsamen Ausschreitungen und Polizisten wurden verletzt. 3. Die französische Regierung hat die Ausgangssperre aufgrund der Ausbreitung des Virus erweitert.

Table 63: This NeMo summary only reports information from the first part of the article, and does not mention what most of the article (highlighted in blue) is about. For space reasons, the original article (which can be found in GitLab) is not reported completely; the part not appearing here would also be highlighted in blue.

Text ID	20min-242 (in GitLab)
Summary	Der Schweizerische Fussballverband sucht bis Ende August einen neuen Nationaltrainer, da bereits im September die WM-Qualifikation beginnt. Das Suchen eines neuen Coaches ist für den Präsidenten Dominique Blanc, Direktor Pierluigi Tami und Generalsekretär Robert Breiter eine neue Aufgabe, da sie dies seit Amtsantritt noch nie machen mussten. Die Frage, wer Nachfolger von Vladimir Petkovic wird, beschäftigt die Schweizer Fussball-Fans. In einer Umfrage gaben über 33'000 Menschen ihre Stimme ab, wobei Ex-Weltmeister-Coach Jogi Löw mit 29 Prozent die meisten Stimmen erhielt und Union-Coach Urs Fischer mit 20 Prozent folgte.
Article	Nach dem Abgang von Vladimir Petkovic braucht der Schweizerische Fussballverband (SFV) bis Ende August einen neuen Nationaltrainer. Denn: Bereits im September geht es weiter mit der WM-Quali. Und so sucht der SFV erstmals seit acht Jahren einen neuen Coach. Es ist eine neue Aufgabe für den Präsidenten Dominique Blanc, den Direktor Pierluigi Tami sowie den Generalsekretär Robert Breiter. Sie alle mussten, seit sie ihr Amt inne haben, keinen neuen Nati-Trainer suchen. Die Frage, wen sie für dieses Amt auswählen, beschäftigt die Schweizer Fussball-Fans. Zu sehen ist das etwa bei der 20-Minuten-Community. Bei der Umfrage, wer Nachfolger von Petkovic werden soll, gaben über 33"000 Menschen ihre Stimme ab. Am meisten Stimmen bekam Ex-Weltmeister-Coach Jogi Löw, erhielt er doch 29 Prozent. Union-Coach Urs Fischer folgt dahinter mit 20 Prozent. Doch welche Namen sind noch im Umlauf? Wir geben eine Übersicht, wer es noch werden könnte. Beginnen wir bei dem Mann, der auf Platz zwei landete: Urs Fischer. Seit drei Jahren steht der Zürcher an der Seitenlinie von Union Berlin. Er tut das mit derart grossem Erfolg, dass er bei einigen Clubs Begehrlichkeiten weckt. Auch beim SFV, das ist kein Geheimnis. Aber auch wenn ein Grossteil der 20-Minuten-Community Fischer gerne als Nati-Coach sehen würde, wird das wohl nicht passieren. Einerseits verlängerte der 55-Jährige zuletzt seinen Vertrag bei Union vorzeitig bis 2023. Andererseits soll er gemäss dem "Kicker" dem SFV eine Absage erteilt haben. Nicht verwunderlich. Bereits im Interview mit "20 Minuten" sagte er zuletzt: "Nati-Trainer ist für mich zurzeit kein Thema". Kommen wir zum 61-jährigen Mann, den sich die 20-Minuten-Leserinnen und Leser sehnlichst wünschen. Jogi Löw, der Mann, der Deutschland 2014 zum Weltmeister-Titel führte. 29 Prozent wünschen sich den Deutschen als Nati-Coach. Klar ist: Diese Lösung wäre nahezu eine Sensation. Denn auch wenn er mit Deutschland zuletzt nicht zu überzeugen wusste, Löw ist ein Coach, der sein Handwerk versteht. Viele Fussball-Fans werden sich beispielsweise noch an den 7:1-Sieg der Deutschen an der WM 2014 gegen Brasilien erinnern. Ob es auch eine realistische Möglichkeit ist, einen Weltmeister-Trainer für die Nati zu gewinnen? Das wird sich zeigen. Ein Pluspunkt für die Schweiz: Der 61-Jährige wohnt im Raum Freiburg, unweit der Schweizer Grenze . Und Löw hat auch Erfahrung im Schweizer Fussball. Seine letzten drei Stationen als aktiver Profispieler waren Schaffhausen, Winterthur und Frauenfeld zwischen 1989 und 1995. Der 47-jährige Weiler hat viel Erfahrung. Er stand bei verschiedenen Clubs im Ausland unter Vertrag, arbeitete in Nürnberg, Anderlecht oder bei Al Ahly. Mit dem belgischen und dem ägyptischen Verein holte er jeweils den Meistertitel. Weiler ist ein moderner Coach mit viel Selbstvertrauen. Nicht immer kommt seine Art gut an. Mit dem FC Luzern hatte er keinen grossen Erfolg, die Beziehung hielt nicht lange. Viele 20-Minuten-Leserinnen und Leser wünschen sich den Ex-Dortmund-Coach Lucien Favre als Petkovic-Nachfolger. Doch auch wenn das der Wunsch von vielen ist: Es wird wohl nicht passieren. "Lucien Favre und ich haben in diesen Tagen viele Anfragen von Journalisten erhalten. Aber wie es Lucien bereits am letzten Freitag erklärt hat, wird er in diesem Sommer keine Mannschaft übernehmen - das gilt auch für die Schweizer Nationalmannschaft", führte sein Agent Payot anfangs Woche aus. Kommt der ganz grosse Name? Wie der "Blick" schreibt, befindet sich Arsène Wenger im Kreis der Favoriten. Wenger ist einer der erfahrensten Trainer in Europa. Der 71-Jährige trainierte 22 Jahre lang Arsenal. 2018 hörte er auf. Am Ende seiner Karriere stand er 828 Mal in der Premier League an der Linie, 208 Mal in der Champions League. Derzeit steht er bei der Fifa als Direktor für globale Fussballförderung unter Vertrag. Als Hitzfeld als Nati-Trainer aufhörte, wollte der SFV Koller als Coach haben. Doch der damalige Wunschkandidat sagte ab, erfüllte seinen Vertrag als Trainer von Österreich. Und jetzt? Nun die Lage, sie ist anders. Der 60-jährige Koller trainiert derzeit keinen Verein und hätte Zeit. Das Problem: Koller lässt einen ganz anderen Fussball als Petkovic spielen. Während der neue Bordeaux-Coach stets offensiv spielen liess, präferiert Koller einen eher defensiveren Fussball. Raphael Wicky ist Trainer bei Chicago Fire. Davor arbeitete der 44-Jährige als Basel-Coach, schaffte mit den Baslern eine grandiose Champions-League-Saison. Dann wurde er nach einem schwachen Saisonstart entlassen. Wicky steht aber für einen offensiven Fussball, der zur aktuellen Nati passen würde. [...]

Table 64: Example of NeMo summary containing a factual mistake ("75", highlighted in red – the correct number, 72, is highlighted in green in the original text. Moreover, this summary contains a hallucination (sentence highlighted in red).

Text ID	orangesum-1625 (in GitLab)
Summary	La propagation du coronavirus en France a entraîné une diminution de l'activité grippale dans la plupart des régions métropolitaines, avec un pic épidémique atteint probablement la semaine dernière. Toutefois, l'épidémie continue de sévir en Guyane et à Mayotte. Le virus de la grippe a causé 744 cas graves admis en réanimation depuis le 4 novembre 2019, avec une baisse significative des consultations liées à la grippe (-22%) la semaine dernière. La vaccination contre la grippe est recommandée pour les personnes présentant des facteurs de risque de complications. Le nombre de décès liés à la grippe saisonnière s'élève à 75 cet hiver en France métropolitaine, dont 10 enfants de moins de 15 ans.
Article	Si le coronavirus se propage dans l'Hexagone, le virus de la grippe, lui, reflux. Dans son dernier bilan publié mercredi 4 mars, Santé publique France relève une "diminution de l'activité grippale dans la majorité des régions en métropole". L'Île-de-France est passée en phase post-épidémique et le pic épidémique a probablement été atteint dans toutes les régions de métropole la semaine dernière, souligne l'agence sanitaire. En revanche, l'épidémie se poursuit en Guyane et touche Mayotte. Santé : quelles différences entre la grippe et le Covid-19 ? par franceinfoDepuis le 4 novembre 2019, 744 cas graves de grippe admis en réanimation en métropole ont été signalés. La semaine dernière, 47 malades ont été admis en réanimation contre 71 la semaine d'avant. La majorité (74%) de ces patients présentaient des facteurs de risque de complications, et quasi autant (69%) n'étaient pas vaccinés parmi ceux pour lesquels ce renseignement a pu être obtenu. Au total, 72 personnes sont décédées de la grippe saisonnière cet hiver en métropole, dont 10 enfants de moins de 15 ans. Trois décès en réanimation ont également été répertoriés en Martinique et deux en Guadeloupe. Les consultations liées à la grippe ont baissé significativement (-22%) la semaine dernière par rapport à la semaine précédente. Le réseau de surveillance Sentinelles a relevé un taux de consultations de 175 pour 100.000 habitants contre 225 pour 100.000 la semaine précédente. SOS Médecins a également relevé une baisse de 11% de ses consultations pour syndrome grippal. Le nombre de passages aux urgences pour grippe ou syndrome grippal rapporté par le réseau de surveillance Oscour a également "nettement diminué" (-14%) par rapport à la semaine précédente (6.100 contre 7.132) ainsi que le nombre d'hospitalisations (724, soit - 11%). Parmi les hospitalisations pour grippe après passages aux urgences, les plus représentés restent les enfants de moins de 5 ans (25%) et les personnes âgées de 75 ans et plus (29%).

Table 65: Example of NeMo summary with improvable information coverage. The summary talks about a speech that Macron held on work ethics, but it does not mention in which occasion he held it. This missing information (highlighted in green in the original text) is not strictly necessary, but it would still help contextualize the speech.

Text ID	orangesum-960 (in GitLab)
Summary	Emmanuel Macron a déclaré que les troubles en France sont parfois dus au manque de sens de l'effort chez certains citoyens. Il a également estimé que les droits individuels doivent être équilibrés par des devoirs et un engagement au travail pour maintenir la force, la cohésion et l'histoire de la France. Le président a fait l'éloge de l'apprentissage et de l'engagement professionnel, notamment dans le domaine de la boulangerie artisanale, qui représente 33 000 établissements en France. Enfin, il a critiqué les actes de vandalisme contre les commerces durante les manifestations des "gilets jaunes", soulignant que ces professionnels jouent un rôle important dans le lien social et l'équilibre des communautés locales.
Article	C'est une phrase qui risque de jeter de l'huile sur le feu. Emmanuel Macron a estimé vendredi 11 janvier que "les troubles que notre société traverse sont parfois dus au fait que beaucoup trop de nos concitoyens" ont perdu "le sens de l'effort".“Notre jeunesse a besoin qu'on lui enseigne un métier et le sens de cet engagement qui fait qu'on n'a rien dans la vie si on n'a pas cet effort", a-t-il dit, devant les maîtres-boulangers réunis à l'Élysée pour la traditionnelle galette des rois. "Les troubles que notre société traverse sont aussi parfois dus, liés au fait que beaucoup trop de nos concitoyens pensent qu'on peut obtenir sans que cet effort soit apporté. Parfois on a trop souvent oublié qu'à côté des droits de chacun dans la République - et notre République n'a rien à envier à beaucoup d'autres - il y a des devoirs. Et s'il n'y a pas ce sens de l'effort, le fait que chaque citoyen apporte sa pierre à l'édifice par son engagement au travail, notre pays ne pourra jamais pleinement recouvrer sa force, sa cohésion, ce qui fait son histoire, son présent et son avenir." Le chef de l'État a fait l'éloge de l'apprentissage, qui "permet à chaque jeune de trouver sa place dans la société". Il enseigne "l'engagement des matins tôt et le soir tard, pour arriver à l'excellence", à une époque "où on pense qu'on peut tout apprendre en quelques jours". Il a aussi félicité les lauréats des concours de la meilleure baguette de tradition française – Laurent Encatassamy, boulanger à Saint-Paulin à La Réunion – et de la meilleure baguette de Paris, attribuée à Mahmoud M'Seddi, boulanger du XIVe arrondissement, parmi les 33.000 boulangeries artisanales en France. "C'est un maillage unique dans les territoires" de ces professionnels qui "font du lien social", a-t-il ajouté. "Notre pays, dans ces moments difficiles, dit que ce lien, il ne veut pas le voir s'abattre. Quand je vois des gens qui s'attaquent aux commerces, ils ont compris l'inverse de ce pourquoi ils se battent parfois", a-t-il critiqué, en références aux dégradations contre des commerces commises lors des manifestations des "gilets jaunes". Les boulangeries permettent "quelques minutes d'un échange quotidien qui change tout de l'équilibre d'un village, d'une ville d'un territoire", a-t-il dit, rappelant que douze millions de clients vont chaque jour dans des boulangeries pour acheter du pain et que 32 millions de baguettes sont vendues chaque jour.“Le travail n'est pas seulement un élément économique, c'est le sens qu'on donne à sa vie. Parce que ce sont des heures passées pour guetter un sourire, la satisfaction d'un client", a-t-il poursuivi, avant de partager avec les professionnels de la boulangerie deux galettes géantes "républicaines" – et donc sans fève ni roi, comme le veut la tradition depuis plus de 40 ans à l'Élysée.

Table 66: Example of LexRank summary lacking syntactic and semantic connection between the sentences.

Text ID	pubmed-487 (in GitLab)
Summary	residents who personally performed repairs of their properties as well as various skilled and unskilled laborers working in the construction and building maintenance trades were at risk for inhalation exposures to dust containing microbial and other agents during demolition, removal, and repair of flood-damaged and contaminated infrastructure and building materials [68]. the unadjusted prevalence rate ratios for each symptom or condition for those doing any restoration work versus those not doing any restoration work were calculated within smoking categories based on 2 2 contingency tables. almost 75% of the study participants reported having performed some restoration work activity after hurricane katrina (n=587), and details on the actual time spent in these activities were self-reported by 474 or 81% of this group (table 2). statistically significant elevations in prevalence rate ratios for those having done any restoration work were not observed for pneumonia, dyspnea, copd, and being below lln for any of the lung function parameters. current smokers who did restoration work showed lower overall predicted lung function compared to smokers who did not; however, multiple linear regression analysis yielded no statistically significant correlations of any of the lung function parameters with restoration work time after adjustment for smoking, gender, asthma status, and use of respiratory protective equipment. in this study, there was an observable elevation in prevalence of new onset asthma after hurricane katrina which increased with increasing quartiles of restoration work time but was not statistically significant.

Table 67: Example of NeMo summary only focusing on the results of the study, and not introducing the topic or the research question.

Text ID	pubmed-79 (in GitLab)
Summary	**Summary:** - **Physical Activity & Outdoor Play:** - No significant associations found between outdoor play/active transport/sports club membership and childhood overweight. - Children playing outside ≥1x/week had a higher overweight risk, but this was a small minority. - **Screen Time:** - Consistently and significantly associated with childhood overweight. - Dose-response relationship observed; reducing screen time to <1 hour/day could potentially reduce overweight prevalence by up to 2 percentage points. - **Diet:** - No association found between soft drink/fast food consumption and overweight. - Unexpected inverse association seen for snack consumption, likely due to selective underreporting or reverse causation. - Parents may perceive reducing snacks as a method to prevent childhood overweight. - **Socioeconomic & Maternal Factors:** - Low maternal education and maternal overweight strongly associated with child's overweight. - Behavioral risk factors (like screen time) explained part of the association between low maternal education and child's overweight but not maternal overweight. - **Subgroup Analysis:** - Screen time remained a significant factor in both high-risk (mothers with overweight/low education) and low-risk subgroups, but associations were weaker. - Potential impact of interventions promoting outdoor activities on overweight prevalence is limited. - **Future Research:** - Need for further study to assess food intake-overweight associations in observational studies due to potential reverse causation.
Original abstract	objective. to prospectively identify behavioral risk factors for childhood overweight and to assess their relevance in high risk sub groups (children of mothers with overweight or low education). methods. in the piama birth cohort (n=3963), questionnaire data were obtained at ages 5 and 7 on screen time, walking or cycling to school, playing outside, sports club membership, fast food consumption, snack consumption and soft drink consumption. weight and height were measured at age 8 years. results. screen time, but none of the other hypothesized behavioral factors, was associated with overweight (aor 1.4 (ci: 1.21.6)). the adjusted population attributable risk fraction for screen time>1 hr/day was 10% in the high risk and 17% in the low risk sub groups. conclusion. reduction of screen time to<1 hr/day could result in a reduction of overweight prevalence in the order of 2 percentage points in both high and low risks sub groups.

Table 68: Example of NeMo summary focusing on the quantitative results of a paper (highlighted in blue both in the summary and in the original abstract) more than on the qualitative ones (green), even though the latter play a bigger role in the original text. Also notice that the summary is in English even though the original paper is in French.

Text ID	termith-772-scientext (in GitLab)
Summary	This text discusses a reception test conducted with thirty engineering students at the University of Technology in Troyes, France. The aim was to gather insights into how viewers perceive and understand emotional cues in media. Here's a summary of the key points: 1. **Test Design**: A short video sequence was shown without any prior commentary, followed by a questionnaire consisting of three questions. Participants had to identify the primary emotion expressed by the main character (a young girl), list five secondary emotions, and indicate five indices used to identify these emotions, ranked in order of importance. 2. **Primary Emotion**: Most participants (20 out of 30) identified the primary emotion as despair. Other responses included disgust, depression, unease, disappointment, anger, and bitterness. 3. **Indices for Emotion Identification**: - The content and discursive procedures used by the young girl were considered the most important indices by many participants (15 out of 30). - Vocal parameters were also deemed important by some participants (10 out of 30), as well as mimogestuality (5 out of 30). - Notably, no participant considered film techniques as an emotion inducer for viewers. - Secondary indices included facial expressions (mentioned 10 times) and tears (7 times). Some students mentioned vocal behavior (12 times), but only one indicated film techniques. 4. **Findings**: The study highlights the multimodal nature of emotional communication in television. Emotions are constructed through the interplay of various phenomena, such as visuals, audio, and language. In this particular case, the empathic goal of the show is served by emphasizing despair, a highly empathetic emotion according to Cosnier's hierarchy. 5. **Unexpected Results**: Participants did not consider the film techniques and constructional work done by filming significant in determining their reception. This suggests that viewers might still perceive television as purely capturing reality rather than actively shaping it. Despite this, viewers acknowledge the authenticity of emotions fabricated through these means. In conclusion, the reception test confirms many aspects of the initial analysis but raises questions about how different components contributing to emotional expression are hierarchically perceived by viewers.
Original abstract	Résumé — Cet article traite des émotions dans une émission de télévision, et plus précisément de la mise en place d'un ≪ dispositif émotionnel ≫ au cours d'un talk-show politique. Ce travail est basé sur l'analyse multimodale de l'émission Demain les jeunes ,diffusée sur France 2 en mars 94. L'analyse de cette séquence émotionnelle (une jeune fille parlant de son désespoir) montre que sa force émotionnelle dépend d'un dispositif global et multimodal. L'expression des émotions est basée sur le verbal, le vocal et le mimogestuel, mais est aussi induite par un cadrage préalable (l'animateur parle de l'angoisse de la jeunesse) et renforcée par le filmage (gros plans). Cette analyse est complétée par un test de réception. Ce travail s'inscrit dans le champ de la pragmatique des interactions médiatiques.

Table 69: Example of LexRank summary of an original text with two ordered lists. The lists are only partially extracted: in the first case, the text of 3 out of 4 list items is (partially) extracted, but only the first item with its number. It the second case, only the first item number is extracted. List content is highlighted in blue in the summary, and the corresponding sentences are highlighted in blue in the original text.

Text ID	school-de-9 (in GitLab)
Summary	[...] Generell unterscheidet das Bundesministerium für Wirtschaft und Klimaschutz vier Dimensionen von Digitalisierung: 1. Digitale Produkte: Dienste oder Produkte sind nicht-physisch datenbasiert und können in der Regel ohne menschliche Einbringung erbracht werden. Digitale Prozesse: Prozesse können teilweise oder ganz datenbasiert gesteuert und organisiert werden. d. : oder teils) automatisiert durchgeführt werden“ (Bundesministerium für Wirtschaft und Klimaschutz). Digitale Vernetzung: Hierbei geht es um die Verknüpfung der digitalen Prozesse in ein digitales Gesamtsystem. [...] Die Kultusministerkonferenz (KMK) reagierte 2016 auf die Herausforderungen der digitalen Transformation für Schulen, Schülerinnen und Schüler sowie Lehrkräfte mit ihrer Strategie „Bildung in der digitalen Welt“, in der sie verschiedene verbindliche Kompetenzbereiche und Handlungsfelder festhielt – die Vermittlung der Kompetenzen in diesen Bereichen ist seither in Schulen und Ausbildung verpflichtend: 1. Die Vermittlung und der Erwerb von digitalen Kompetenzen sind daher unabdingbar für ein Leben und Arbeiten in einer digitalisierten Welt.
Original text	[...] Generell unterscheidet das Bundesministerium für Wirtschaft und Klimaschutz vier Dimensionen von Digitalisierung: 1. Digitale Produkte: Dienste oder Produkte sind nicht-physisch datenbasiert und können in der Regel ohne menschliche Einbringung erbracht werden. Beispiel: Künstliche Intelligenz (KI) in ChatBots für die Beantwortung von Kundenfragen einsetzen. 2. Digitale Prozesse: Prozesse können teilweise oder ganz datenbasiert gesteuert und organisiert werden. So können im Beruf zum Beispiel „Webshops betrieben, Kunden analysiert oder Beschaffungs-, Absatz- und Produktionsprozesse völlig (Anm. d. Verf.: oder teils) automatisiert durchgeführt werden“ (Bundesministerium für Wirtschaft und Klimaschutz). Beispiel: Papierbelege durch digital ablaufende kaufmännische Prozesse ersetzen und somit nachhaltiger arbeiten. 3. Digitale Vernetzung: Hierbei geht es um die Verknüpfung der digitalen Prozesse in ein digitales Gesamtsystem. Beispiel: Einer Online-Bewerbung für eine Ausbildung in einem Handwerksbetrieb ein digitales Schulzeugnis aus einer Cloud anfügen – hier wird also ein digitaler Bewerbungsprozess mit einer digitalen Zeugniserstellung wie -ablage verknüpft. 4. Digitale Geschäftsmodelle: Kundinnen und Kunden können digital(e) Produkte erwerben. Die Kundenprozesse sind hier stark digitalisiert und der Kunde wird über Prozessdaten einbezogen. Beispiel: Einen Online-Shop mit einer Webseite eröffnen, in dem eine Kundschaft einkaufen kann. Digitalisierung zeigt sich privat wie beruflich als ein sich stetig wandelnder Prozess in allen vier Dimensionen. [...] Die Kultusministerkonferenz (KMK) reagierte 2016 auf die Herausforderungen der digitalen Transformation für Schulen, Schülerinnen und Schüler sowie Lehrkräfte mit ihrer Strategie „Bildung in der digitalen Welt“, in der sie verschiedene verbindliche Kompetenzbereiche und Handlungsfelder festhielt – die Vermittlung der Kompetenzen in diesen Bereichen ist seither in Schulen und Ausbildung verpflichtend: 1. Suchen, Verarbeiten und Aufbewahren: Wie suchen, filtern, aus- und bewerten, speichern und abrufen? 2. Kommunizieren und Kooperieren: Wie interagieren, teilen, zusammenarbeiten, miteinander umgehen und an der Gesellschaft aktiv teilhaben? [...] Die Vermittlung und der Erwerb von digitalen Kompetenzen sind daher unabdingbar für ein Leben und Arbeiten in einer digitalisierten Welt.

Table 70: Example of NeMo summary with insufficient information coverage, which retrieves most information from the first half of the text. In the original text, sentences from which information was retrieved are highlighted in blue.

Text ID	school-de-9 (in GitLab)
Summary	Die Digitalisierung bezieht sich auf die Verwendung digitaler Technologien in Gesellschaft, Wirtschaft und Politik, um wirtschaftliche und soziale Prozesse zu beschleunigen und zu verbessern. Sie ermöglicht beispielsweise das Homeschooling oder Fernunterricht an Schulen sowie das Streaming von Filmen und Musik. Die Bundesregierung definiert vier Dimensionen der Digitalisierung: digitale Produkte, Prozesse, Vernetzung und Geschäftsmodelle. Die digitale Transformation bezieht sich auf die Integration digitaler Technologien in alle Bereiche von Wirtschaft und Alltag, um Effizienz, Nachhaltigkeit und Wettbewerbsfähigkeit zu steigern sowie Transparenz und Nutzerfreundlichkeit zu verbessern. Der digitale Wandel betrifft auch das Bildungssystem und erfordert die Vermittlung digitaler Kompetenzen in Schulen und Ausbildung.
Original text	Grundlegend ist darunter die digitale Abbildung eines analogen Gegenstands zu verstehen, zum Beispiel der Brief (analoges Objekt) und die E-Mail (digitaler Brief). Heutzutage benutzt man den Begriff „Digitalisierung“ aber vielmehr für die Verwendung digitaler Technologien in Gesellschaft, Wirtschaft und Politik, um wirtschaftliche und soziale Prozesse zu beschleunigen und zu verbessern. Alltäglich gesprochen ermöglicht sie zum Beispiel in der Schule das Arbeiten von zuhause im Homeschooling oder Fern- beziehungsweise Distanzunterricht. Filme, Serien und Musik können gestreamt werden und müssen nicht mehr als physische DVDs oder CDs vorliegen. Durch Digitalisierung soll – wie durch jede technische Innovation – der Alltag für Individuen, Gesellschaften und Unternehmen leichter und flexibler werden. Man kann Kosten und Zeit sparen und Arbeitsumgebungen produktiver und effizienter gestalten – auch in der Schule. Generell unterscheidet das Bundesministerium für Wirtschaft und Klimaschutz 1. Digitale Produkte: Dienste oder Produkte sind nicht-physisch datenbasiert und können in der Regel ohne menschliche Einbringung erbracht werden. Beispiel: Künstliche Intelligenz (KI) in ChatBots für die Beantwortung von Kundenfragen einsetzen. 2. Digitale Prozesse: Prozesse können teilweise oder ganz datenbasiert gesteuert und organisiert werden. So können im Beruf zum Beispiel „Webshops betrieben, Kunden analysiert oder Beschaffungs-, Absatz- und Produktionsprozesse völlig (Anm. d. Verf.: oder teils) automatisiert durchgeführt werden“ (Bundesministerium für Wirtschaft und Klimaschutz). Beispiel: Papierbelege durch digital ablaufende kaufmännische Prozesse ersetzen und somit nachhaltiger arbeiten. 3. Digitale Vernetzung: Hierbei geht es um die Verknüpfung der digitalen Prozesse in ein digitales Gesamtsystem. Beispiel: Einer Online-Bewerbung für eine Ausbildung in einem Handwerksbetrieb ein digitales Schulzeugnis aus einer Cloud anfügen – hier wird also ein digitaler Bewerbungsprozess mit einer digitalen Zeugniserstellung wie -ablage verknüpft. 4. Digitale Geschäftsmodelle: Kundinnen und Kunden können digital(e) Produkte erwerben. Die Kundenprozesse sind hier stark digitalisiert und der Kunde wird über Prozessdaten einbezogen. Beispiel: Einen Online-Shop mit einer Webseite eröffnen, in dem eine Kundschaft einkaufen kann. Digitalisierung zeigt sich privat wie beruflich als ein sich stetig wandelnder Prozess in allen vier Dimensionen. Digitale Transformation bedeutet digitaler Wandlungsprozess in Wirtschaft, Gesellschaft und Politik, denn hierbei geht es um das Integrieren digitaler Technologien und Lösungen wie Datenanalytik, Künstliche Intelligenz, Augmented und Virtual Reality oder Sensorik in allen Bereichen und Belangen von Wirtschaft und Alltag. Wirtschaftlich betrachtet ist so eine Steigerung der Effizienz, Nachhaltigkeit und Wettbewerbsfähigkeit sowie eine Verbesserung der Transparenz und Nutzerfreundlichkeit möglich. Alle Unternehmen sind von dieser Transformation betroffen, denn die Geschäftsmodelle und -prozesse werden sich in den nächsten Jahren aufgrund der zunehmenden Digitalisierung verändern: Kommunikation mit Kunden, Beschaffung, Vertrieb, etc. Wie und in welchem Ausmaß diese Wandlungen vonstattengehen, ist dabei individuell verschieden. Digitale Prozesse – zum Beispiel über das Internet – werden von Unternehmen sowie Kundinnen und Kunden genutzt. Weiterhin betrifft der Wandel auch Anwendungen wie Software und bildet sich in mehreren Dimensionen ab: in Betriebsmodell, Produkt, Service und Kundenbeziehung. Wichtig dafür sind die Enabler („Ermöglicher“), also die technischen Voraussetzungen, sowie die Akteure, die menschlichen Anwenderinnen und Anwender der digitalen Transformation. Benötigt sind demnach einerseits eine gute digitale Infrastruktur und digitale Anwendungen, die Angebote digital zur Verfügung stellen können. Diese müssen dann wiederum zur Wertschöpfung (wirtschaftliche Leistung) genutzt und das Verwertungspotenzial (neue Möglichkeiten für Unternehmen, die sich aus technischen Entwicklungen ergeben, zum Beispiel neue Geschäftsmodelle) ausgeschöpft werden, um den digitalen Wandel voranzubringen. Der digitale Wandel der Gesellschaft, Politik und Wirtschaft verändert auch die Lehr- und Lernprozesse im Bildungssystem – in der allgemeinbildenden wie beruflichen Bildung. Die Kultusministerkonferenz (KMK) reagierte 2016 auf die Herausforderungen der digitalen Transformation für Schulen, Schülerinnen und Schüler sowie Lehrkräfte mit ihrer Strategie „Bildung in der digitalen Welt“, in der sie verschiedene verbindliche Kompetenzbereiche und Handlungsfelder festhielt – die Vermittlung der Kompetenzen in diesen Bereichen ist seither in Schulen und Ausbildung verpflichtend: 1. Suchen, Verarbeiten und Aufbewahren: Wie suchen, filtern, aus- und bewerten, speichern und abrufen? 2. Kommunizieren und Kooperieren: Wie interagieren, teilen, zusammenarbeiten, miteinander umgehen und an der Gesellschaft aktiv teilhaben? 3. Produzieren und Präsentieren: Wie etwas entwickeln, produzieren, weiterverarbeiten, integrieren und dabei rechtliche Vorgaben beachten? 4. Schützen und sicher agieren: Wie sicher in digitalen Umgebungen agieren, persönliche Daten und Privatsphäre sowie Gesundheit und Umwelt schützen? 5. Problemlösen und Handeln: Wie technische Probleme und eigene Defizite lösen, Werkzeuge einsetzen, digitale Tools und Medien zum Lernen, Arbeiten und Problemlösen nutzen sowie Algorithmen erkennen und formulieren? 6. Analysieren und Reflektieren: Wie Medien analysieren, bewerten, verstehen und reflektieren? 2017 wurden derartige Ansätze auf europäischer Ebene durch den „Europäischen Referenzrahmen für digitale Kompetenzen von Lehrenden“ etabliert. Die KMK-Strategie von 2016 wurde 2021 noch erweitert: Wie können neue Technologien Entwicklungsprozesse in Schule und Unterricht ermöglichen? Wie können Lehrkräfte entsprechend didaktisch und technisch ausgebildet werden? Was müssen Schülerinnen und Schüler können, wissen und umsetzen, um sich auf das Heranwachsen und Leben in einer sich rasch wandelnden digitalen Welt vorzubereiten? Und was ist dafür an digitalen Tools, Medien, Knowhow, Prüfungsformaten, (berufsschulischen, betrieblichen und überbetrieblichen) Verknüpfungen sowie an überunterrichtlichen Organisations-, Personal- und Kooperationstechnologien wie -konzepten notwendig? In jedem dualen Ausbildungsberuf müssen digitale Kompetenzen im Betrieb und in der Berufsschule vermittelt werden. Die Vermittlung und der Erwerb von digitalen Kompetenzen sind daher unabdingbar für ein Leben und Arbeiten in einer digitalisierten Welt.

Table 71: Example of LexRank summary containing a misleading header and sentences written by different people. Sentences used in the summary are highlighted in bold in the original email thread. For space reasons, only a part of the original email thread is reported here, and some line breaks have been removed. The whole thread has already been reported in Table 51.

Text ID	emails-unpr-4 (in GitLab)
Summary	——Original Message—— From: Rangel, Ina Sent: Thursday, March 07, 2002 12:56 PM To: Badeer, Robert Subject: FW: Badge Access Ina, We can most likely have this by Monday morning and he can pick this up at the 5th floor reception. Will he be able to have his badge by Monday morning and where will he go to pick it up.
Original thread	[...] ——**Original Message—— From: Rangel, Ina Sent: Thursday, March 07, 2002 12:56 PM To: Badeer, Robert Subject: FW: Badge Access** When you get here on Monday morning, come to the 5th floor reception of the new building. [...] ——Original Message—— From: Curless, Amanda Sent: Thursday, March 07, 2002 2:50 PM To: Rangel, Ina Subject: RE: Badge Access **Ina, We can most likely have this by Monday morning and he can pick this up at the 5th floor reception.** If he has any problems he can call me. Thanks! Mandy ——Original Message—— From: Rangel, Ina Sent: Thursday, March 07, 2002 2:39 PM To: Curless, Amanda Subject: RE: Badge Access << File: Badge Access Form.doc >> I filled out all of the information that I had on him. **Will he be able to have his badge by Monday morning and where will he go to pick it up.** Ina [...]