Radboud University Nijmegen

Faculty of Social Sciences

On the factual correctness and robustness of deep abstractive text summarization

MSc Artificial Intelligence

Author:
Klaus-Michael Lux
Student number:
s1012898

Academic supervisor:
Martha Larson

External supervisor:
Maya Sappelli

Second reader:
Iris Hendrickx

August 20, 2020

Contents

1 Introduction

FD Mediagroep is an Amsterdam-based publisher whose main products are Het Financieele Dagblad (FD), a daily business newspaper, and BNR Nieuwsradio. Supported by a grant by the Google digital news initiative, the company introduced the Smart Journalism project [1]. The aim of the project is to offer a personalized landing page for users that serves automatically generated summaries of FD articles tailored to the specific interests of the user. The summarization task envisioned can be outlined as follows: Generate a single-document, multi-sentence abstractive summary that is served to the reader as a bullet-point list, where every bullet-point is a full sentence. In line with general requirements on summarization, the generated summaries should be grammatical and concise, i.e. only contain the most relevant information from the article. The most important aspect however is factual correctness: No incorrect information should be contained in the summary, as this would severely undermine FD's credibility and reputation as a provider of serious, accurate financial news.

While the past few years have seen an explosion of research interest in neural abstractive text summarization, a recent critique by [2] highlights a number of pressing issues in the field that have so far been insufficiently addressed, among them the issue of factual correctness. The authors find that even though recently developed abstractive methods perform well according to widely used automatic metrics which rely mostly on word overlap with reference summaries, they still produce a high number of factually incorrect summaries. There currently is no comprehensive typology of the factual errors produced by text summarization systems. Individual authors generally tend to provide a few examples or verbal descriptions of frequent errors, such as [3], who state that *"[c]ommon mistakes are using wrong subjects or objects in a proposition [...], confusing numbers, reporting hypothetical facts as factual [...] or attributing quotes to the wrong person."*. More generally, we lack an understanding of how the systems use language, what kind of errors they make and how those affect the reader.

This situation means it is currently difficult to gauge for FD Mediagroep whether an abstractive summarization system should be used at all and if so, which one would be most suitable. This research will tackle this knowledge gap by investigating the performance of four recently introduced summarization systems and analyzing system errors in a systematic fashion. Creating a typology of errors and comparing systems is highly relevant to guide decisions on which system to use, especially if the severity of different types of errors varies. Additionally, this typology can guide future research, enabling the design of new methods and strategies to tackle sources of specific types of error.

Previous research into the prevalence of errors has narrowly focused on systems that were trained from scratch on the summarization task. However, transfer learning, i.e. the use of knowledge gained in one task in another related task, has been on the rise in the language domain. Its emergence began after [4] demonstrated that a pre-trained language model can be used effectively for a multitude of downstream applications, such as text classification and question answering. In contrast to earlier approaches such as Word2Vec or GloVe, the pre-trained model is not merely used as a feature extractor for a different model. Rather, the whole language model is retained and then depending on the nature of the downstream task, additional layers are added at the top. Domain data for the desired downstream task can then be used to fine-tune the model to perform this task. This approach is attractive for FD's use case: As there is less training data available due to the smaller audience of a Dutch-language media outlet compared to what is available for big English-language media companies, more needs to be done

with less. Pre-training can boost sample efficiency, i.e. fewer samples are required to obtain good performance and it can thus be applied successfully to data-sparse domains such as FD's summarization task. However, there is no prior research into the factual correctness of summaries generated using this approach and thus it is currently unclear whether it is suitable.

The aim of this research is to provide a linguistic analysis of the errors neural abstractive summarization systems produce and how they affect the factual correctness of summaries. Setting out to ensure diversity, we select a total of four different abstractive summarization systems by different authors, two of which leverage transfer learning. The summarization task was the same for all systems and the same dataset was used, allowing us to inspect their output on the test set to establish a typology of summary errors. Its validity is evaluated using the agreement between multiple annotators and then used to annotate a larger number of summaries to get an understanding of the prevalence of errors among different systems. Finally, we look into the performance of the systems for totally new data from the same source as used originally – how robust are they to changes in the topics covered in the articles?

2 Background and related work

2.1 Automatic text summarization

Automatic text summarization is the task of automatically generating a summary of textual information. We can distinguish between *single-document* summarization and *multi-document* summarization. While the former concerns generating a summary for a single given document, the latter requires integrating information from multiple documents. For the use case of FD, single document summarization needs to be performed, as it is planned to generate a single summary for every article published.

Existing approaches for automatic summarization can be grouped into *extractive* and *abstractive* methods. In extractive summarization, the summary is composed solely of sentences present in the source document(s). For example, a very widely cited method of this type by [5] represents sentences as nodes in a graph, connected by their cosine similarity. The most central nodes are then picked and concatenated into a summary. Alternatively, extractive summarization can also be construed as a binary sentence classification task, where for each sentence we predict whether it should be included in the summary. A fairly recent approach in this vein by [6] uses a recurrent neural network to encode sentences and documents into vector representations. The document representation and sentence representations are then fed into a sigmoid readout layer that predicts the inclusion for each sentence.

In contrast, abstractive approaches do not yield output composed solely from source sentences, rather, they rephrase the source into possibly novel sentences. Historically, systems of this type have been rarer, as the desired behaviour was difficult to achieve before recent advances in natural language processing. Due to the more open formulation of the task, abstractive approaches offer the promise of higher-quality summaries that could potentially be similar to human-written content, flowing naturally and rephrasing information concisely. Rather than just a combination of document sentences not written to be part of a summary and requiring post-processing steps, these methods could directly deliver readable and informative summaries. Due to this high appeal, FD decided to put the focus on abstractive summarization. However, there is a plethora of approaches that differ in a number of ways. Before further considering different systems,

it is necessary to first delineate criteria for selecting the summarization approaches to include in the evaluation. The next section describes criteria that have previously been used for evaluating the quality of generated summaries.

2.2 Evaluating summary quality

The decision for one summarization system over the other will largely be based on summary quality. However, this is multi-faceted, involving aspects such as concision, readability and factual correctness. The following sections discuss various aspects in summary quality and how they have been previously been evaluated to facilitate comparisons between summarization systems.

2.2.1 Evaluation aspects and practices

Once we have a summary for a given document, how do we decide how good it is? What aspects are important to consider? Over the years, there has been a large number of works relying on different paradigms, e.g. various forms of human evaluation and automatic measures. The field of evaluation was deeply affected by summarization challenges offered at two conferences, namely first the stand-alone Document Understanding Conference (**DUC**, which ran till 2008) and then later the Text Analysis Conference (**TAC**) of which DUC became a subtrack. While exact criteria at these challenges have varied, three elements have remained constant over the years, namely *readability*, *informativeness* and *non-redundancy* [7]. Readability describes the linguistic quality of the summary, i.e. how easy it is to read and understand. Informativeness describes how useful a summary is to a reader in terms of the information from the article contained within it. Finally, non-redundancy deals with the conciseness of the summary, punishing summaries that are repetitive.

While readability and non-redundancy are usually evaluated using manual methods (e.g. by asking assessors to rate summaries), there is more variability in practices for rating informativeness. According to [8], three types of metrics for informativeness can be distinguished, namely questionnaire-based metrics, overlap-based metrics and other metrics. Metrics from the first category require some form of human input, e.g. by requiring humans to answer questions based only on the summary text and then having raters judge the quality of the answers or by asking participants to judge summary quality on a Likert scale.

In contrast, overlap-based metrics are fully automatic. By comparing a candidate summary to a reference summary and investigating the overlap of content units at different levels, these metrics allow a fast and easy evaluation. A number of overlap-based metrics have been proposed over the years. Though its low correlation with human ratings has been pointed out by [9], the most widely reported metric in recent papers is still ROUGE. Originally presented in 2004 [10] and used at DUC, the metric comes in a number of different shapes, but the general principle always involves computing the overlap between elements of the candidate summary and the reference summary, with a higher overlap being viewed as preferable. ROUGE-N computes the overlap between n-grams. All n-grams in the reference summary are obtained and one then computes what proportion is also found in the candidate summary. Widely used instantiations of this metric are referred to as ROUGE-1 (for $n = 1$, unigrams), ROUGE-2 (= 2, bigrams). ROUGE-L first finds the longest common subsequence in the two summaries and then divides it by the length of the candidate summary or the length of the reference summary to obtain the precision or the recall, respectively. The harmonic mean of the two quantities then yields the final score (also known as ROUGE-L-F1).

2.2.2 Factual correctness

How does the aspect of factual correctness interact with different aspects in evaluation? There appears to be no prior research on this issue and both recent surveys considered [7], [8] do not mention this issue at all. This can presumably be explained by the history of the field of text summarization: Before the advent of deep learning, there were no black-box systems with unclear properties trained in an end-to-end fashion. Instead, most systems were extractive, incurring only a small risk of introducing factual errors via problems with rewriting. Alternatively, when systems were abstractive, they had to rely on methods that were built manually leveraging ideas from information extraction or natural language processing (c.f. [11], who build a pipeline for mapping articles into semantic graphs, reduce them to achieve compression and then use text generation to produce summaries). Approaches in this vein were well controlled and unlikely to produce factual errors on a significant scale. As a consequence of this lack of knowledge, we cannot estimate whether sufficient performance as evidenced by various metrics can guarantee that summaries are indeed factually correct. Especially for overlap-based metrics like ROUGE, this seems unlikely: As they only measure surface-level overlap and do not capture the retention of semantic aspects, a factually incorrect summary could still receive a high score if has sufficient word-level overlap with a reference document.

Indeed, research has demonstrated that even though recent abstractive systems score high on automatic metrics for summary evaluation such as ROUGE, they have a propensity to generate factually incorrect summaries. An analysis by [12] of a recent neural system finds that up to 30 % of generated summaries contain "fabricated facts". Similarly, the authors of [3] evaluate three different state-of-the-art systems (all trained from scratch) and find that between 8 and 26 % of the generated summaries contain at least one factual error, even though ROUGE scores indicated good performance.

As factual correctness is of high importance to FD Mediagroep, the application of these approaches to the task at hand could be problematic. [3] and [12] propose post-processing steps to reduce the prevalence rate of incorrect summaries, but these require additional resources (such as entailment models and entity extraction frameworks) and are therefore not directly applicable to the task. A different avenue of research involves relying on pre-training: Rather than training a model to generate summaries from scratch, a pre-trained model is utilized. This model is trained on a language modelling task using a large corpus and then fine-tuned on the summarization task, as exemplified in [13], [14] and [15]. The technique is promising with respect to factual correctness: The authors obtain good results on automatic metrics, additionally [15] also conduct an evaluation based on a QA paradigm that asks humans to answer factual questions based on generated summaries, reporting an improvement in how well they were able to do this when the pre-trained model was used. If a summary enables humans to give correct answers to reference questions, it can generally be assumed to be factually correct, as long as the set of questions covers a sufficiently diverse set of statements in the text. Together with the intuition that pre-training should allow the model to learn useful facts about language in general that could help it better perform the summarization task, it seems plausible that pre-training could be an effective method to avoid factual errors in summaries.

2.3 Abstractive neural summarization systems

While automatic text summarization has been a research subject for decades, interest in the field has increased strongly in recent years, mostly due to advances in natural language processing (NLP) and the successful application of deep learning allowing for

end-to-end learning of abstractive summarization. There is now a plethora of approaches using various neural architectures and training schemes for this purpose (for two recent surveys, see [16] and [17]).

We first describe the basic paradigm underlying most current work in neural text summarization in Section 2.3.1. After that, Sections 2.3.2 - 2.3.5 describe the details of the four approaches, sorted by date of original publication. Section 2.4 contains a comparison between them and discusses some of the expected differences in terms of factual errors. Section 2.5 discusses the aspect of robustness to changes in the articles. Finally, we describe the dataset that was used for training and evaluation of all approaches in Section 2.6.

2.3.1 Sequence to sequence approaches

While exact architectures and training objectives vary, almost all approaches for neural abstractive summarization share the same basic paradigm: Summarization is treated as a sequence to sequence task. Originally developed for the use in machine translation, this paradigm views both the source document x and the summary y sequences of tokens, written as $(x_1, x_2, ..., x_d)$ and $(y_1, y_2, ..., y_s)$, respectively. Most approaches task an encoder with translating the source x to a sequence of hidden states h^e and a decoder with generating the summary y based on these hidden states as its input. Both the encoder and the decoder can be instantiated by different network architectures, including convolutional and recurrent neural networks. The first paper to transfer the Seq2Seq paradigm to the summarization domain was [18] and since then, numerous authors have proposed variations on the underlying theme.

2.3.2 Pointer-generator model with coverage

The earliest and most widely cited system in consideration is the pointer-generator model presented by [19]. The authors criticize Seq2Seq models presented at the time for a tendency to include repetitions and for being "liable to reproduce factual details inaccurately". They claim that their approach overcomes both these problems, but this finding is based only on casual observations and not backed up by any systematic evaluation of factual correctness. The novelty of the approach is derived from two ideas. The first is the inclusion of a pointer mechanism. In contrast to the basic Seq2Seq paradigm, which generates one word from the vocabulary at every decoder step using the probability distribution P_{vocab}, the model can decide to instead copy a word from the source document by means of the pointer. The decision between these two actions is made based on the generation probability p_{gen}, which is treated as a soft switch. For a given word w at time t, the probability of generation is defined as:

$$P(w) = p_{gen} * P_{vocab}(w) + (1 - p_{gen}) * \sum_{i:w_i=w} a_i^t$$

The first term in the sum uses the output distribution of the Seq2Seq decoder weighted by the probability of generating rather than copying. The second term is the copying term. Here, the probability of producing the word via copying is obtained by the complement of p_{gen} (i.e. the probability of copying) multiplied by the activation of the word in the attention distribution a^t of the model. Hence, the model is a hybrid of a Seq2Seq model and a copying mechanism, which allows it to also produce words that are not part of its vocabulary.

p_{gen} is not a hyperparameter, but rather computed for every timestep based on the context vector h_t^*, the decoder state s_t and the decoder input x_t, as the sigmoid of the

sum of these vectors weighted by learnable weights and a learnable bias. These learnable components enable the model to obtain the ability to dynamically adjust p_{gen} based on information represented in the context and the decoder.

The second novelty in the approach is its adaptation of a coverage mechanism. By keeping a sum of attention distributions over time, we get a measure of how much of the input in the sentence was already "covered". This coverage vector is used as an additional input to the attention mechanism, weighted by learnable weights. The intuition behind this is that it might help the model spread its coverage of the sentence better, thus avoiding repetitions. In practice, the authors find that additionally, a separate coverage loss term is necessary to induce sufficient coverage.

The authors conduct an entirely automatic evaluation of their proposed changes, comparing them to the results observed by [18] on ROUGE with reference summaries. In an ablation analysis, they find a slight improvement compared to baseline by adding the pointer mechanism and a larger improvement by also incorporating coverage. However, even the combined system fails to outperform a simple LEAD baseline, in which the first three article sentences are treated as the summary. Looking at abstractiveness in terms of novel n-grams, they find that the generated summaries are far less abstractive than reference summaries, e.g. less than 6 % of 3-grams in generated summaries are novel, compared to close to 70 % for reference summaries. Factual correctness is only alluded to, but not evaluated in any way.

2.3.3 Reinforce-selected sentence re-writing

[20] introduce a number of innovations to the basic pointer-generator paradigm. Most saliently, they split summarization into an extraction and an abstraction step, claiming this mimics the way humans summarize documents. The extraction step selects a subset of sentences, each of which is then re-written in the abstractive step. The authors claim this change reduces the risk of redundancy and improves the speed of summarization, as it no longer necessary to maintain an attention distribution over the whole document when abstracting individual sentences. The extractive step is handled by the extractor agent, a encoder-decoder with a pointer network which computes the extraction probability for each step. For the abstractive step, the abstractor agent is used. This is another encoder-decoder network that includes a copying mechanism. However, as the original dataset provides only articles and abstractive summaries, proxy training data for the two components has to be generated. For the extractor, the authors select the most similar document sentence (according to ROUGE-L recall) for any given reference sentence and assign label it as extracted, while all other sentences are assigned 0. In the same way, for the abstractor, the authors create pairs of each reference summary sentence and the closest document sentences and treat one as the abstracted version of the other for the purposes of training. Both components are first trained using maximum-likelihood training.

After that, reinforcement learning is applied, training the extractor further in an end-to-end fashion. This is done by introducing the notion of timesteps. At every timestep, the extractor selects a sentence that is then rewritten by the abstractor, yielding a summary sentence whose similarity to a document sentence (according to ROUGE-L-F1) is then used as the reward for training. The authors claim that this practice causes the extractor to extract more relevant sentences, as it would receive a high reward only for well-selected sentences that can be re-written to be similar to a reference summary sentence. As there is no natural point when to stop extracting sentences, the authors add a

"stop" action to the action space of the extractor agent. When this is selected, no more sentences will be extracted. They set the reward such that the model is encouraged to select the action when there are no remaining ground-truth sentences, assuming that this will teach the model to adaptively select the right number of sentences for a given article, eliminating the need to manually tune a cutoff parameter. The final innovation is the use of a re-ranking strategy: For every sentence, k candidates are generated, where k is the beam size of the beam search used for decoding. These candidates are retained and when all n sentences have been generated, all k^n combinations of beam search candidates are obtained as candidate summaries. These candidates are then reranked based on how many repeated n-grams they contain, with a lower number being preferred. The best candidate is selected as the final summary. This strategy is intended to remove redundancy similar to the coverage mechanism described above.

The authors rely on automatic evaluation, looking at ROUGE scores on the test set. They compare their system (and various ablations thereof) to [19] and a number of simpler baselines. Out of all abstractive methods, the full combination of maximum likelihood training, reinforcement learning and re-ranking performs best, outperforming both the best result by [19] and the LEAD-3 baseline, though the latter only barely. Additionally, they conduct a more detailed head-to-head comparison with [19], performing among others human evaluation by crowd-workers and a comparison of abstractiveness according to novel n-grams. These results show that their method is slightly preferred when summaries are judged on relevance. There is also a very small difference when summaries are judged on readability. The statistical significance of the differences is not reported. In contrast, the differences in abstractiveness are much more pronounced, with more than 22 % of 3-grams being novel compared to 6 % for [19]. While being only modestly better according automatic evaluation and human judgments, this difference makes the method stand out, offering much more abstractive summaries of a similar general quality. However, factual correctness is not a topic of interest to the original authors, with no experiments looking into this aspect.

2.3.4 Methods leveraging pre-trained language models

Since the authors of [4] demonstrated the potential of using pre-trained neural language models for various language processing applications, there has also been strong research interest to apply them for abstractive summarization. [21] showed that a pre-trained language model can even generate sentences resembling summaries when directly applied to the summarization task with no further fine-tuning (*zero-shot learning*), though they do point out automatic metrics are fairly low and summaries "often focus on recent content from the article or confuse specific details such as how many cars were involved in a crash or whether a logo was on a hat or shirt." The language model they used is known as GPT-2. It is based on the Transformer architecture [22] and trained to predict the next word on a large corpus of text. A language model such as GPT-2 can be directly applied to summarization if one takes a different view of the task, construing it as a conditioned generation task rather than one of translation. One feeds the (truncated) document into the model and then uses a special token to induce generation based on the document seen so far. The generated text is then treated as the summary. This approach is known as *decoder-only*. Alternatively, one can retain the traditional encoder-decoder framework and just instantiate either (or both) components by the (pre-trained) language model. [13] compare the merits of these different approaches. Looking at the effects of pre-training, they find that it generally improves performance on ROUGE regardless of the approach chosen. Comparing the encoder-decoder framework to the decoder-only approach, they find the latter to be marginally better. Both

are competitive with existing approaches in the field. However, by means of training models on various subsets of the training data, they demonstrate the decoder-only approach to be much more sample-efficient, achieving substantially higher ROUGE scores when trained on very small sets. Analyzing the differences between the two approaches under these conditions, they claim that the decoder-only approach can better leverage information from the source and is less likely to hallucinate information. They do not analyze whether this is also true for situations where more training data is made available.

A very similar paper by [14] only looks at the effects of pre-training on the decoder-only architecture. They find pre-training to generally improve performance and decoder-only models to be on par with existing methods, while being much simpler, obsoleting the need for many techniques such as "sequence-to-sequence modeling, coverage mechanisms [or] direct ROUGE optimization via reinforcement learning [...]".

For this comparison, we were interested in including a decoder-only model leveraging a pre-trained language model, as both these properties could have an affect on factual correctness of generated summaries. We approached the main author of [13], asking for trained models. Unfortunately, those were not longer available, but she graciously pointed us to the "sister paper" of her publication [23]. The authors leveraged a pre-trained GPT model for a decoder-only summarization system and report ROUGE scores that are only slightly worse than those reported by [13]. No evaluation of abstractiveness is performed. The trained model was made available to us after e-mail conversation with the main author of the paper. None of the papers implementing this approach have conducted any investigation into the factual correctness of generated summaries.

2.3.5 Text summarization with pre-trained BERT

There is not just one type of neural language model. Rather, multiple researchers have proposed different paradigms that vary in their exact architecture and training objective. Another widely used model by [24] is known as BERT. It also uses Transformer layers, but enables bidirectional self-attention, allowing both left and right context to be taken into account at every step. The training objective is also different. Rather than predicting the next word based on its left context, words are randomly masked and context from both sides is used for the prediction. Beyond these aspects, there are also differences in how exactly fine-tuning is conducted. Similar to GPT-2, pre-trained BERT has been demonstrated to be highly useful for various applications. [15] use the model for abstractive summarization and report high ROUGE scores that are the state-of-the-art as of the end of 2019. The authors describe two different set-ups for using the model, namely an extractive and an abstractive setup. For the extractive setup, a modified BERT model called BERTSum is combined with a head consisting of transformer sentence layers and sigmoid classifier. This is used to predict whether a sentence should be extracted. Again, the authors are faced with the problem that no ground-truth labels are available for this task, so they pick the most similar document sentence for each reference sentence. The abstractive setup relies on the standard Seq2Seq paradigm and uses BERTSum as the encoder an a randomly initialized Transformer as the decoder. A specific training schedule is applied for this component to ensure that there is no mismatch due to the different amount of pre-training. Both setups can be combined, yielding a condition the authors call BERTSumExtAbs. One first fine-tunes BERTSum on the extractive task and then some further on the abstractive task. This performs better than the abstractive setup, though being outperformed in terms of ROUGE scores by only using the extractive setup. This system is not very abstractive, with around 15

% of 3-grams being novel, compared to close to 70 % in reference summaries.

The authors additionally also conduct a human evaluation study using a QA paradigm that involves asking human subjects to answer a set of reference questions only relying on the generated summary. The method performs significantly better on this benchmark than a number of models trained from scratch and the LEAD baseline for three different datasets.

2.4 Comparison of summarization systems

We set out to get a representative estimate of the error prevalence of a number of recently proposed summarization systems in order to judge whether the current state-of-the-art is suitable for the needs of FD. We selected a total of four different neural abstractive systems, visualized in Figure 1. Among them are the least and most error-prone system according to a study into factual errors [3], specifically See [19] and Chen [20], respectively. Additionally, two recent approaches relying on a pre-trained language model are included in the analysis, namely LM [23] which relies on a GPT transformer and treats summarization as a language modelling task and PreSumm [15], which uses BERT as a pre-trained language model and adopts a more traditional view of summarization as a translation task featuring an encoder and a decoder.

The pointer-generator architecture (henceforth referred to as **See**), the RL-inspired rewriting paradigm (**Chen**), the language-modelling approach leveraging GPT (**LM**) and the approach leveraging pre-trained BERT encoders (**PreSumm**) were all trained on the same split of the non-anonymized version of the CNN/Daily Mail dataset (see Section 2.6). For See and Chen, there has been some prior inquiry into their propensity of generating factually incorrect summaries, namely [3], for the other two systems, no data of this sort is available. Table 1 contains a detailed comparison of the systems in consideration. The table shows minor variation in ROUGE scores and stronger variation in terms of abstractiveness. We can see that See is most in line with a traditional encoder-decoder architecture, while Chen and PreSumm are variations of the theme. The former contains two separate encoder-decoder systems, one for an extractive and one for an abstractive task. PreSumm has only one encoder that is shared for an extractive and abstractive task. LM in contrast does away with the whole paradigm, replacing it with an approach where summarization is treated as a language modelling problem.

The aims of our analysis are two-fold: For the purposes of FD, we are interested in the differences between systems, asking which is most suitable for current needs. Having picked a diverse array of candidates, we expect to find relevant differences. The other aim is to get a deeper understanding of how different design decisions affect performance of summarization systems. We thus identify the two most salient high-level differences between the four systems. These are whether transfer learning is used and how (if at all) the system involves an extractive step. See and LM directly train on the abstractive task and do not involve extraction. In contrast, PreSumm performs initial fine-tuning an extractive task and Chen even involves an extractive sub-step directly in the pipeline. This allows us to investigate the effects of these two design decisions, though our insights will be of a more preliminary character, as these aspects are not the only relevant differences between systems. We formulate some initial hypotheses, subject to some revision after the error typology is established and validated.

Regarding **pre-training**, we predict that in general, the incidence rate of factual errors will be reduced. Some prior research even allows us to even speculate about the

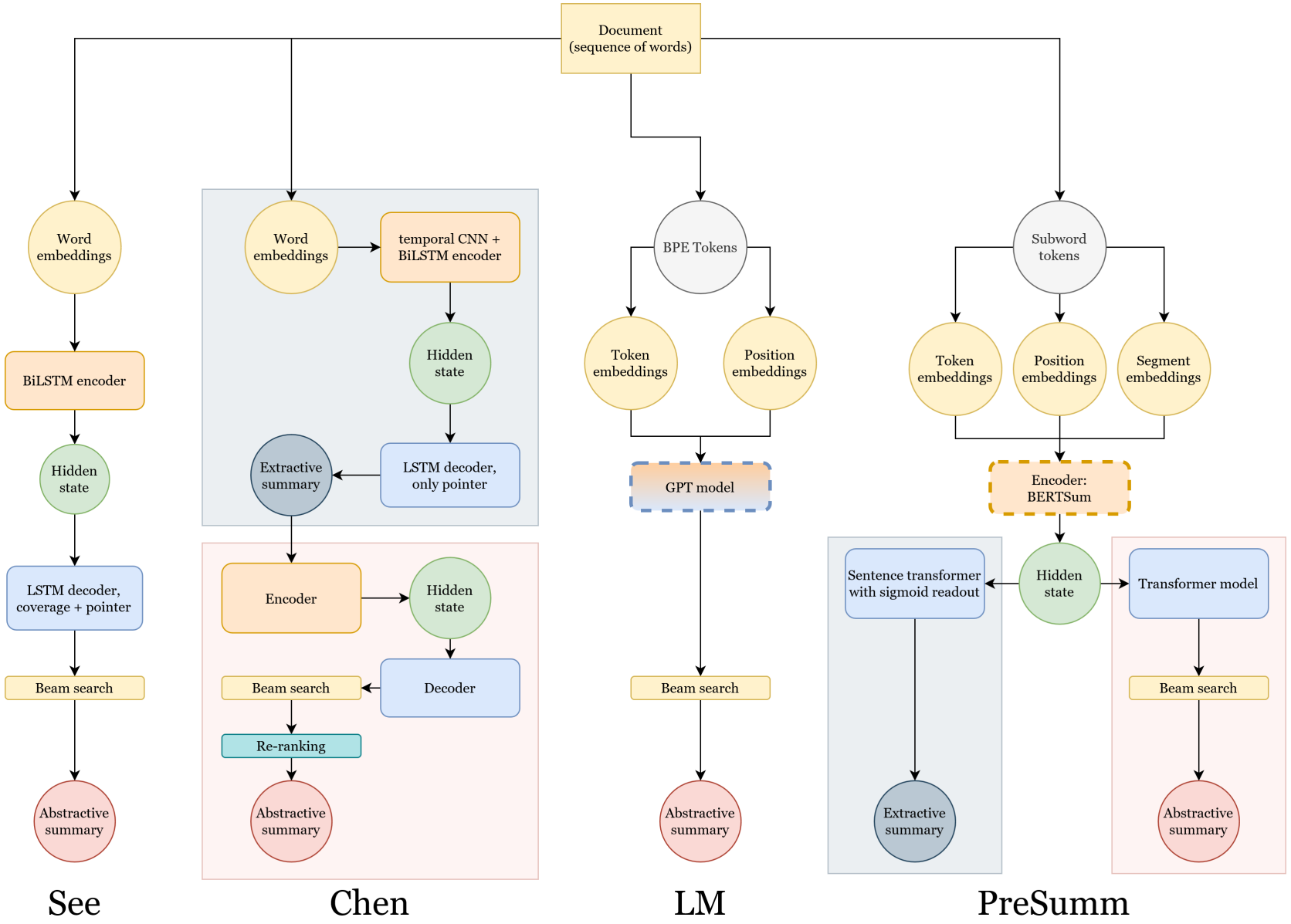


Figure 1: A schematic representation of the four summarization systems. Encoders in orange, decoders in blue. Extractive components in dark blue, abstractive components in red. Pre-trained components indicated by dashed lines.

effects on different types of factual errors. Some types may be reduced, while others might become more prevalent. Specifically, [3] and [12] do not provide a quantitative analysis into different types of errors, but from the examples listed in these papers and their descriptions it is evident there are two types of error whose prevalence can be expected to be reduced by the prior linguistic knowledge that is encoded in a model due the language model pre-training:

1. Errors that reflect an insufficient grasp of the dependency structure of the target input, such as generated summaries that contain subjects and predicates not from the same sentence in the target. Research into the attention heads within pre-trained Transformer models [25] shows that some of them capture dependency relations such as nominal subject (*nsubj*) and this implicit knowledge could help a pre-trained system to avoid errors of this kind.

2. Errors that yield a sentence that is semantically implausible. As the language model is pre-trained on a large array of text, it can be expected to capture co-occurrence patterns that can help it not to generate semantically implausible sentences such as *"bosnian moslems postponed after unhcr pulled out of bosnia"* as reported by [12].

However, there are also reasons to believe other errors might be more prevalent: [13] point to instances where the model hallucinates information not present in the source document. This can be a cause for factual errors in at least two ways: The model might create some untrue, topically related fact that is nowhere to be found in the document. Alternatively, it might include some background information on entities found in the text, some of which might be no longer accurate, such as referring to Barack Obama as the current US president. The original authors do not conduct an evaluation into the different types of hallucination or their relative prevalence, but if sufficiently prevalent, they would also be highly problematic for the proposed task.

Regarding the **inclusion of an extractive step**, the effects are harder to predict. The reasoning by [20] that their two-step approach nature is a more natural, human-like way to model the task is compelling, however, there is already some evidence that their system performs poorly in terms of factual correctness, though we cannot directly tie this aspect to the separation between extractor and abstractor. However, there are reasons to think this could be the case: More complexity is introduced (e.g. heuristics need to be chosen for picking training pairs of re-written sentences) and the abstractor does not have access to the full underlying document, making it hard for it to appropriately model effects introduced by context. Using the extractive task only for fine-tuning in line with [15] does not suffer from the issues, so it could be predicted that this might be a preferable method to involve extraction.

2.5 Robustness to article changes over time

Concept drift refers to a problem in supervised learning where the relation between input data X and the prediction target y over time changes. This can have a detrimental effect on model performance. One distinguishes between **real concept drift** where it is truly the relation between input and output $P(y|X)$ that changes, while the input distribution $P(X)$ itself stays the same and **virtual drift**, where the relation stays the same, while the input distribution changes [26]. In the context of summarization, the former would mean that what constitutes a good summary y for two similar articles x_1 and x_2 from different time points is not the same. The latter would refer to the more general fact that articles might change over time, which might then also have a downstream performance impact if models are not stable to this change. There has been

	See	Chen	LM	PreSumm
Transfer	No	No	Yes	Yes
Ext. and abs.	Not separate	Separate components of the pipeline	Not separate	Separate tasks during training
Encoder	Single layer bidirectional LSTM	**Extractor**: temporal convolutional model + single layer bidirectional LSTM, **Abstractor**: standard encoder	Summarization as a language modelling task, only one GPT model	BERTSum, a modified version of a pre-trained BERT encoder
Decoder	Single layer unidirectional LSTM with attention and coverage mechanism	**Extractor**: LSTM with pointer mechanism, **Abstractor**: Standard decoder with copy mechanism	-	**Extractive task**: Sentence transformer with sigmoid readout, **Abstractive task**: Randomly initialized 6-layer transformer
Decoding method	Beam search, size: 4	Beam search, size: 5 trigram blocking, diverse beam search	Beam search, size: 3, trigram blocking	Beam search, size: 5, trigram blocking, length normalization
Input truncation	400 tokens per document, start higher and then decrease	100 tokens per sentence	512 tokens per document, 110 tokens per summary	512 tokens per document
Output truncation	120 tokens (almost never reached because of self-stopping)	Not truncated	Not truncated	Not truncated
Input Embeddings	Word embeddings, trained from scratch	Word embeddings initialized to W2V trained on the same set, trainable	Token and position embeddings, both available pre-trained	Token, position and segment embeddings; the former pre-trained, latter alternate between sentences
Vocabulary	50K, words	30 K, words	GPT's BPE (40k merges) vocab	BERT's 30k subword vocab
ROUGE-1	39.53	40.88	38.67	42.13
ROUGE-2	17.28	17.80	17.47	19.60
ROUGE-L	36.38	38.54	35.79	39.18
Novel 1,2,3-grams	0.1, 2.2, 6.0	0.3, 10.0, 21.7	Not available	0.4, 10.0, 15.0

Table 1: Comparison of systems in consideration

no investigation into if and how this phenomenon is present for the domain of summarization. Generally, models are evaluated only on held-out from one point in time, correspondingly, we lack an understanding into what happens if a model is deployed and then left as is. As training summarization systems is quite expensive in terms of necessary computation and requires expert supervision, it seems conceivable that the following scenario might happen: An organisation might train and deploy a model once, but then fail to monitor or adapt the system over time. It is thus important to gauge what the possible consequences of this might be.

For this research, we will focus on virtual drift, as this is more straightforward to investigate, given that the mapping $P(y|X)$ (i.e. what constitutes a good summary) is difficult to define. It is hard to estimate what changes in the input distribution might occur, this will likely also depend on the nature of the news outlet. An outlet might start producing articles with a different target audience and differences in writing style, e.g. when starting a series of background articles on certain topics or introducing shorter, bullet-point style coverage of certain events. Any of these changes might cause problems in a summarization component if it has not sufficiently generalized to the task and relies on the presence of latent stylistic properties which are difficult to model. Correspondingly, whenever changes of this sort are made, system output should be closely monitored. One more predictable and frequent change is the emergence of new topics and concepts that were not present in the training data. Governments change, companies boom and bust and there is a constant stream of novel trends and fads. Systems might be brittle in the face of this, failing to adequately summarize articles that contain novel topics. We are not aware of any prior study that has investigated how robust abstractive summarization systems are to changes of their input. For this reason, we will conduct a small pilot study, obtaining recent articles from the original sources and evaluating the performance of the four systems.

2.6 CNN/Daily Mail dataset

The **CNN/Daily Mail dataset** (CNN/DM for short) contains more than 310,000 article-summary pairs that were crawled between 2007 and 2015 from the websites of the American broadcaster CNN and the British tabloid newspaper Daily Mail. Both these websites feature abstractive bullet-point summaries written by editors which are presented to the reader at the top of article pages. The dataset was generated by Google researchers for a study on question answering [27], later adapted for the summarization task by [18] and has since then been used to train numerous summarization systems presented in the literature. It comes with a pre-defined split: The validation set contains all articles extracted from the sites in March 2015 (13,368 articles), the test set the articles published in April 2015 (11,490) and the training set consists of articles published prior to these dates (287,226).

Both media outlets are still active and continue to operate the websites that were scraped by the original researchers. Consequently, more recent articles can be obtained and allow for an investigation into the robustness of different systems.

2.6.1 Issues and criticism

Even though widely used in a number of publications and a de facto standard dataset due to its sheer size, the CNN/DM dataset has a number of flaws and issues that are potentially problematic for training an end-to-end system. [2] criticize

- the **under-constrained nature** of the task. They claim that presenting the

model with a single summary for each article makes the task too ambiguous, as there might be multiple good summaries for a given article and as prior knowledge and the expectations of different readers are not modeled.

- the **layout bias** present in the data. Due to the way news articles are written, important sentences tend to be clustered towards the beginning of the article. The authors demonstrate that this is also the case for the CNN/DM dataset and lament the fact that rather then being viewed as a possible impediment towards generalization, the bias has even been leveraged in the heuristics of current summarization models.

- the high amount of **noise** in scraped datasets in general. As content is extracted automatically and manual review of the large amount of content is infeasible, the quality of the data depends largely on heuristics and post-processing steps taken by the original extractor of the dataset. For CNN/DM dataset, the authors report that 0.47 % of training, 5.92 % of validation articles and 4.19 % of test articles were found to contain noise such as "links to other articles and news sources, placeholder texts, unparsed HTML code, and non-informative passages in the reference summaries"

Over the course of this study, we also inspected a number of articles and could validate that noise and artifacts are indeed present. Beyond this, we also found that occasionally, articles were duplicated, as the editors of the Daily Mail page had apparently republished them on a different date with a slightly different headline. Additionally, we found that image captions were included along the rest of the text, even though images are not included in the input and this often results in repetitions when image captions also appear as document sentences. Clearly, for a system to work on news articles from any source, such peculiarities in the data should be avoided as much as possible. Systems might learn to bank on the fact that certain important sentences are repeated, thus incorrectly generalizing and possibly failing if new data does not conform to this structure.

A more theoretical aspect not considered in prior literature is the assumption that what is served to the reader on the CNN and Daily Mail sites is even intended as a self-contained summary. This does not seem to ever have been questioned. The collection of the data was conducted by independent researchers that have no apparent connection to either publishing outlet and who did not intend to use the data for summarization, thus not reflecting on this aspect. However, there are grounds for the assumption that this does not really capture the purpose of the text. On both sites, the short bullet-points are presented not on an overview page, but only after the article has been clicked (c.f. Figure 2 for an illustration. On the Daily Mail page, the bullet-points directly follow the headline, on the CNN page, they are included in a separate box titled "Story highlights" that is placed to the left of the first paragraph of the article body, also below the headline and the caption of the first image if any is present. It thus seems overwhelmingly likely for the reader to have read the headline (and possibly the image caption) before referring back to the article summary. Editors writing the reference summaries can be expected to rely on this unless explicitly instructed to write a fully self-contained summary. Inspecting some articles, we found a number of reference summaries that explicitly assumed the presence of the headline and were difficult to understand in isolation. For example, the reference summary of a Daily Mail article from 2011[1] (part of the training set) reads:

[1] `http://web.archive.org/web/20110726014413id_/http://www.dailymail.co.uk/news/article-2018608/Whataburger-Carol-Karl-Hoepfner-visit-720-favourite-fast-food-chain-stores.html`

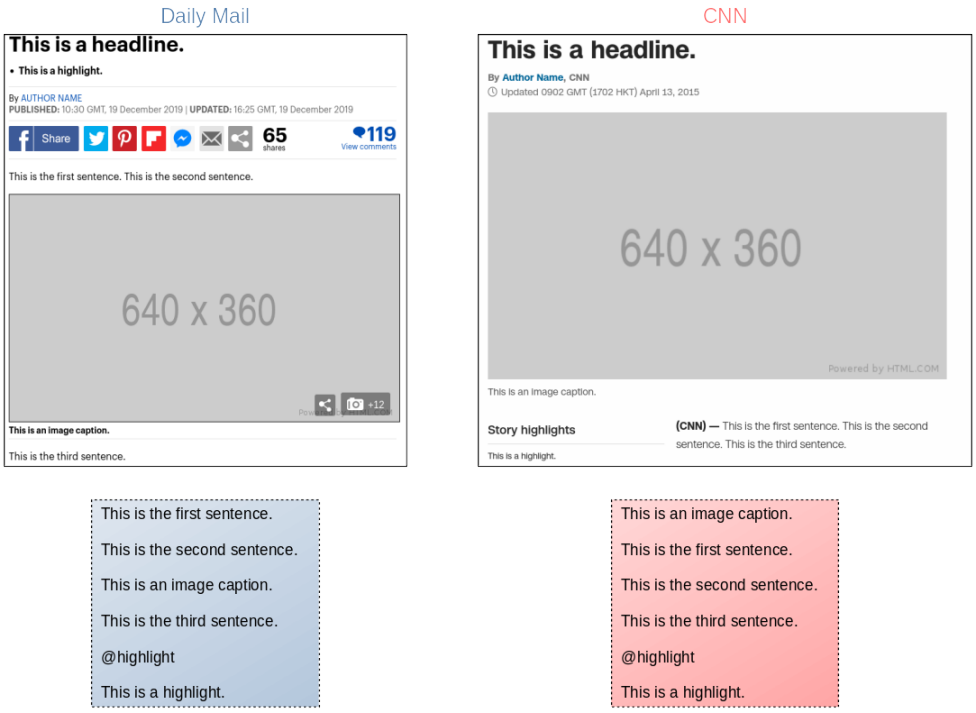


Figure 2: Visualization of the effect of the pre-processing on the article representation for DM and CNN. Both headlines are omitted and the image caption is integrated into the text body.

- Carol and Karl Hoepfner have already visited 225 of 722
- Couple had first Whataburger meal almost 50 years ago
- Awarded $10,000 prize of "Whataburger's Biggest Fans'
- Eaten more than 7,000 meals at their local in Texas

This is hardly a self-contained summary. The first bulletpoint is confusing as there appears to be a missing reference. Only by reading the following sentences carefully and then making a number of assumptions can the possible meaning be inferred. Now consider reading the bullet-points after having read the headline "Whataburger! Retired couple to visit 722 restaurants of favourite fast food chain across 10 states". With the additional context, the summary becomes easy to understand, as the crucial pieces of information "722 restaurants" and "favorite fast food chain" are now present.

This effect is also present in the CNN articles. The reference summary of a CNN article from 2015[2] (part of the test set) reads:

- Scientists in southern Italy have known about him since 1993
- Researchers worried that rescuing the bones would shatter them

[2] `http://web.archive.org/web/20150615083935id_/http://edition.cnn.com/2015/04/13/europe/italy-oldest-neanderthal-dna-sample-altamura/index.html`

Without the original headline "Neanderthal who fell into a well gives scientists oldest DNA sample" there is no way to understand the meaning from the summary sentences alone.

These findings mean that unless combined with the headlines, the summaries on the pages are not self-contained and often hard to parse. However, the headline is simply discarded by the authors who originally extracted the dataset and thus also not used at all by any approach that uses the CNN/DM dataset for training. In essence then, all systems in consideration are trained on reference summaries that are possibly incomplete and difficult to parse. It seems conceivable that this has effects on the downstream performance, making summaries less coherent and less likely to make sense in isolation.

3 Research questions

We selected four abstractive neural summarization systems whose differences allow us to investigate the effects of different strategies for involving extraction and of pre-training. All systems have been trained on the same section of the CNN/DM dataset. Consequently, their outputs can be directly compared on the held-out test portion of the dataset. We will create a typology of errors on a subset of these articles and describe how they relate to factual correctness. The typology will be validated by means of measuring the agreement between multiple annotators. This yields two research questions:

RQ1: What is the nature of errors produced by abstractive summarization systems? What errors can be distinguished and how can they be categorized? How do they affect factual correctness?

RQ2: Can we achieve human agreement on what constitutes an error in the setting of abstractive summarization?

After generation and validation, we will compare the prevalence of error types between different systems. A larger number of summaries will be annotated and the error prevalence will be compared. Focusing on the most salient differences between the systems in question, we pose the following research questions:

RQ3: How do different methods of involving an extractive step (not all all, only during training or as a separate component of the model) affect the prevalence of different types of errors?

RQ4: Do summarization systems that leverage pre-trained language models differ in systematic ways in the prevalence of different types of errors when compared to models that are trained from scratch?

It is one thing to produce correct summaries on articles that are fundamentally similar to the training data in formatting, style and the distribution of topics covered, as can be expected for the held-out set of the CNN/DM dataset. Since the training set contains data from the months directly before the cut-off point, it can be expected that there are no strong differences in these aspects. As time progresses, the set of topics frequently featured in the news will likely change, and this might have a negative impact on how well the summarization systems work. In this thesis, we will inspect summaries generated for more recent articles from the original sources to evaluate whether current

automatic summarization systems are robust to changes in their input. This yields the following two research questions:

RQ5: Does the prevalence of error types change when summaries are generated for recent articles from the original sources?

RQ6: Do the models differ in how robust they are to this change?

4 Generating an error typology

In this section, we set out to answer **RQ1** by systematically describing and analyzing the errors made by current summarization systems. We first describe how we used a card-sorting approach to group erroneous sentences and then describe the results and multiple revisions to increase mutual exclusivity and linguistic grounding of the error groups. Finally, we present the typology, illustrating different error types by means of example sentences.

4.1 Grouping errors by card sorting

We first collected the output of the summarization systems on the test set of the CNN/DM dataset which was provided by the original authors. Each summary was then matched to the underlying article and ingested into a database. Even though previous approaches had only performed summary-level annotation, we decided to conduct the annotation on a sentence level. This was done as to improve flexibility, allowing us to look at more fine-grained differences or to optionally aggregate sentence-level errors to the level of the summary.

To identify different types of errors in summary sentences, we employed a card-sorting approach [28]. This is a method frequently used in various domains to group a number of objects into meaningful categories. Each object is printed on a card and placed on a surface accessible to a number of participants, who are then tasked with grouping cards into categories. As we had no strong prior intuitions, we used an open card-sorting, i.e. we did not define any error categories beforehand, but instead allowed participants to freely create, merge and remove categories during the sorting process.

For each of the four summarization systems, we randomly sampled 30 of its summaries, ensuring that all were for different articles. A filter was employed, omitting summaries that contain only sentences directly copied from the document, i.e. purely extractive summaries. Articles and summaries were printed to a DIN A5 template that showed a portion of the article at the top and summary sentences at the bottom. The overlap between summary sentences and article sentences was computed and indicated visually to the sorters: Each summary sentence was assigned a different color. Words copied from the article into the sentence were color-coded both in the article and in the summary. We determined the sentence furthest into the document that contained words copied into the summary. Articles were cut off two sentences after that sentence. If an article still overran the available space, it was cut off at that point. When a summary contained multiple sentences that were not extractive, one copy of the card was printed for each of those sentences. Each of these sentences was additionally marked with a star.

The card-sorting was conducted at the offices of FD, with a total of six people in attendance. Three of the participants were Data Scientists at FD, one was the product owner of the Smart Journalism project and two (including the author) were master's

Index	**Category name**	**Number assigned**
1	Ungrammatical	20
2	Nuance missing	10
3	Context missing	18
4	Hallucination	1
5	Wrong word re-writing	3
6	Wrong subject	20
7	Word(s) missing	16
8	Wrong combination of sentence parts	5
9	No error	251

Table 2: Initial error typology after first card-sorting pass

students working as interns in the Data Science team at the company. All participants are proficient users of the English language, as this is the language used for everyday communication at the office, though none are native speakers (four participants have Dutch as their first language, one Vietnamese, one German). They were instructed to carefully read the article and each of the sentences. Whenever a sentence struck them as wrong or inappropriate in some way, they were asked to try to identify the underlying error. If they felt that there was no suitable error category present yet, they could create a new category by writing its name on a sticky note. They then placed the card for the erroneous sentence next to the category they chose. After all summaries had been presented, we reviewed the categories that had been created.

After the initial pass of the card sorting, we found a small number of instances where two document sentences had been merged, this was later to be determined to be found to be to a bug in the printing process. We thus proceeded to analyze the merged sentences in isolation and they were added to another appropriate category. A small number of sentences could not be judged from the portion of the article included in the print-out, we retrieved the full article text for them and sorted them into other categories as appropriate. Another small set of sentences was labeled under 'Meaning in article unclear', these were discussed in the plenum and then also sorted into other categories. The resulting error categories are presented in Table 2. It can be seen that a relatively large number of error types was identified, differing in relative frequency.

4.2 Initial revision: Ensuring exclusivity

After some closer inspection, it became evident the initial categories lacked mutual exclusivity, as some were hard to delineate from one another. For example, some of the cases labeled as *word(s) missing* were also ungrammatical. Similarly, the distinction between missing nuance and context was not clear. While some of the categories seemed to focus more on the surface nature of the error (*wrong word re-writing, word(s) missing*), some others dealt more with the consequences (*context missing, wrong subject*). It is possible that this situation was caused by participants attending differently to these aspects. As this situation would make it hard to achieve a clear categorization of new examples by annotators, the categories underwent a substantial revision conducted by the author.

The new error typology distinguishes two dimensions of summary error: The *mapping dimension* describes the surface level, looking at how the summary system used words and phrases from article sentences to create the erroneous summary sentence. Four

different cases can be distinguished:

- **Omission**: The system copies words from summary sentence, but omits certain words or phrases. The omission causes an error. This category is based on the *word(s) missing* category.
- **Wrong combination**: The system copies words or phrases from multiple article sentences and combines them into an erroneous sentence. This category is based on the *wrong combination of sentence parts* category.
- **Fabrication**: The system introduces one or multiple new words or phrases that cause an error. This category is based on the *wrong word re-writing* category.
- **Lack of re-writing**: The system fails to adequately re-write sentences, e.g. by not replacing referential expressions with their original antecedents in the text. When the antecedents are also not present in the preceding summary context, this causes an error. This category is based on the *context missing* category.

In contrast, the *meaning dimension* describes the effect of the error on how (and if) the reader understands the sentence. There are four different high-level descriptions of the effects an error can have:

- **Unnatural language use**: A sentence that is either syntactically or semantically unnatural and would not be uttered by a competent speaker. It might be malformed, i.e. it does not comply with the rules of syntax. Alternatively, it might be obviously nonsensical due to semantic errors. Sometimes, the error causes the sentence not to have any clear meaning, i.e. a reader will not be able to understand it. This category is based on the *ungrammatical* category.
- **Meaning changed**: A sentence that claims something that is in no way entailed by the article. This category is based on the *wrong subject* category.
- **Implication changed**: A sentence whose implication structure is altered when compared to corresponding sentence in the article. The reader will still be able to correctly infer the meaning, but might be misled to assume implications that were not present in the original article. This category is based on the *Nuance missing* category.
- **Dangling anaphora**: Expressions such as *the group* or *california firm* are present in the sentence, but the entity they refer to (their antecedent) is not present in the surrounding summary context. The effect on the reader is trouble understanding the meaning of the sentence.

Table 3 shows the breakdown of errors in the initial set according to these dimensions. All summary sentences were revisited and sorted into the category space spanned by the two dimensions. In the process, it was discovered there was an additional mapping not accounted for, it is included separately: **Error in the article** refers to cases for which the summary error was already present in the article.

4.3 Second revision: Linguistic grounding

The revised typology was used to annotate a small set of unseen summaries, relying on two annotators, including the author. While the mapping dimension appeared generally clear, we discovered that further changes to the meaning dimension were necessary to

Mapping dimension	Meaning dimension	Examples from	Number assigned
Omission	Unnatural language use	1,7	20
”	Meaning changed	6,7	12
”	Implication changed	2,3,7	12
”	Dangling anaphora	-	-
Wrong combination	Unnatural language use	1,8	9
”	Meaning changed	6,8	18
”	Implication changed	-	-
”	Dangling anaphora	-	-
Fabrication	Unnatural language use	-	-
”	Meaning changed	4,5	4
”	Implication changed	2	1
”	Dangling anaphora	-	-
Lack of re-writing	Unnatural language use	-	-
”	Meaning changed	3	1
”	Implication changed	-	-
”	Dangling anaphora	3,6,7	26
Error in article	Various	1,7,8	3
None		9	238

Table 3: Distribution of summaries among error types after revision of initial card-sort results.

achieve a less ambiguous annotation. Specifically, we had not paid sufficient attention to a number of linguistic processes that could have an effect on how the reader understands a summary. For instance, a sentence previously labeled as **dangling anaphora** could be perceived as fine, if one assumes the existence of linguistic accommodation, i.e. the reader trying to use contextual information and world knowledge to infer the reference. Similarly, it was pointed out that the category **Unnatural language use** was too broad, encompassing both syntactical and semantic errors that seemed to be qualitatively different. We also found issues with the **Meaning changed** category: Occasionally, summary sentences would make claims that were not contradicted by the article, but also could not be assumed to be true. Situations like this might occur when a summarization system uses knowledge gained from training set articles as the basis for test set summaries, inserting information it has previously seen. Arriving at these findings, we realized that the existing typology was insufficiently grounded in linguistic theory. We thus took a step back, and by involving an existing account of sentence processing ([29]), arrived at a flowchart representation of the process (Figure 3). In addition to providing more grounding, this typology also offers an intuitive conception of *error severity*, distinguishing between **malformed** and **misleading** sentences.

According to this view, summary sentences are first parsed into a structural representation of their syntax (we are agnostic about the exact details of this representation). When this process fails, a sentence is judged to be ungrammatical. In the next step, the sentence meaning is inferred, using accommodation and repair strategies in the process. Should these fail, no meaning can be inferred. Generally, inference can fail because the sentence has no truth conditions given its context. In a more specific case, the sentence has no possible truth conditions under world knowledge, it is then judged to be semantically implausible. All errors up to this step (color-coded in yellow) can be considered to be **malformed**, causing the average reader to stumble and question the quality of

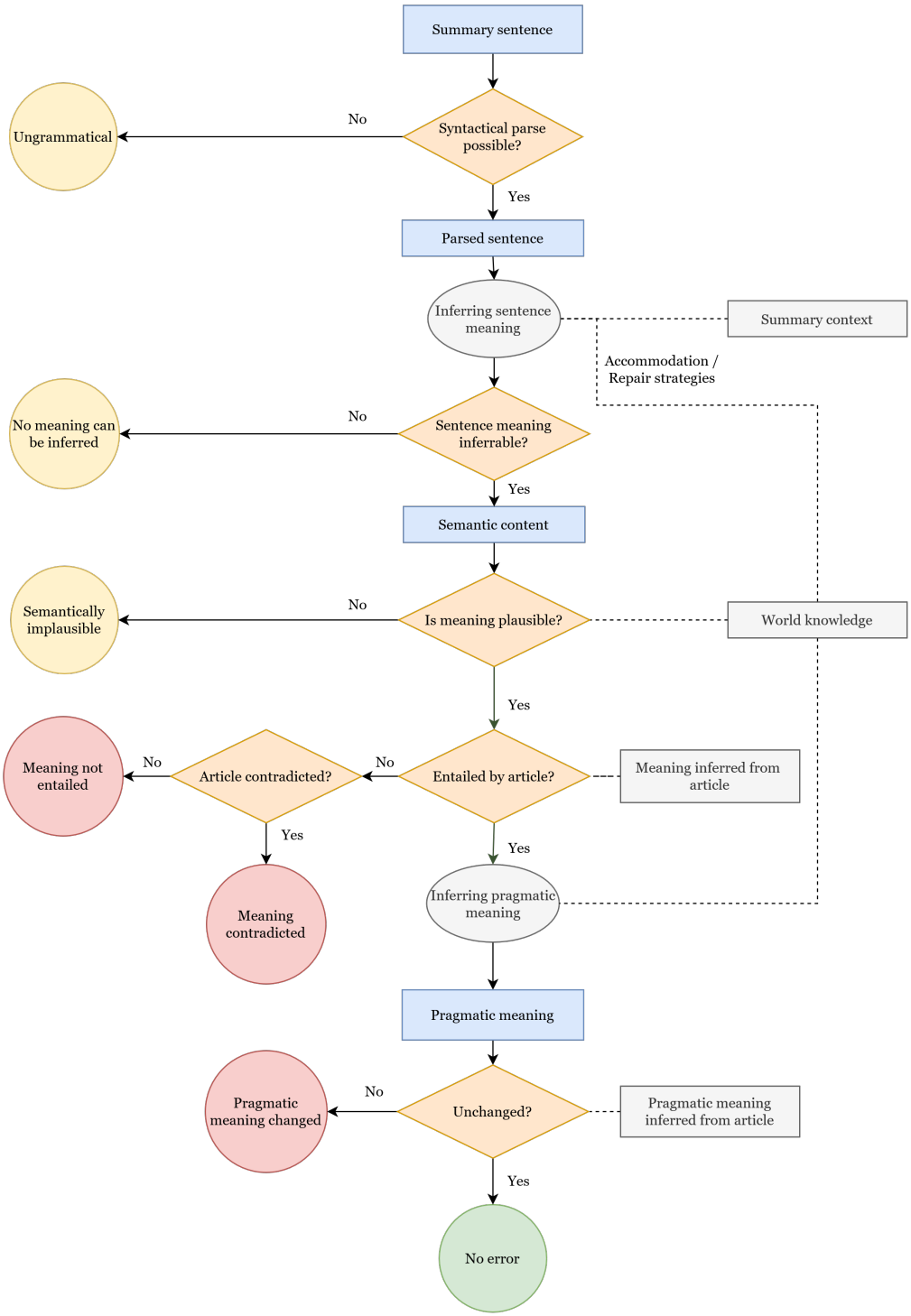


Figure 3: View of linguistic processing taken for revision of the typology. Meaning dimension categories in circles, color-coded by severity (malformed = yellow, misleading = red).

the summary, but they will not mislead the reader in any way. After this point, without recourse to the article, the reader has no means of spotting additional errors. If inference succeeds, the reader arrives at the semantic content of the sentence. Pragmatic inference processes kick into play, generating additional aspects of meaning beyond the semantic content. By comparing the semantic content and pragmatic meaning licensed by the summary sentence to what would be inferred *if the reader had full access to the original article*, we can detect cases of divergence. Specifically, if the semantic content inferred is not entailed or contradicted by the article or if the pragmatic meaning inferred differs, we arrive at additional failure cases for the summary sentence. Errors after successful inference (color-coded in red) can be considered to be **misleading**: in contrast to malformed sentences, they cannot directly be noticed by the reader as they do not contain any obvious cues.

Based on these insights, we completely revised the meaning dimension. We split **Unnatural language use** into **Ungrammatical** and **Semantically implausible**, allowing the scheme to reflect different levels of linguistic capacity that a system might be lacking. While the former demonstrates an incomplete grasp of syntactical rules, the latter reflects negatively on the world knowledge of the system. The category **Dangling anaphora** was completely removed, replaced by a more general **No meaning can be inferred** that explicitly accounts for the accommodation process and only encompasses sentences for which it fails entirely. **Meaning changed** was split, distinguishing between cases for which the new meaning was clearly in contradiction to claims made in the article and cases for which the new meaning was merely not entailed by the article. We also hypothesize that these errors could differ in ease of detection by a human editor checking automatically generated summaries before publishing. Sentences in clear contradiction to the article should be easier to spot than those are merely not entailed. **Implication changed** was renamed to **Pragmatic meaning changed** to allow for it to cover a wider range of pragmatic phenomena. In sum, the resulting typology now encompasses a total of six categories, namely:

- **Ungrammatical**: A sentence that is syntactically unnatural and would not be uttered by a competent speaker. It is syntactically malformed, i.e. it does not comply with the rules of syntax.
- **Semantically implausible**: A sentence that is semantically unnatural and would not be uttered by a competent speaker. It is obviously nonsensical due to semantic errors.
- **No meaning can be inferred**: A sentence that is grammatically correct, but to which no meaning can be assigned, even after accommodating.
- **Meaning changed, not entailed**: When read in the context of the surrounding summary, the semantic content assigned to a sentence is not entailed by the original article.
- **Meaning changed, contradiction**: When read in the context of the surrounding summary, the semantic content assigned to a sentence is in contradiction to what is said in the article.
- **Pragmatic meaning changed**: When read in the context of the surrounding summary, the sentence gains a pragmatic meaning that was not present in the original article. Alternatively, a pragmatic meaning present in the original article is not faithfully retained in the summary.

4.4 Final error typology

To further elucidate the final typology and to give the reader a qualitative understanding of what each of the error type amounts to, we present a number of examples for each of the categories. These are sorted by the meaning dimension, but also annotated according to the mapping dimension. This was done as the meaning dimension was judged to be more relevant for practical issues, such as deciding which system to use.

4.4.1 Ungrammatical

Sentences in this category are entirely ungrammatical, as they lack words or phrases that are syntactically required. They are easily detected by human readers even without reference to the original article, but their presence is indicative of a lack of syntactic abilities in a summarization system that produces them.

Article context: [...] she suffered from the rare disease progeria which ages the body at eight times the normal rate. [...]
Summary sentence: she suffered from rare disease progeria which ages the body at eight times.

Example 1: Ungrammatical. Omission. System: Chen. *By deleting only parts of the prepositional phrase "at eight times the normal rate", an ungrammatical sentence is created.*

4.4.2 Semantically implausible

This category refers to summary sentences that are unnatural in their composition. They would not be produced by a competent user of the English language. Sentences in this category are grammatical on the surface, but obviously nonsensical.

Article context: [...] and among the most curious viewers of a royal night out, released next month to coincide with the anniversary of ve day on may 8 , 1945 , will be a woman who knows better than anyone what really happened on that extraordinary night. [...]
Summary sentence: the anniversary of ve day on may 8 , 1945 , will be a woman.

Example 2: Semantically implausible. Omission. System: Chen. *Due to large-scale deletions, parts of a relative clause providing a closer description of the film are merged with parts of the verbal phrase of the main clause, creating a totally new and obviously nonsensical sentence.*

4.4.3 No meaning can be inferred

The reader is not able to infer the meaning of sentences in this category, even after accommodating. Some sentences in this category contain a referential expression that the reader cannot resolve, as its antecedent is not present in the surrounding summary text (c.f. Example 4). Contextual information is not sufficient to make an educated guess about what the expression could refer to.

Article context: [...] female crews representing oxford and cambridge universities rowed the same stretch of the river thames in london as the men for the first time in the 87 years they have competed.. [...] oxford dominated cambridge in the men 's boat race today, claiming a third consecutive victory in a row [...]
Summary sentence: the men 's boat race took place on the same stretch of river thames as the men for the first time in 87 years.

Example 3: Semantically implausible. Wrong combination. System: LM. *The system merges information from two document sentences. These are only superficially related (both mention of the names of the universities competing in different tournaments), but the system merges the subject from the second sentence into the first and adds a new verbal phrase to combine them, yielding a nonsensical sentence.*

Article context: [...] it turns out a corporation can indeed be prosecuted like a person.. it 's a practice the supreme court has approved of for over a century. [...]
Summary sentence: it 's a practice the supreme court has approved of for over a century.

Example 4: No meaning can be inferred. Lack of re-writing. System: See. *By not including the antecedent of the referential expression "it" either by copying the sentence that contains it or by re-writing, the system creates a sentence whose meaning cannot be inferred.*

4.4.4 Meaning changed, not entailed

Sentences in this category make a new claim that is found nowhere in the article, but not directly in contradiction to it. This could be explained by summarization systems "learning" facts about the world that they then reproduce in other summaries. In Example 5, another prior article present in the training set might have contained the claim that the woman mentioned in the article is expecting her first child. This seems less likely in the case of Example 6: There would have to be another thematically very similar article by Sally Kohn that contains a similar claim and features an attribution to her in the summary text. In actuality, a reverse search revealed the article in question was not authored by Sally Kohn, but by another author. Thus, the behaviour of the model does not reflect true knowledge about the world, but rather a superficial tendency to insert author names into summaries of articles that might be thematically similar to articles they in fact have authored.

One can take two views on whether examples in this category should be considered errors. Under a "closed-world" interpretation, a summarization system is tasked only with summarizing what is in the article and should not inject any additional information. According to this view, if not entailed, summary sentences are erroneous and should not be produced at all. Under an "open-world" stance, in contrast, a summarization system would be akin to a human editor, being able to creatively rewrite the article content to generate a pleasing summary, with the caveat that all information contained in it should still be true based on some external source of knowledge. However, while a human editor would have the capacity to point to said source of knowledge should statements in his summary not be covered in the article, black-box neural abstractive systems currently possess no such capacity. When they engage in behaviour of this sort, the outside world has no guarantee it reflects their world knowledge of any sort. For this reason, while

instances in this category might be qualitatively different from those that involve a direct contradiction, we will take the closed-world interpretation and consider them to be errors as well.

Article context: [...] for in the wake of amelie mauresmo's announcement that she is pregnant, the world no 3's trial with his prospective assistant coach jonas bjorkman has assumed a greater importance.. mauresmo, who is to give birth some time in august, will be around eight months' pregnant during wimbledon this summer, which is not finishing until july 12 due to the new three week gap between roland garros and the championships.. [...]
Summary sentence: amelie mauresmo is eight months' pregnant with her first child.

Example 5: Meaning changed, not entailed. Fabrication. System: LM. *The article contains no information about how many children Mauresmo has.*

Article context: [...] yet for decades , many have stood by the belief that such programs, known as syringe exchange or syringe services programs, promote and encourage drug use. [...]
Summary sentence: sally kohn: for decades , many have stood by the belief that such programs promote drug use

Example 6: Meaning changed, not entailed. Fabrication. System: PreSumm. *The article contains no attribution to an author, but the summary attributes its claims to Sally Kohn, a political commentator for CNN.*

4.4.5 Meaning changed, contradiction

Beyond just being not entailed by the article, sentences in this category make a new claim that is directly contradicted by the article. For instance, information from different sentences is merged incorrectly (c.f. Example 7) or actions are attributed to the wrong subject (c.f. Example 8). While some sentences in this category might be easy to detect as false even without access to the article, others are plausible and thus prone to go unchecked by a reader. For this reason, their presence is highly worrying for the use case of FD. Readers could be seriously misinformed when taking summary sentences at face value, especially when reading under the assumption that summaries are written by human editors.

Article context: [...] if the player misses the girl, she starts to lose weight until she eventually dies.. [...] gamers have to throw food at the girl who appears in one of nine holes before she disappears again. [...]
Summary sentence: gamers have to throw food at the girl who appears in one of nine holes before she dies

Example 7: Meaning changed, contradiction. Wrong combination. System: PreSumm. *Information from two sentences is merged incorrectly, giving rise to a contradiction with original article meaning.*

Article context: [...] for in the wake of amelie mauresmo's announcement that she is pregnant, the world no 3's trial with his prospective assistant coach jonas bjorkman has assumed a greater importance. [...] andy murray celebrates taking a set against novak djokovic in the final of the miami open.. the british no 1 is set to marry his long-term girlfriend kim sears in dunblane on saturday. [...]
Summary: jonas bjorkman has assumed a greater importance.
the british no 1 is set to marry his long-term girlfriend kim sears in dunblane.

Example 8: Meaning changed, contradiction. Lack of re-writing. System: See. *The referential expression in the second sentence is not rewritten, misleading the reader to believe it refers to Jonas Bjorkman, rather than Andy Murray.*

4.4.6 Pragmatic meaning changed

In contrast to sentences that contain an outright change in semantic content, sentences in this category contain subtler changes. They are still entailed by the article, but changes in them lead to the addition or removal of pragmatic meaning. In Example 9, two statements entailed by the article are juxtaposed in a way to generate an additional aspect of pragmatic meaning not present in the original article. This is an example whose consequences are rather mild, as the reader is not seriously misled. In contrast, in Example 10, as aspect of pragmatic meaning is not faithfully retained due to a deletion. In this example, the effects are more worrying for the use case of FD: Reading only the summary and missing the pragmatic meaning that the publication does not endorse the claim as true, readers could be more likely to believe it. In case it is eventually disproved, they might trace their belief back to the publication, and thus it might lose credibility.

Article context: [...] the eco-friendly collection is made entirely of 100 per cent recycled material.. [...] the collection is inspired by the ocean, with coral reef scenes featuring throughout. [...]
Summary sentence: inspired by the ocean, the collection is made entirely of 100 per cent recycled material

Example 9: Pragmatic meaning changed. Wrong combination. System: PreSumm. *By juxtaposing the participial phrase with the main clause from another sentence, new pragmatic meaning is generated, namely that the material-sourcing was inspired by the ocean.*

4.5 Conclusion

For **RQ1**, we were interested in a linguistic analysis of errors of automatic summarization systems. By systematically analyzing the output of abstractive summarization systems, we arrived at a typology that describes their nature. The mapping dimension looks into how the summary sentence relates to sentences from the article. The meaning dimension describes how the erroneous sentence interferes with the processing of the sentence by the reader and one the consequences this can have on its meaning.

Additionally, the question asked how errors and factual correctness interact. The

Article context: [...] male colleagues at the bank were promoted ahead of her and she was 'mocked' and subjected to ' gratuitous derogatory' comments about her childcare arrangements, she alleges. [...]
Summary sentence: male colleagues at the bank were promoted ahead of her and she was ' mocked'

Example 10: Pragmatic meaning changed. Omission. System: Chen. *By deleting the phrase "she alleges" that the journalist used to distance himself, the summary changes the pragmatic meaning of the utterance. The additional meaning about the author not endorsing the main claim of the sentence is lost.*

meaning dimension provides insights into that: On the one hand side, a variety of errors involve malformedness and cannot mislead the reader in any way. On the other, there are a number of errors that have the potential to mislead the reader, giving rise to incorrect beliefs that would not have been produced by the article alone. These errors can be equated with factual errors in traditional parlance. They are highly worrying for FD and any system that consistently exhibits them would not be suitable for its needs.

5 Computing inter-annotator agreement

After having arrived at a typology, we now set out to answer **RQ2**, investigating whether it was possible to reach meaningful inter-annotator agreement (IAA) between annotators on the specified error categories.

5.1 Methods

To allow the annotation of large numbers of summaries by multiple annotators, articles and summaries were ingested into an online database (based on MongoDB) and then made accessible via an SSH connection. We developed a graphical user interface for summary annotation, known as *SummaryInspector* (see Figure 4). This tool presents the article in the upper half of the screen. The lower half is split into columns, each containing the summary of the article by one of the four systems in consideration. The order of the placement of the systems is shuffled for every sample and system names are hidden by default, such that annotators cannot be biased by hidden assumptions they might have about different systems. Summaries are split into sentences and each sentence can be annotated separately by means of check buttons. Convenience features for the annotation include color highlighting of summaries (similar in style to the examples presented above, allowing the annotator to quickly see where words from the summary originate in the article) as well as a functionality to quickly retrieve the original article on the web via a Google search of a couple terms from it. Annotations are stored in the central database.

We selected a random subset of 30 articles from the CNN/DM dataset. Annotators were provided with the annotation specification included in the appendix. They could annotate the summary sentence on their own computer, taking as much time as needed and revising annotations at any point if they so desired.

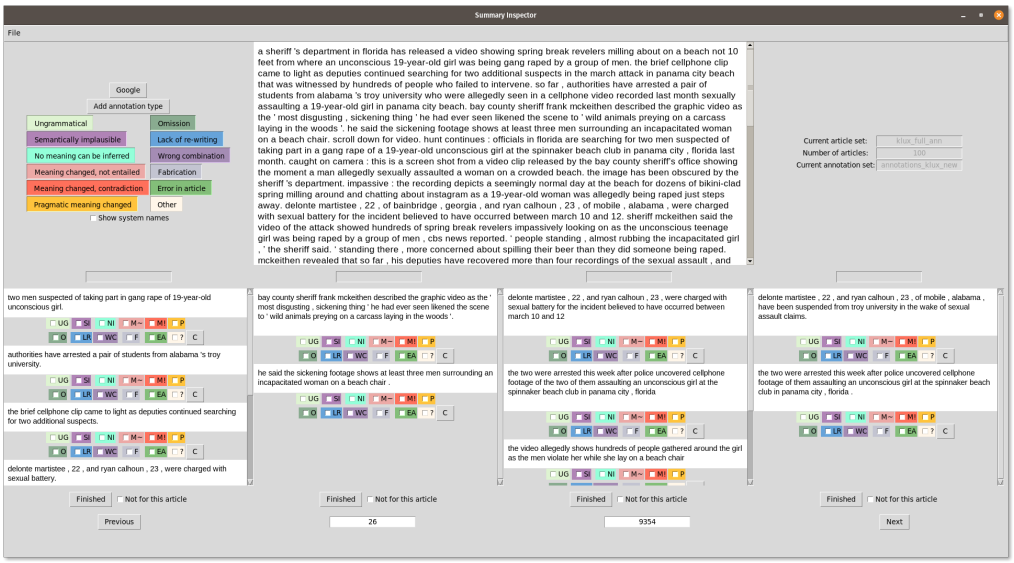


Figure 4: A screenshot of the SummaryInspector tool developed for this thesis to allow annotation of summaries

5.2 Results

After completion of the annotation, we used the stored annotations to compute the IAA of the group. This analysis was conducted separately for the two dimensions. We used the implementation of a standard IAA metric (Cohen's Kappa κ for multiple annotatators) included in the NLTK package.[3]. The metric reflects whether the observed agreement is substantially different from what would be expected by chance alone. $\kappa > 0.41$ reflects moderate agreement, while $\kappa > 0.61$ reflects substantial agreement. Beyond just determining the agreement for the original annotation schemes, we also experimented with reduced schemes by 1) merging categories and 2) aggregating sentence level errors to the summary level where appropriate. The results are presented in the following section.

5.2.1 Meaning dimension

Figure 5 shows the distribution of meaning dimension categories across sentences for the three different raters. While the exact prevalence of categories differs slightly, the overall picture is similar for all raters: Almost 80 % percent of summary sentences are judged not to contain an error of any kind. Less than 10 % percent of sentences are viewed as ungrammatical. The incidence rate for sentences that directly contradict the article is in a similar range. All other error categories are infrequent, affecting less than 5 % of summary sentences. It can be seen that some labels were never used by some annotators, which could be explained by the low incidence rate of errors of this kind. Looking at the inter-annatator agreement, the most fine-grained setting (i.e. the full scheme on the level of individual sentences) performs worst, reaching $\kappa = 0.44$. Table 4 shows that for four of the fine-grained classes, there was not a single case for which all raters unambiguously agreed it belongs to the respective class. For these classes, cases where at least two raters agreed are also rare, with most of their occurrences being accounted for by a single rater using them for a given summary sentence. Although some of this

[3] `https://www.nltk.org/_modules/nltk/metrics/agreement.html`

effect might be explained by a combination of a low incidence rate of certain error types and individual raters sometimes diverging randomly, it still seems to be problematic for ensuring a coherent annotation. Thus, various ways of aggregation were attempted to obtain a more reliable annotation scheme as reflected by fewer very sparse categories with little agreement and higher IAA, reflected in Table 5.

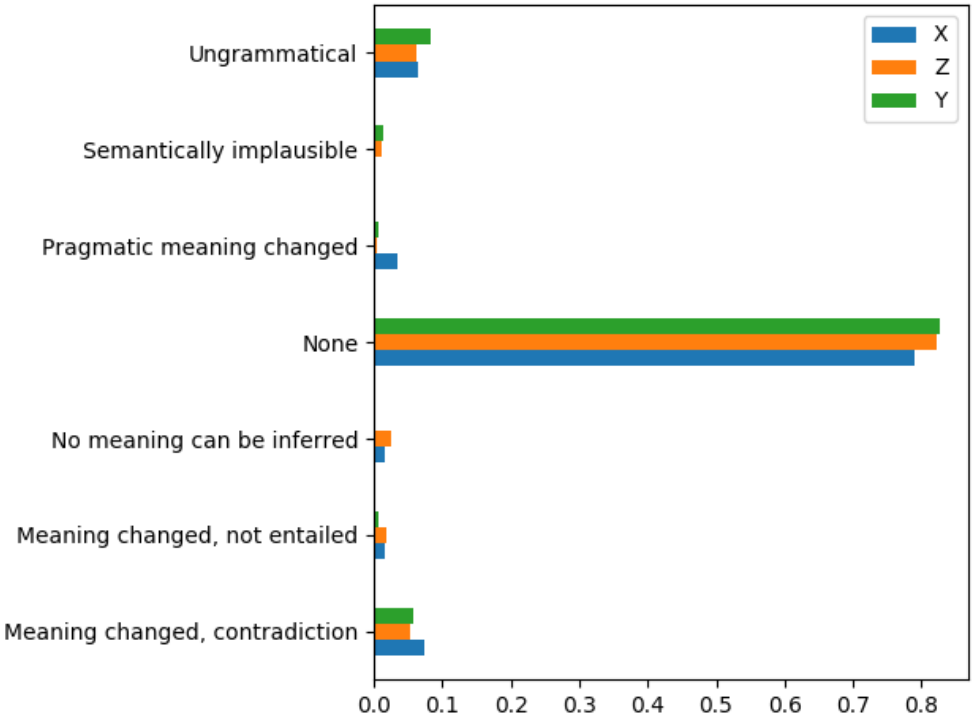


Figure 5: Summary-level incidence rate of different error effects for three raters.

	One rater	Two raters	Three raters
No error	38	55	284
Ungrammatical	26	17	9
Semantically implausible	10	1	0
No meaning can be inferred	14	2	0
Meaning changed, not entailed	12	3	0
Meaning changed, contradiction	21	11	11
Pragmatic meaning changed	17	1	0

Table 4: Absolute counts of sentences annotated by a single rater, two raters or all three raters as belonging to a particular meaning dimension category.

Grouping annotations into a ternary scheme based on error severity, i.e. distinguishing only between error-free, malformed and misleading sentences offers categories that can be annotated in a more reliable fashion ($\kappa = 0.46$). A binary split that merges all errors of any kind and contrasts them with error-free sentences yields $\kappa = 0.50$. This finding demonstrates that most of the disagreement is located not between effects, but rather caused by raters not agreeing whether a sentence contains an error at all.

There is no meaningful way to aggregate error effects from the full scheme to the summary level, as there is no way to decide which meaning dimension category the summary should receive in case multiple errors are present in its sentences. For the ternary scheme in contrast, this is possible due to the severity scale: A summary is labeled as error-free if and only if all sentences do not contain an error. It is labeled as malformed if contains at least one malformed sentence, but no misleading sentence. It is labeled as misleading if it contains at least one misleading sentence. Performing this procedure yields an agreement of $\kappa = 0.51$. The most coarse-grained scheme possible provides the upper-limit for what agreement is possible when almost all distinctions are leveled: By labeling a summary as erroneous if it contains one error of any kind, we achieve $\kappa = 0.57$. Thus, all schemes investigated only reach an IAA in the moderate range.

	Full split	Ternary split	Binary split
Sentence-level	0.44	0.46	0.50
Summary-level	-	0.51	0.57

Table 5: Effects of different granularities of annotation on inter-rater agreement: Multi-rater kappa scores, rounded to two digits of precision.

5.2.2 Mapping dimension

Figure 5 shows the distribution of categories in the mapping dimension across sentences for the three different raters. The general picture of prevalence rates is again similar between raters: As expected due to the design of the annotation specification scheme, the number of error-free sentences matches the one observed in the meaning dimension, accounting for roughly 80 % of sentences. About 10 % of errors are judged to have been caused by omissions. A somewhat smaller percentage is attributed to wrong combinations. The other categories are all fairly rare, occurring in less than 5 % of cases.

Looking at the agreement between raters (c.f. Table 6), we can see that agreement on the full typology is somewhat higher than for the meaning dimension, reaching $\kappa = 0.46$. For validation, mapping dimension annotations were also binarized and cumulated to summary level. As expected because every erroneous sentence was annotated for both dimensions, observed kappa scores are identical to the meaning dimension here.

	Full split	Binary split
Sentence-level	0.46	0.50
Summary-level	-	0.57

Table 6: Effects of different granularities of annotation on inter-rater agreement.

5.3 Analyzing disagreement

When raters disagreed, what did they disagree about? In line with the low gain in IAA obtained from binarizing annotations, the confusion matrices included in the appendix demonstrate that individual raters often agree on what the annotation for an error is - provided they also agree they agree there is an error at all. The vast majority of disagreements are caused by one rater indicating an error of some sort, while the other

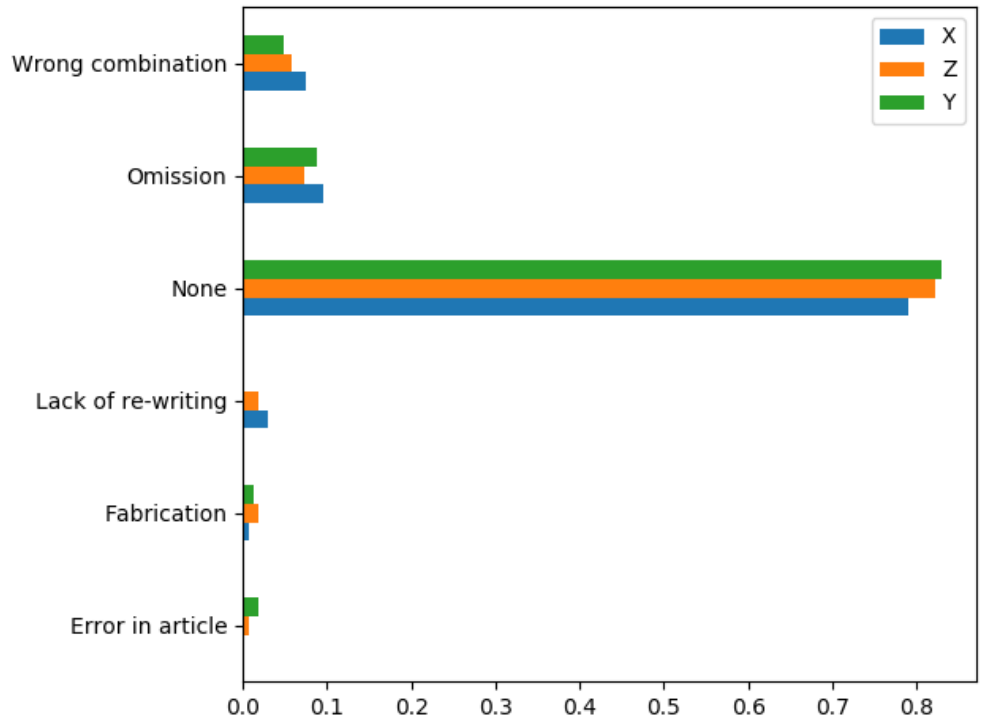


Figure 6: Sentence-level incidence rate of different categories in the mapping dimension for three raters.

indicated no error at all. For example, raters Y and Z disagree on the meaning dimension a total of 79 times and in 64 of these cases (more than 80 %) one of raters accepted the sentence while the other rejected it. For raters X and Y, disagreements of this sort make up 85 % of the total, for raters X and Z the percentage is somewhat lower at 76 %. When we see an error, we can usually agree on what on kind it is. There are two explanations of this effect. A subset of errors might be edge cases, causing raters whose inherent error tolerance is higher to let it pass, while others with a lower tolerance flag it up. Alternatively, some less glaring or more pernicious errors might also be harder to spot, causing some raters to occasionally miss them.

We can get a better understanding of these effects by pooling annotations of all three annotators and inspecting cases where one or two of the raters saw an issue, but there was no total agreement. Cases flagged up by two raters are more likely to just have been missed by the disagreeing rater. In contrast, cases flagged up by only a single rater might be especially pernicious and difficult to spot. We will first look at debated misleading cases and then at malformed cases.

5.3.1 Misleading sentences

We identified a total of 27 sentences that had been rated as error-free by two raters and as misleading by a third. 13 had been labelled as contradiction, ten as a change in pragmatic meaning and four as not entailed. These were manually inspected. In twenty cases, raters exhibited differing standards, with the disagreeing rater being less

accepting of edits than the others, c.f. Examples 11 and 12. In the remaining seven cases, an error had been missed by two raters, c.f. Examples 13 and 14.

Article context: [...] matthew upson closes down a leicester team-mate in training on thursday and could face west brom.. boss nigel pearson has no further injury worries as his rock bottom side continue to fight for barclays premier league survival. [...]
Summary sentence: nigel pearson has no further injury worries as his rock bottom side fight for survival .

Example 11: One rater labeled this as pragmatic meaning changed, two disagreed. Example of different standards.

Article context: [...] charlie adam (centre) lets fly from inside his own half as he scores one of the goals of his career at chelsea.. thibaut courtois watches on as he is unable to stop adam 's shot from hitting the back of the net for 1-1. [...]
Summary sentence: thibaut courtois is unable to stop adam 's shot from hitting the net for 1-1.

Example 12: One rater labeled this as meaning changed, not entailed, two disagreed. Example of different standards.

Article context: [...] michelle schwab, who has three sons and a degree in therapeutic childcare, has been charged with child endangerment after she allegedly dangled her child over the 10-foot-deep enclosure in cleveland metroparks zoo before he slipped and fell.. [...] on monday, a spokesman for kindercare, a nationally-acclaimed education , care and resource provider, confirmed schwab has taken a leave of absence from her management role at one of the centers in columbus, ohio. [...]
Summary sentence: schwab is a nationally-acclaimed education , care and resource provider.

Example 13: Meaning changed, contradiction, missed by two raters.

Article context: a florida mother has accused a school of threatening to suspend her five-year-old autistic son because of ' essential oils ' he wears to help combat his illness.. jessica kemp from eustis says teachers at seminole county elementary warned they would remove kindergartner logan from class because the products, manufactured by doterra, smell and are a distraction to youngsters around him. [...]
Summary sentence: teachers at seminole county elementary warned they would remove kindergartner logan from class.

Example 14: Pragmatic meaning changed, missed by two raters.

We found a total of seven cases that featured two raters agreeing a sentence was misleading, while the other one labeled it as error-free. Four cases were labelled as contradictions, two as not entailed and one as a change in pragmatic meaning. In all seven cases, the review showed the sentence to be indeed erroneous, c.f. Examples 15 and 16.

Article context: [...] its central trunk is hollow with six smaller ones branching off - possible due to disease - and locals fear it is on its way out due to wilting branches and falling leaves.. [...] experts say it is 3,500 to 4,000 years old. [...]
Summary sentence: experts fear it is on its way out due to wilting branches and falling leaves .

Example 15: Meaning changed, not entailed, missed by one rater.

Article context: [...] the are no premier league clubs in the last eight of the [champions league] after chelsea , arsenal and manchester city were eliminated at the last 16 stage, while everton were dumped out of the europa league in the previous round. [...]
Summary sentence: chelsea , arsenal and manchester city were dumped out of the europa league.

Example 16: Meaning changed, contradiction, missed by one rater.

These findings demonstrate that the annotation task is not trivial and requires maintaining close attention: A total of 14 misleading sentences were missed entirely by at least one rater. Often, these sentences are perfectly plausible at the surface (c.f. Examples 13 - 16). Similarly, there is often at least some judgment involved in deciding whether a given sentence is actually misleading. We found 20 examples judged misleading by one rater and acceptable by two others where the differences were caused by different personal views on whether certain edits had faithfully retained original meaning. It thus appeared that this aspect is less clear-cut than previously believed.

5.3.2 Malformed sentences

We found a total of 27 sentences that were malformed according to one rater but fine according to two others. In 15 of these cases, the rater viewed the sentence as ungrammatical, eleven cases were judged as *No meaning can be inferred* and one case was judged to be semantically implausible. 16 cases were found where two raters had agreed the sentence was malformed, while one rater disagreed. 14 of these cases were labelled as ungrammatical, one as semantically implausible and one as a case of *No meaning can be inferred*. These findings demonstrate that there is considerable disagreement in what constitutes an ungrammatical sentence. An inspection of cases revealed that disagreement were often related to telegraphic language style, c.f. Example 17. Even though explicitly instructed not to flag up this style of language, raters sometimes diverged, presumably because their tolerance was lower. As this style is quite prevalent in reference summaries, especially for the Daily Mail, systems can be expected to imitate it, meaning that future annotation efforts should try to make raters more familiar with this somewhat unusual style if this dataset is to be used. Similarly, raters seem to exhibit different standards regarding the conditions under which meaning can be inferred

from an article. This can also be explained by the peculiarities of the dataset used for training the models: As mentioned in Section 2.6.1, some reference summaries are hard to parse without access to the headline. If models pick up this aspect, summaries might exhibit similar properties, implying that raters who differ in how much they are willing to accommodate will disagree frequently.

Article context: [...] west brom.. craig dawson is set to return for west brom after serving a one-match ban. [...]
Summary sentence: craig dawson set to return for west brom after serving one-match ban

Example 17: Labeled as ungrammatical by one rater. Example of telegraphic language style.

5.4 Conclusion

RQ2 asked in whether there is meaningful human agreement on errors in generated summaries. In conclusion, we managed to reach *moderate* agreement between raters using the original fine-grained typology for both dimensions of the typology. A number of aggregation operations are available and increase observed agreement, but are not sufficient to reach an agreement considered *substantial*. Some specific properties of the CNN/DM dataset can be speculated to have negatively affected agreement about the malformedness of sentences, namely telegraphic language style and the issue of reference summaries lacking relevant context. Looking specifically at misleading errors, we can see that there while there are a number of cases where all raters unambiguously agree, there is also at least some disagreement. Some of it is caused by especially pernicious summary sentences that subtly alter the meaning and are difficult to spot, potentially risking that even if manual control of summary outputs were imposed, some misleading sentences might still be released. A more substantial source is general disagreement on what constitutes a good retention of article meaning.

However, observed agreement is still high enough to allow a meaningful analysis of the differences between individual summarization systems. For this purpose, we will rely on the ternary scheme, as it offers the best trade-off between agreement and granularity. The next section first investigates how our initial hypotheses about the differences between the systems map to the annotation scheme and then analyzes the differences in error prevalence on a larger set of articles.

6 Comparing systems on the original test set

How do our initial hypotheses about differences between systems map to the ternary annotation scheme? For **RQ3**, we are mostly interested in how different ways of involving an extractive step affect the error rates, with no specific prior intuitions about different error types. For **RQ4**, looking into the effect of pre-training, we had three specific intuitions:

1. A reduction in errors that reflect an insufficient grasp of the dependency structure. We did not create an error category that maps directly to this idea. Errors labeled as a wrong combination seem to be mostly directly related to this aspect, but other errors can also be conceived as to be caused by it, e.g. a sentence that is ungrammatical because some necessary phrase was deleted.

2. A reduction in semantically implausible sentences. This category is directly part of the full annotation scheme, but in the ternary scheme, it is no longer present, having been merged into the malformed category.

3. An increase in fabrications. We have an mapping dimension category that directly captures this aspect.

While the third intuition can directly be investigated, the first two required some adjustments. The first is based on the assumption that pre-training helps the model to obtain syntactic knowledge, possibly avoiding ungrammatical sentences. The second deals with surface-level semantic knowledge, helping to avoid implausible sentences. We capture both these types of error in the malformed category, whose absence can be thought to reflect general surface-level linguistic skills. In sum, our initial intuitions will be converted into the following predictions:

1. We will combine Intuition 1 and 2, yielding a prediction that the **incidence of malformed sentences should be lower** for pre-trained models.

2. The aspect of dependency in Intuition 1 will be further captured by predicting a **lower incidence rate of errors caused by wrong combination and omission** generally (for both malformed and misleading sentences).

3. Intuition 3 is captured by stating our expectation that **fabrications will be more prevalent** for pre-trained models.

6.1 Methods

We have established a typology for sentence-level error annotation of summaries and outlined how it relates to our prior expectations. We now set out to use it for an annotation of a set of summaries generated for the original test set articles. We annotated a total of 200 articles with four summaries each, yielding a total of 800 summaries with roughly 2600 annotated sentences. For thirty articles, annotations from three annotators were available. Cases where the majority of raters agreed were assigned to the class preferred by the majority. For cases where there was no general agreement, arbitration was used to reach agreement. After that, 170 randomly selected additional articles were annotated by the main author and then combined to yield the final dataset for analysis.

6.2 Results

Table 7 presents the most coarse-grained view of differences between summaries, looking into the binary error rates for the four different systems. We can see that on the sentence-level, PreSumm and See are the least error-prone, with LM in third place and Chen last, suffering from an error rate that is 9 percentage points higher than that of PreSumm. On the summary level, the error rates are higher, with around 40 % summaries containing at least one error of any kind for See, LM and PreSumm and Chen faring much worse at almost 75 %. We also look at the average summary length in sentences and compute the expected number of summary errors if sentence-level errors were completely independent as:

$$exp_{summ} = 1 - ((1 - er_{sent})^{n_{sent}})$$

where exp_{summ} is the expected rate, er_{sent} is the observed sentence error rate and n_{sent} is the average number of sentences.

For all systems, the observed error rate is closely aligned with the expected error rate, implying that sentence-errors are likely randomly distributed across summaries.

The numbers also demonstrate that Chen's high summary level error is explained by a combination of its high sentence level error rate and the fact that sentences are much longer with an average of almost five sentences. LM is outperformed by the other two systems on the sentence level, but performs best on the summary level, as average summary length is very low at 2.27.

System	See	Chen	LM	PreSumm
Sentence error rate	0.16	0.24	0.20	0.15
# sentences / summary	2.91	4.93	2.27	3.33
Expected summary error rate	0.39	0.74	0.39	0.41
Observed summary error rate	0.38	0.73	0.36	0.39

Table 7: Binary error rates, sentence- and summary-level.

To better understand the differences between systems, we break down the errors using the meaning dimension. First, Figure 7 shows the incidence rates of malformed and misleading sentences for the systems. We can see that all systems produce both types of error, but the distribution is quite different. See produces the fewest misleading sentences (incidence rate: 4.5 %), while the rates for errors of this type are higher for the three other systems, with Chen and PreSumm both at roughly 8.0 % and LM producing errors of this type almost 2.5 times as frequently as See at 11.2 %. Malformed sentences are much more common for See and Chen, at 11.3 % and 16.0 %, respectively, while LM and PreSumm can often avoid them, obtaining incidence rates of 8.4 % and 6.7 %, respectively.

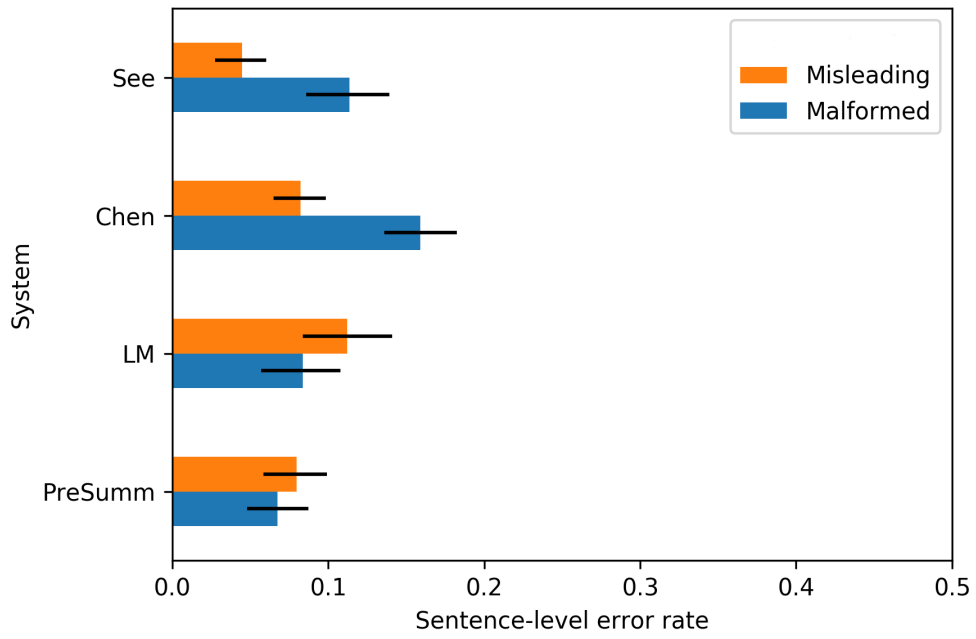


Figure 7: Sentence-level error type incidence rates by system. 95 % confidence intervals (CI) obtained by bootstrapping.

The general pattern is similar on the summary level (c.f. Figure 8), although the incidence rates there are higher, in line with the observations made above. Again, See has

the lowest rate (11.5 % of summaries contain at least one misleading sentence), with the other three systems performing worse. However, under this scheme, LM and PreSumm (19 % / 22.5 %) perform better than Chen, which has the highest rate at 32.0 %. In other words, between roughly 1 in 10 and 1 in 3 summaries generated by a number of current summarization systems contain at least one misleading statement. This level of analysis also allows a comparison with the numbers reported by [3], who gave estimates for the incidence rates of factual errors on the summary level. Their estimates for See (8 %) and Chen (26 %) are both somewhat lower than our observations, but the general trend is reflected. Looking at malformed summaries, we can see that the relative performance of the systems is the same as on the sentence level, with higher incidence rates across the board. PreSumm and LM perform best, producing malformed summaries roughly 1 in 6 times at 16.0 % and 16.5 %, respectively. See is somewhat worse at 26 %, while Chen is much worse at 41 %.

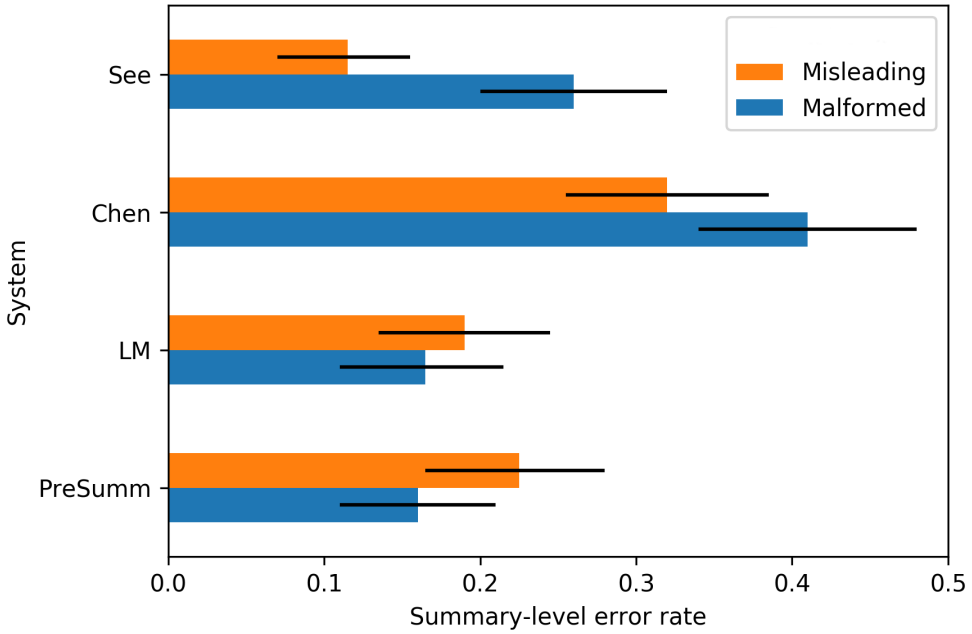


Figure 8: Summary-level error type incidence rates by system. 95 % CI.

Next, we look into the interaction between the two dimensions. Figure 9 shows the sentence-level counts of different categories in the mapping dimension across systems, separated by the meaning dimension. We can see that categories differ markedly both in absolute prevalence as well as in their association with categories in the meaning dimension: Omissions are most frequent and quite strongly associated with malformedness (roughly two thirds of omissions are malformed sentences). Lack of re-writing is less frequent and even more strongly associated sentences being malformed (almost 80 percent of sentences of this type are malformed) - this can be explained by a large number of sentences being removed from their surrounding context without adequate re-writing causing problems with inferring sentence meaning. Wrong combination is about as frequent, but more strongly associated with misleading sentences (roughly 60 % fall into this category). Fabrication is less frequent still, but has the strongest association with sentences being misleading (almost 70 %). Systems also differ markedly in the interac-

tion between the two dimensions, c.f. Figure 10. When a sentence is misleading, this is almost never due to a fabrication for See and Chen, but in almost a quarter of cases for PreSumm and LM. In contrast, both for misleading and malformed sentences omissions and wrong combinations are more prevalent for See and Chen, and less frequent for LM and PreSumm. The difference is more pronounced for malformed sentences. Chen suffers markedly from omissions, these account for almost 80 % of all misleading sentences, while being much rarer for the other systems. The pattern is similar for malformed sentences, demonstrating that Chen has a general tendency to omit words to a variety of effects.

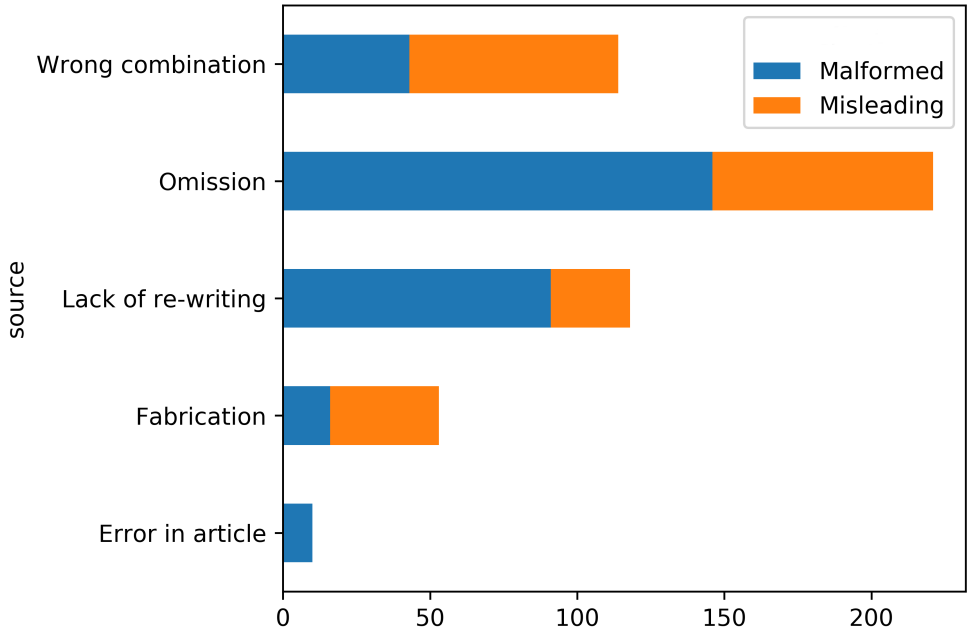


Figure 9: Incidence rates of mapping dimension categories, separated by meaning dimension.

While sentence-level error rates are interesting for researchers intent on understanding what aspects of a system might be involved in its performance, practitioners are likely to focus more on the big picture, asking which system is least likely to make an error across the board. To answer this question, we need to compare summary quality across articles. Summary-level rates can offer some guidance already, but we also perform a head-to-head comparison on individual articles: As there is a relative preference for an error-free summary over a malformed one and for a malformed one over a misleading one, we can rank the four summaries produced by the systems, awarding the best rank if a summary is error-free and lower ranks for malformed and misleading summaries, respectively. The average ranks of systems across articles can then be compared using a non-parametric test known as Friedman test. Its null hypothesis is that there is no difference between ranks, meaning that in the limit, one would not be better off using any system over the other. We performed this analysis, first ranking system summaries across the available 200 articles and then computing average ranks and the test statistic. In line with standard practice, we set the decision threshold $\alpha = 0.05$. As we find $p = 0.0189 < \alpha$, we refute the null hypothesis and conclude there is a sta-

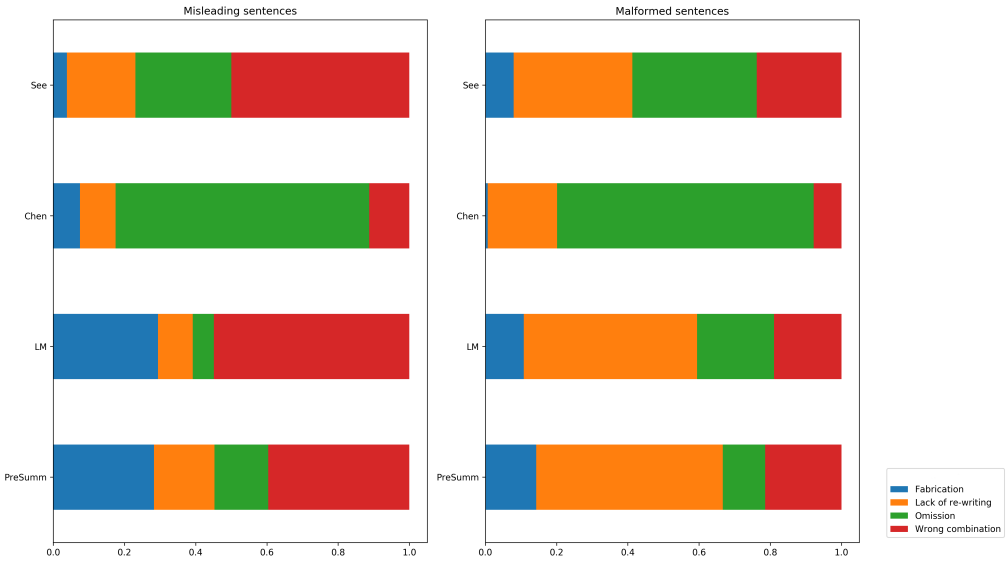


Figure 10: Distribution of mapping dimension categories by system, separately for different meaning dimension categories.

tistically significant difference between the systems. We are thus licensed to perform a post-test that compares the differences between systems. The results of the Nemenyi test are visualized in Figure 11. We can see that Chen is substantially worse than the other three systems, which are not significantly different from one another.

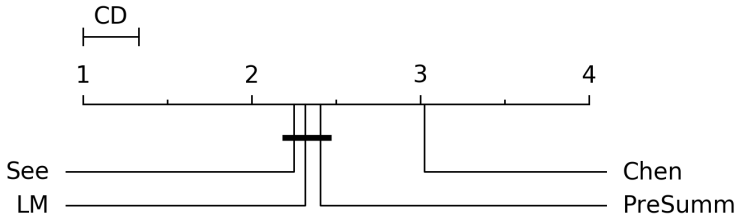


Figure 11: Results of Nemenyi test. Average ranks of different systems indicated on the axis. Non-significantly different conditions connected by bold line. CD = critical difference.

Are some articles consistently difficult for all systems? When one makes a mistake, is another system more likely to err as well? To answer these questions, we look at the article level, inspecting the distribution of summary errors per article (c.f. Figure 12) and the pair-wise correlations between the error incidences of different systems (c.f. Figure 13). Looking at the distribution of errors, we can see that it is somewhat skewed towards the left, there being more articles for which none or few systems make a mistake than those for which all or almost all do. The figure also contains the distribution we would expect if summary errors were totally random (sampled separately with the observed error rate for each system). We can see that there is a slight divergence between the two distributions, such that in the observed case, there is a somewhat higher than expected chance for all or few systems to err. This provides some preliminary evidence that some

articles might be inherently harder or easier than others. In a similar vein, we can observe positive correlations between error prevalence between all pairs of systems, even though the correlations strength is fairly weak in most cases.

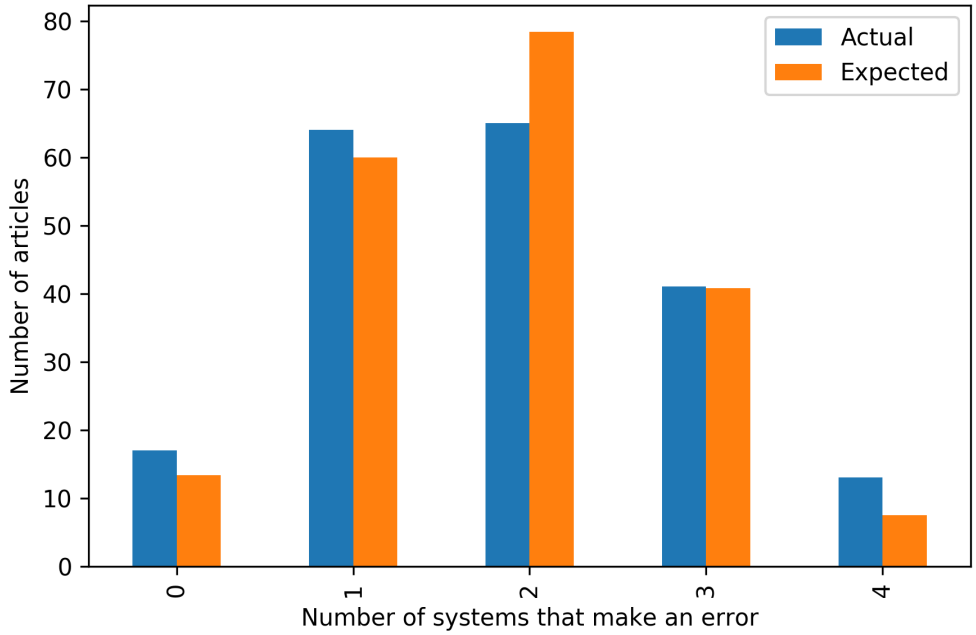


Figure 12: Expected and actual distributions of articles by how many systems make an error for them.

As there have been no prior detailed analyses of sentence-level error rate, there is also no prior investigation into what factors might be mediating any possible relation between system and observed error rates. Looking into mediators can provide us some understanding into what properties of the output of summarization systems are associated with decreased error rates.

Abstractiveness could potentially be a very important mediator between system and error rate. This makes intuitive sense, as a summary that is not abstractive at all is much less likely to suffer from errors. To some extent, this also applies to individual sentences in isolation: When copied over verbatim, the only applicable mapping dimension categories are error are an error in the article and a lack of re-writing, if the original sentence needed its surrounding context to make sense. In this way, there is only a slight chance for the sentence to contain an error, when compared to the ample potential for mistakes if more editing takes place. It could thus be expected that systems that are more abstractive are also more error-prone, unless they are inherently more capable of correctly abstracting sentences than others.

We first need a better understanding of how much variance in abstractiveness there is between systems and how they compare to reference summaries. Figure 14 contains density plots of two important quantities for all sentences in the reference summaries / the system output on the annotated dataset. For each sentence, we automatically select the closest document sentence in terms of word overlap. We then compute ROUGE-L, i.e. the ratio between the length of the longest common subsequence and the length of the article sentence / summary sentence, respectively. Normalizing by the length of the

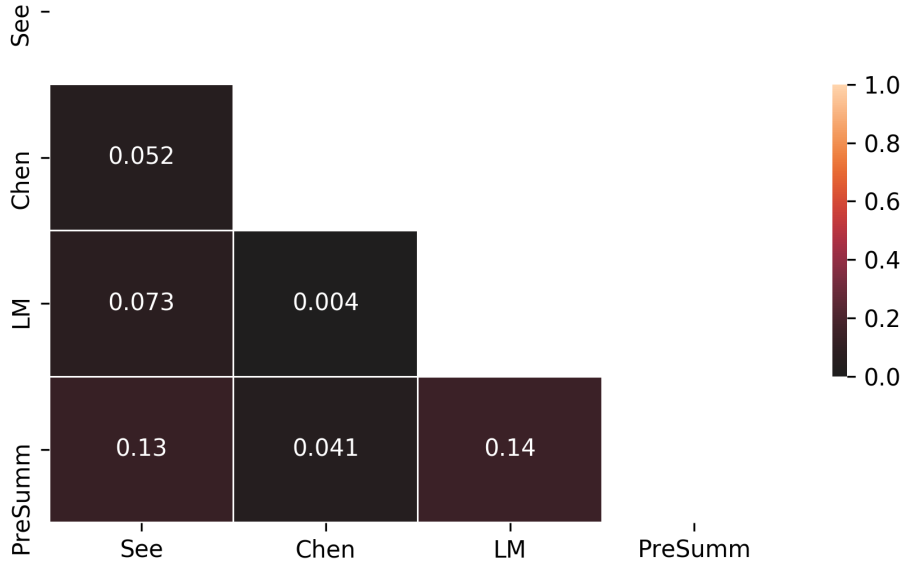


Figure 13: Pair-wise correlation of article-level errors between systems.

article sentence gives the precision of ROUGE-L and can thus provide us an understanding into how much of the article sentence is retained. In contrast, normalizing by the summary length gives the recall of ROUGE-L, elucidating how much of the summary originates from the closest document sentence. Looking at the reference summaries first, we can see a great deal of abstractiveness, with the majority of summary sentences clearly incorporating words from more than one sentence and few document sentences being retained in their entirety. The picture is starkly different for all four systems: See is the least abstractive, with the overwhelming majority of summary sentences scoring close to 1 on both precision and recall and thus being almost exact copies of a document sentence. Chen is very similar in terms of recall, but shows more variance in terms of precision. Sentences with a high recall but a smaller precision are examples of deletions: Some words from the article sentence are deleted, but there is no influx of new words. Sentences of this type are fairly frequent for Chen, and thus the system can be said to be prone towards deletions. PreSumm occupies a middle ground between See and Chen: Precision is somewhat lower than for See, but higher than for Chen. This system is also fairly extractive, but engages somewhat more in deletions than See. Recall is somewhat lower, so there are some sentences that cannot be traced back in their entirety to one article sentence. Finally, LM exhibits the greatest variance across the spectrum. Recall is generally a bit higher than for Chen, but varies considerably, as does precision. This system can thus be said to be most abstractive, generating sentences that miss some words from a source document and contain words from other sentences as well. Clearly, there are marked differences in abstractiveness between the systems. Do these also affect error rates?

To tackle this question, we first computed the F1 score, the harmonic mean of precision and recall for all ROUGE values. Sentences were then binned into two equal size bins, yielding a threshold of 0.705. We label sentences with a ROUGE-L-F1 below the

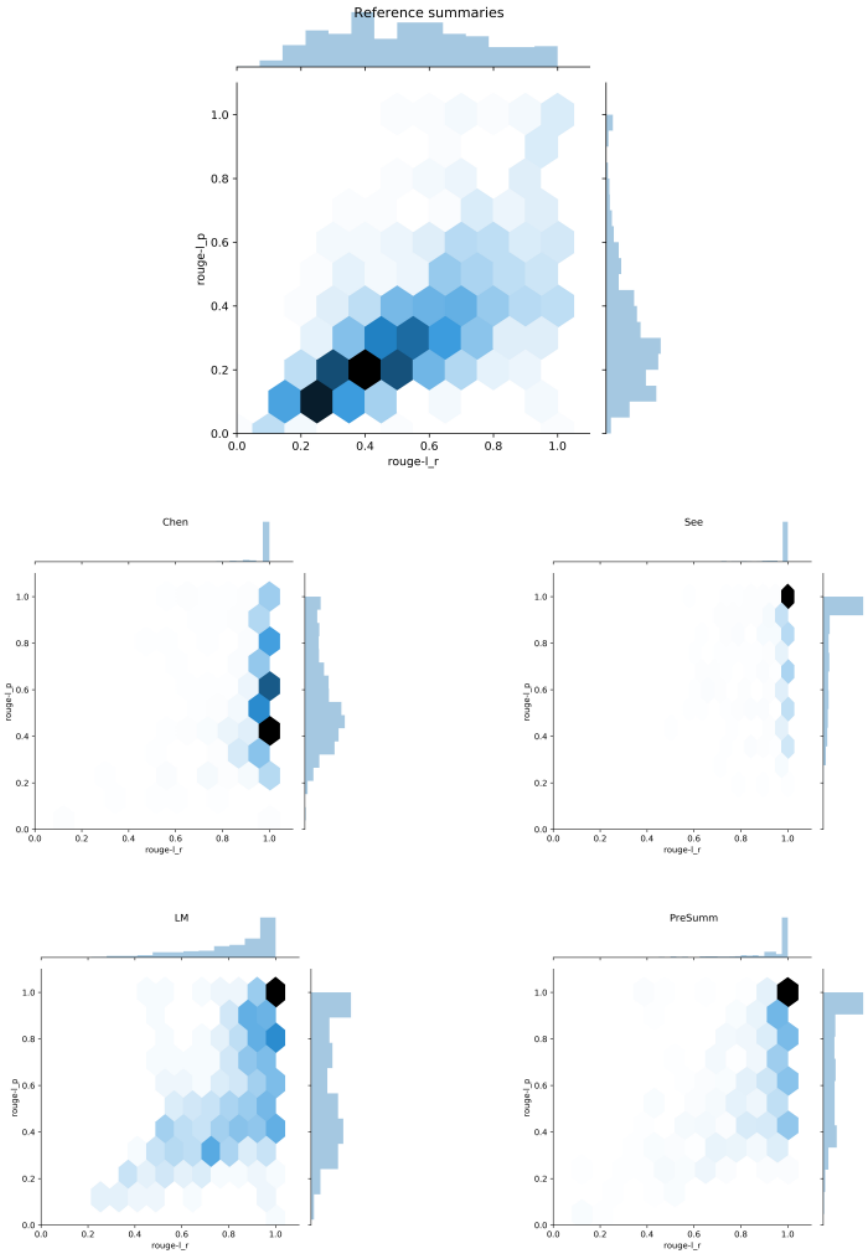


Figure 14: Distribution of sentences across ROUGE-L recall (x-axis) and ROUGE-L precision (y-axis) for reference summaries and summaries by the four different systems. ROUGE to closest document sentence.

threshold as highly abstractive. Figure 15 shows what percentages of sentences for each system falls into either category. Again, we can clearly see that See and PreSumm are fairly extractive, while LM and Chen are more abstractive. Averaged across all sentences, we can also see that higher abstractiveness is associated with a higher error rate (c.f. Figure 16). Sentences that score high in abstractiveness are more than twice as likely

to be misleading and 50 % more likely to be malformed than those that score low. The crucial question is now the exact nature of the relation between abstractiveness, system and error rate. Either, all systems are more likely to make an error when they generate more abstractive sentences. Alternatively, some systems are more abstractive and also more-error prone. The interaction plot in Figure 17 provides a preliminary answer to this question: We can see that the general pattern of higher abstractiveness being associated with higher error rates is stable across systems. We cannot conclusively establish that some systems are inherently better at writing abstractive sentences, instead they all perform about equally in this category. For largely extractive sentences, See, LM and PreSumm perform about equally well, while Chen is markedly worse. These findings thus support the idea that the difference in the propensity of systems to create more abstractive sentences mediates the differences in error rates between systems. In other words, when we observe an absolute difference in sentence error rate between systems, this difference could also be explained not by one system being inherently better, but just being less likely to write more abstractively and thus more error-prone.

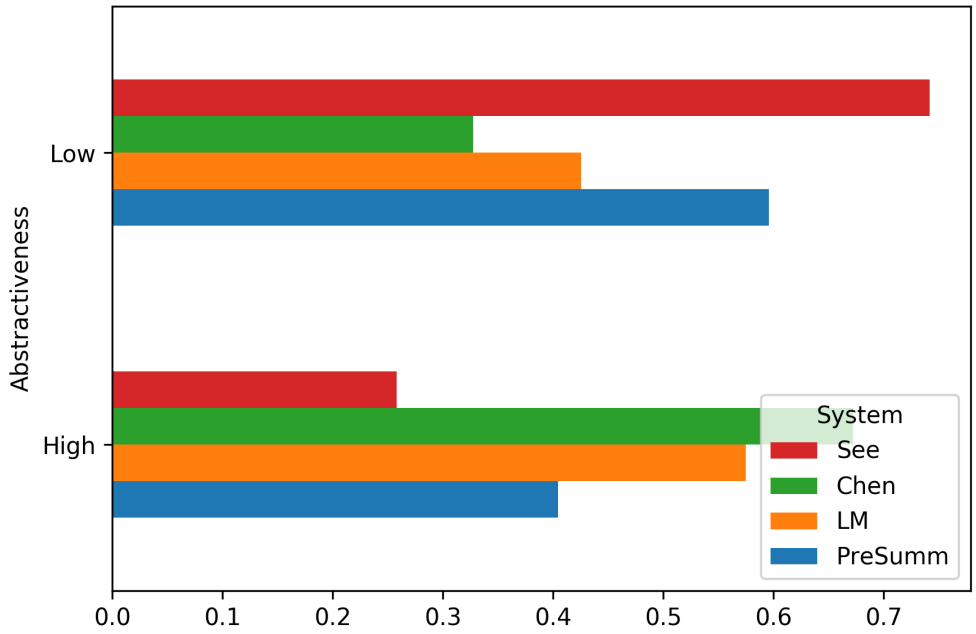


Figure 15: Binned ROUGE-F1 scores (threshold: $ROUGE - F1 = 0.705$). Breakdown of all sentences generated by each system by bin.

6.3 Discussion

In this section, we set out to get an understanding of what errors current state-of-the-art summarization systems make and how they differ from one another in this aspect. Overall, we found that no system is a magic bullet, with all of them fairly frequently generating malformed and misleading summaries. If one were to use any of the current systems in a real-world scenario, readers could frequently end up confused, irritated or worst of all misled to hold incorrect beliefs. We could establish that one system was demonstrably worse in a head-to-head comparison on the article level: Chen's sentence re-writing system has a substantially higher summary-level error rate than other sys-

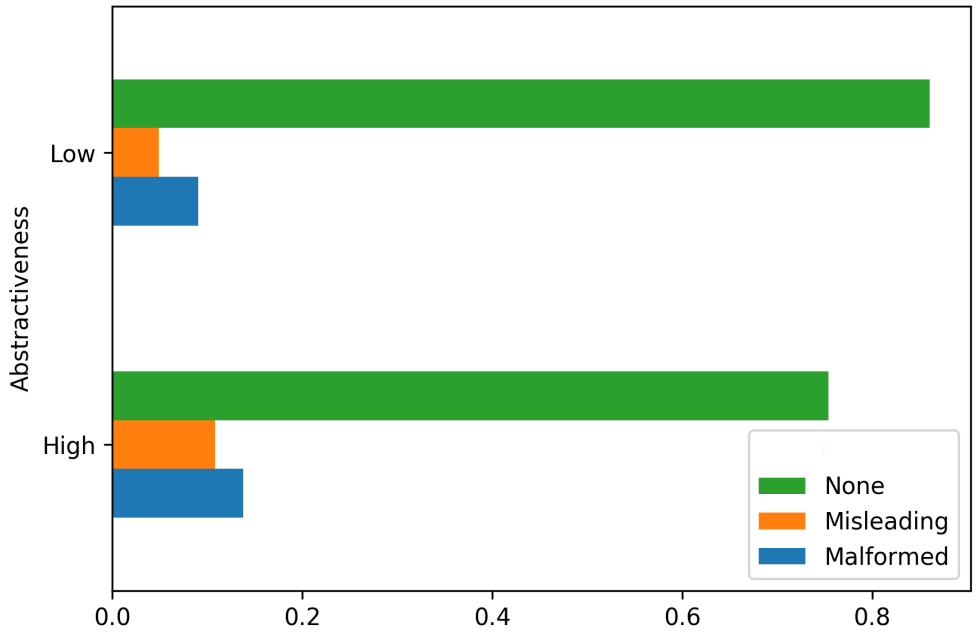


Figure 16: Binned ROUGE-F1 scores, average error rates in bins across all sentences.

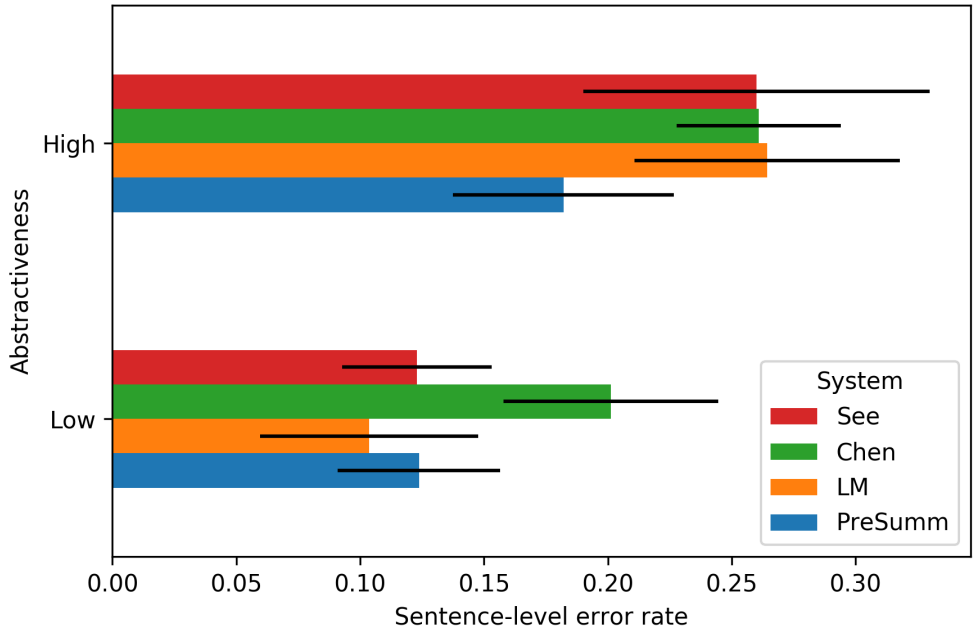


Figure 17: Binned ROUGE-F1 scores, average error rates in bins separately by system. 95 % CI.

tems. This can be explained by individual sentences being frequently malformed due to omissions and the high number of sentences being generated. It appears that the

separately trained abstraction component has not learnt how to correctly re-write sentences, often deleting incomplete phrases and most of its errors being judged to be caused by omissions. PreSumm, which only performs extractive fine-tuning, suffers much less from this problem. The other two systems do not involve extraction at all also perform similarly. This provides our answer to **RQ3**: A complete separation of extraction and abstraction seems to have a negative impact on performance. An extractive fine-tuning step is a better way to involve extraction, but it could not be shown to measurably improve performance compared to systems that do not leverage extraction at all. Chen's abstractor is very aggressive in deleting, as evidenced by the high proportion of summary sentences that have a low ROUGE-L precision with the document sentence. More research into the paradigm should look into when separate abstractors fail and what re-writing patterns they can even learn by inspecting the automatically generated pairs of summary and document sentences: There is quite a substantial assumption involved, namely that summary sentences matched in this way constitute correct re-phrasings. Additionally, the abstractor works in isolation, having no access to the document or summary and thus will not be able to properly model surrounding context, which might also negatively impact performance.

We could not establish significant differences in article-level performance between the remaining systems, meaning that none is directly preferable from a practical perspective. There are a number of interesting patterns regardless. Focusing on the two models leveraging pre-training for **RQ4**, we can see that they produce fewer malformed sentences than either of the systems trained from scratch. This category comprises ungrammatical and semantically implausible sentences. Correspondingly, **prediction 1**, which states that pre-training helps to foster the general linguistic ability of summarization models, is **validated**. Looking at the breakdown of errors by mapping dimension, we established that wrong combinations and omissions were also less associated with misleading and malformed sentences generated by pre-trained systems, **validating prediction 2**. However, on the downside, these systems are no better at avoiding misleading sentences, instead producing them at a rate similar to the worst system trained from scratch in case of PreSumm and even more frequently in case of LM. We can see that a larger proportion of these errors are due to downright fabrication of words for both pre-trained systems. This evidence **supports prediction 3** (pre-trained systems are more prone to fabricating). Summing up the answer to **RQ4**, the effects of pre-training appear to be a mixed blessing: While it makes for summaries of a higher linguistic quality that are easier to parse, it does not help avoiding misleading sentences and even produces errors of a nature that might make them harder to detect.

One important caveat in this analysis is the fact that there is still an important property which we could not control for, namely the decoding algorithm. While all systems use beam search, the beam size and some tweaks used (e.g. Trigram blocking) vary between systems. It would have been preferable to standardize these aspects, but this was not possible, as it would have required extensive modification of the existing code. For current purposes, we assume that as the properties were selected by the original authors to offer what they consider good performance, the differences should not have a strong impact on the observed differences in performance. Moreover, while some variance due to beam size might be expected, the values used were in a relatively small range, so the effects are presumably small as well. However, future research should aim to standardize this aspect further.

We also found substantial differences in abstractiveness between systems and looked into the relation between how frequently systems produced sentences that diverged

markedly from the closest document sentence and their error rate. It could be established that higher abstractiveness is associated with a higher error rate across the board. No system managed to avoid this pattern and it has implications for future comparisons between systems: The degree of abstractiveness is an important mediator of error rate and needs to be controlled for a fair comparison, otherwise a less abstractive system inherently looks better on comparisons of error rate. Ideally, we should look for a system that has a similar pattern of abstractiveness to the original summaries, but does not suffer from the association between abstractiveness and error rate. If we were to rely on a single system based on current evidence, it would be likely be See, which was demonstrated to be least error prone in line with some prior research. However, this system mainly works well because it just is not very abstractive. Essentially, we might wonder what we gain compared to a purely extractive system - this would presumably produce an even smaller number of errors and given See's low abstractiveness, the differences in output might be negligible. In this way, there is still much more research to be done to get a system that delivers both on abstractiveness and factual correctness.

7 Comparing systems on newer articles

For **RQ5** and **RQ6**, we are interested in the robustness of different systems to changes in their input. Specifically, we select novel articles from the original outlets and inspect the summaries generated, using the same annotation scheme as before. We try to distinguish the effects of mere recency as opposed to topical novelty. We hypothesize that new articles will be associated with higher error rates than test set articles and that this pattern is stable across systems. Additionally, we predict that this effect is most pronounced for topically novel articles, with those that are more topically similar to older articles giving rise to lower error rates.

7.1 Methods

To perform summarization on unseen data, two tasks have to be performed. Firstly, articles have to be obtained, downloaded and the text extracted from them. Afterwards, the code shared by the original authors of the papers have to be adapted as it was not explicitly designed for the purpose of performing on-line inference on new articles, but rather only for bulk inference on a fixed test set. After having arrived at a workable pipeline, we experiment with a method to identify topical novelty. Finally, we annotate a number of articles.

7.1.1 Obtaining and processing new articles

The authors who collected the CNN/DM dataset [27] do not provide any information on how exactly articles were scraped. We found that both outlets offered sitemaps for specific date ranges (month-wise in the case of CNN[4], day-wise for Daily Mail[5]). It could not be established whether these sitemaps were used by the original authors, but they were manually inspected and found to be comprehensive, containing a large number of articles on a diverse set of topics. We were able to access the original source code[6] used by [27]. It was adapted to work directly on URLs from the CNN/DM pages, obviating the need to first archive pages into the Web Archive before downloading. We also made some adaptations to the expressions used for extracting text and story highlights (*i.e.*

[4]Example sitemap for January 2020: `https://cnn.com/article/sitemap-2020-01.html`

[5]Example sitemap in XML format for January 1, 2020: `https://www.dailymail.co.uk/sitemap-articles-day~2020-01-01.xml`

[6]`https://github.com/deepmind/rc-data`

reference summaries), as the HTML structure of both sites had changed slightly since 2015[7]. With these changes in place, we were now able to download articles published over a desired period and to save them in the original story format for further processing.

7.1.2 Performing inference on new articles

For all four systems evaluated above, we were able to obtain trained models. Additionally, there was also source code for preprocessing the original dataset and for performing inference on preprocessed data. We initially experimented with enabling complete online inference (i.e. providing an endpoint to which an unprocessed article can be sent to obtain a summary), but later settled for retaining the original bulk-inference that all systems already provided for performing inference on the validation / test set due to the higher stability of this approach. We apply the preprocessing code to newly extracted stories, swapp out the original test set for the newly created set and then perform bulk-inference by relying on the original code. Generated summaries are then ingested into the already existing database format and can thus directly be inspected and annotated using the SummaryInspector tool.

7.1.3 Measuring topical novelty

As we were interested in the effects of topical novelty, we needed some way to operationalize this concept. Rather than trying to manually establish whether a given article is novel or not, we built an automatic method, leveraging topic modelling via Latent Dirichlet Attribution (LDA) as introduced by [30]. This method finds a pre-defined number of topics in a text and attributes each document to one or more of them. We were especially interested in topics whose presence differed for old and new articles, as those were likely to either be old topics no longer covered or more recent topics not reflected in the original data. For each of the two news outlets, we performed the following procedure to obtain a measure of how topically novel a given article is:

1. Download and parse recent articles. We downloaded all articles from January 2020, this yielded roughly 3600 articles for CNN and over 10000 articles for DM. The latter set was randomly downsampled to a set of the same size as the CNN data.

2. Download and parse old articles. We selected articles from past years. For each year, the same number of article URLs from the training set was randomly selected and articles were downloaded and parsed. We selected 2000 articles per year.

3. Fit a topic model on the combined set of recent and older articles. Articles were preprocessed, performing bigram extraction, stopword removal and removal of words with certain POS tags, details can be found in the appendix. We performed grid search for some LDA hyperparameters, also described in the appendix. The model maximizing the coherence score C_V was chosen, as this score has been shown to have good correlations with human ratings. [31].

4. For each of the topics in the resulting model, we computed how much of the total probability mass allocated to it was found in recent vs. older articles, respectively. We then computed the ratio between these two terms. This ratio informs us whether a topic is more prevalent among recent or older articles. If there was no difference, we would expect the ratio to be similar to the ratio between the number

[7]The tags of article paragraphs had been renamed in the case of CNN. For the Daily Mail, the formatting of story highlights differed, though only in the tags used, not in the graphical presentation on the page.

of recent and older documents. The bigger the ratio, the stronger the bias towards more recent articles.

5. For each article, we obtain its distribution over topics. We compute the sum over the novelty ratio defined above, weighted by the probability for each topic. This yield a topical novelty score for each article. Articles with a higher score can be said to be dominated by topics that are more associated with newer articles, while those with a lower score consist of topics with that are more associated with older articles.

Having arrived at this metric, we should now be able to distinguish articles by how topically novel they are. The empirical distributions of the scores are visualized separately for the two news outlets in Figure 18.

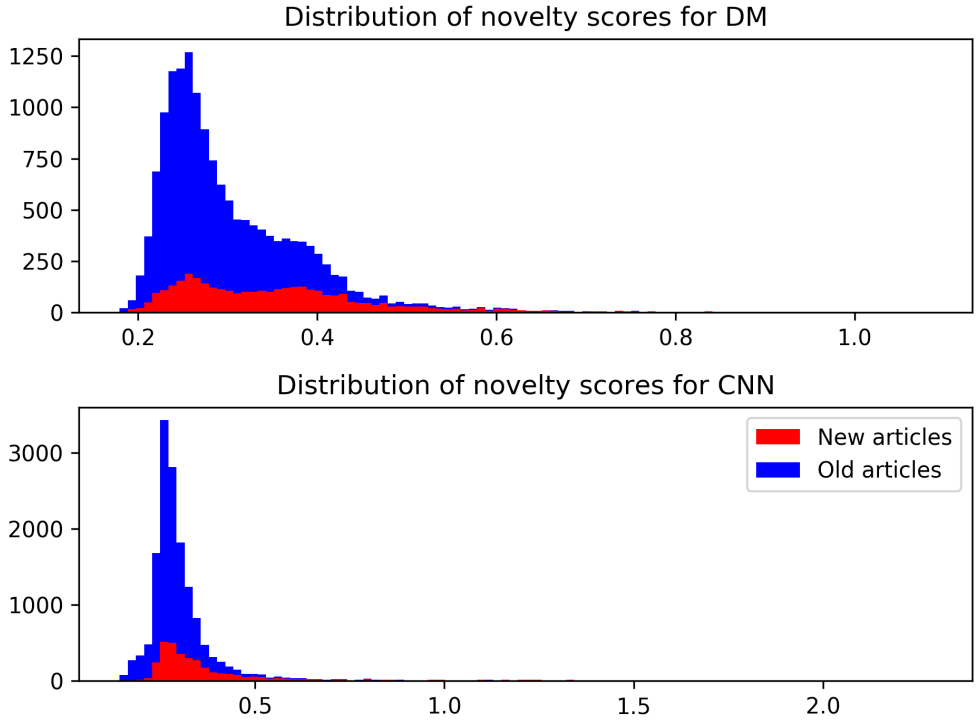


Figure 18: Distribution of topical novelty scores for old and new articles by outlet.

It can be seen that the distributions of the score are fairly similar for old and new articles. While there are a few outliers with fairly high scores for CNN which all represent a novel topic (the impeachment of Donald Trump), the bulk of new and old articles do not differ substantially in novelty as judged by the metric. For the Daily Mail the articles with the highest topical novelty score all seem to deal with red carpet appearances of celebrities at the Golden Globe and the Critics' Choice Award. We were originally interested in using the metric in a hard boundary to select truly novel articles, but these are fairly rare. There are a number of possible reasons why no clear distinction is possible. It might be that true topical novelty is just rare, with most articles covering topics that are broadly familiar (e.g. coverage of crimes, sports or regular events such as award ceremonies). Alternatively, our method is not sensitive enough to reliably identify these topics. We performed some manual inspection of generated topics and found that

while they were generally coherent, they were also quite broad, e.g. for CNN there was one topic that encompassed a wide variety of articles related to air travel, including various accidents. It seems conceivable that more hyperparameter tuning might offer more specific topics, for example by including larger values for the number of topics. Additionally, only selecting the model with the highest coherence might not be enough for current purposes, it might be necessary to also evaluate the distinctiveness of topics between new and old articles and select the model that offers the best trade-off between this metric and more general coherence. Another method for selecting novel topics might be to rely on the fact that certain news sites tag articles by topics (unfortunately, this is not the case for either DM and CNN). Articles on the site of British newspaper *The Guardian* are usually tagged with a variety of tags ranging from specific to more general. In this manner, an article on the Coronavirus outbreak will be tagged both with a tag named *Coronavirus outbreak* and with more general tags such as *Infectious diseases* and *Medical research*. By inspecting the temporal distribution of tags, one should be able to identify novel topics. Even if one does not want to use articles from another outlet, it might be possible to leverage it as a source of information, e.g. by training a classifier to distinguish between labelled topics on its articles and then using the classifier to identify articles on the novel topics from the original outlets. For current purposes, we decided to use the topical novelty score as a factor in the analysis of error rates, though we are aware of its limitations. It might still reflected graded differences in the novelty between different articles and an analysis could thus already offer preliminary insights into the effects of topical novelty. More information about the exact procedure used and intermediate topic modelling results can be found in the appendix.

7.1.4 Annotation

To match the original 200 test set articles we annotated 200 unseen articles from January 2020. We matched the distribution of articles between the different news outlets, selecting 10.5 percent of articles from CNN and the remaining 89.5 percent from DM. The selected articles were then summarized using the four systems presented above and annotated by means of the existing tool. Manual inspection revealed that articles seemed to still conform largely to the same writing style as articles from the test set. Some articles were topically novel, while others dealt with familiar tropes, especially for the Daily Mail share of the data. There, a large number of articles dealt with similar topics, often describing the appearance and clothing of various celebrities.

7.2 Results

The most immediate change we noticed was the increased tendency for summaries to contain repeated sentences. This was in fact so prevalent that we decided to annotate all sentences that were a proper subset of another sentence in the summary as repetitions. The analysis of the data revealed that See and to a much less extent Chen suffered from this problem: 50 % of summaries by See contained at least one repeated sentence, with some of them consisting entirely of the same sentence repeated a number of times. For Chen, this number was 5.5 %. We could not establish the cause of this behaviour, it could either be a genuine failure of the system to deal with novel articles or just a bug in the implementation that causes it to fail for example when encountering certain characters that happened to be absent for the test set articles due to different pre-processing. We decided to remove these sentences, only the first copy was retained. The effect this had on the average number of sentences per summary is reflected in Table 8. We can see that See generates a similar number of sentences as on the original test set, but as many of those are duplicates and have to be removed, the actual average is substantially lower. All other systems generate roughly the same number of sentences on the test data and

the unseen data. We can also see some differences in sentence and summary error rates when comparing to the test data. Sentence error rates seem to vary slightly, but there is not the expected general upward trend, with error rates lower than on the test set for Chen and PreSumm and somewhat higher for See and LM. The same pattern can be seen for summary error rates. Figure 19 contains a more detailed breakdown of error rates by meaning dimension. Most of the fluctuation is too small to be meaningful, such that the observed differences cannot be said to be due to differences between older and newer articles. The only large difference is observed for the rate of misleading sentences for PreSumm, with the rate decreasing from about 8 % to less than 3 %.

System	See	Chen	LM	PreSumm
Sentence error rate	0.20 (0.16)	0.20 (0.24)	0.24 (0.20)	0.10 (0.15)
# sent. / summary (raw)	2.76	5.01	2.00	3.19
# sent. / summary (no reps.)	2.16 (2.91)	4.95 (4.93)	2.00 (2.27)	3.19 (3.33)
Expected summary error rate	0.38 (0.39)	0.67 (0.74)	0.42 (0.39)	0.29 (0.41)
Observed summary error rate	0.33 (0.38)	0.62 (0.73)	0.39 (0.36)	0.27 (0.39)

Table 8: Binary error rates, sentence- and summary-level.

Systems are true to their original style when it comes to abstractiveness: PreSumm and See again engage largely in extraction, while Chen is prone to deletions and LM re-writing the most. Similarly, the distribution of error types in the mapping dimension is also remarkably stable. On a qualitative level, we also noticed that the style of the summaries was quite consistent, with many again suffering from unclear referential expressions and many of the error patterns referenced above still present. The fact that systems seem to deviate so little from the established formula can be thought to be related to the layout bias pointed out by [2]. If the most salient fact that systems pick up is just the position of the relevant sentences in the document, they will continue to apply the same edit strategy regardless of the actual content of the article. The most interesting observation relates to the intuition that pre-trained models with more "prior knowledge" might be more susceptible to this knowledge growing stale. We found a small number of incidents of this sort. Example 18 shows how LM introduces the name of President Obama in an article that only mentions President Trump, whose election happened briefly after the data collection period for the GPT model used. It seems conceivable that frequent mention of the name Obama in corcordances with "president" and "drone strike" biased the model so much that it failed to properly attend to the contents of the article. A similar pattern (though not related to knowledge growing stale) can be seen in Example 19, where the model inserts a more plausible term into a sentence while disregarding the content of the article. As dogs are much more frequently kept as pets, the language model seems to have overruled proper attention to the article. However, as mentioned, the general incidence rate of fabrications did not change, indicating that this problem is present for both old and new articles and recency alone is not enough to cause an increase. Topical novelty could potentially be a better predictor of worse summarization performance, however, as seen in Figure 20, we could not establish any meaningful differences between systems with regard to this factor.

7.3 Discussion

For **RQ5** and **RQ6**, we were interested in the robustness of systems. How do they perform when faced with more recent articles? We selected articles from January 2020, some of which dealt with novel topics. We found that performance of none of the systems degraded substantially, **with all of them producing errors at roughly the same rate** as before. The rate of misleading sentences decreased for one of the systems,

Article context: [...] Singer John Legend has blasted an airstrike ordered by President Donald Trump that killed one of Iran's top generals. [...]
Summary sentence: president donald trump has blasted an airstrike ordered by president obama that killed one of iran's top generals .

Example 18: LM introduces the name of a past president that is not mentioned in the article.

Article context: [...] Jaclyn Tarrant rescued the calf, named Ferdinand, in October - and he's been part of her family ever since.. The bull now goes everywhere with the family - to the dog park, beach, and even on a trip out of Sydney for Christmas lunch. [...]
Summary sentence: the pup now goes everywhere with the family - to the dog park , beach , and even on a trip out of sydney for christmas lunch .

Example 19: LM introduces a less implausible noun, ignoring the contents of the article.

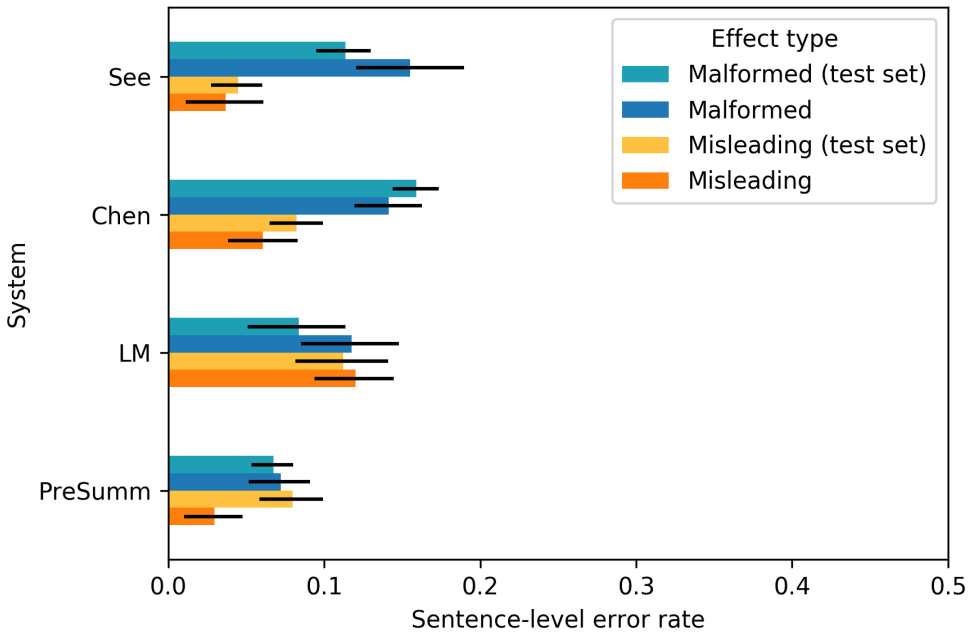


Figure 19: Comparison between error rates on old and new articles.

but no immediate explanation is available, as abstractiveness and edit types have not changed to a measurable degree. This pattern is generally present across systems, indicating that mere recency is not enough to cause substantial changes in the way systems summarize. We also investigated the effects of topical novelty, again finding performance to be similar no matter whether the article was categorized as more novel. Based on currently available evidence, we can then state that **all systems are robust** in the sense that their performance does not deteriorate. There is some preliminary evidence that at least one of the system suffers from the problem of relying on stale knowledge

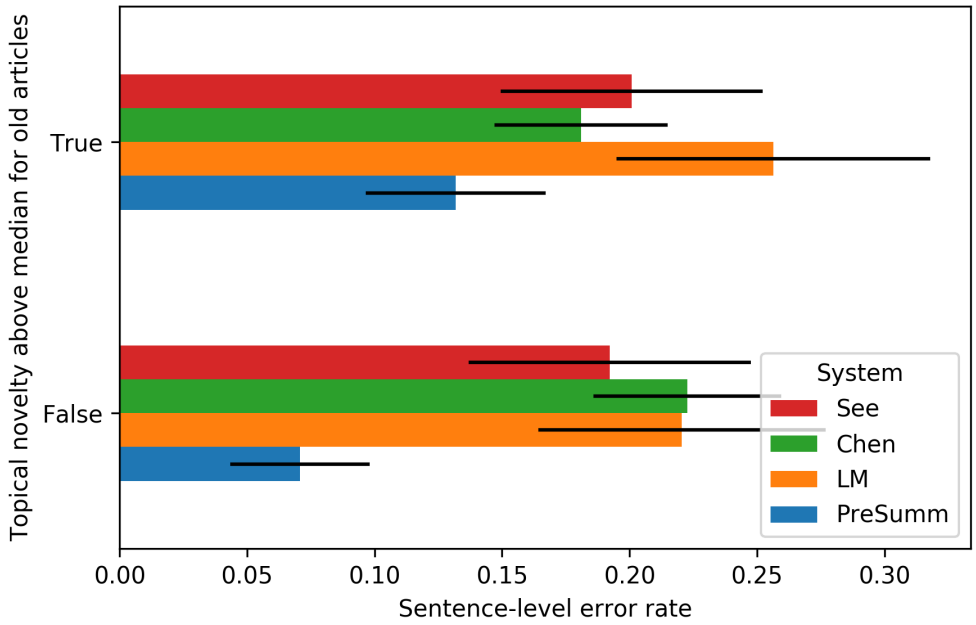


Figure 20: Effect of topical novelty on error rates. 95 % CI.

embedded in its language model, raising implications for future research. However, this part of the current study has a fair number of limitations. Our operationalization of topical novelty was largely ad-hoc and should be replaced with a more principled metric. Similarly, future analyses should try to disentangle the differences between topical novelty and stylistic change better. One could for example designate novel topics and then pick articles on these topics both from the original outlets and from other outlets with different writing styles. Looking at the effect of stale knowledge, sample articles could be engineered for a more detailed study, e.g. by re-writing an old article about a former US president slightly and including the name of the current president.

More generally, the error rates we observed on novel articles are still large, raising more general implications about the field of summarization. These will be discussed in the final section of the thesis.

8 Conclusion and Outlook

Are current summarization systems suitable for large-scale deployment at a major news organization? This was the question that originally motivated this thesis. While our main focus was on avoiding factual errors, we first had to get a better understanding of the nature of errors per se. Towards this aim, we established a novel typology of errors to be used for sentence-level annotation and validated it by means of the inter-annotator agreement. We then proceeded to annotate summaries for a subset of the test set of the original data. We found that systems make a substantial number of errors, with none of them being able to entirely avoid at least occasionally producing misleading sentences. Pre-training is not enough to avoid this issue. While summaries generated by pre-trained systems can be found to generally be less likely to be malformed, they are often misleading, with fabrication of words being the culprit frequently. We identified

an important trade-off between abstractiveness and error rates, with all current systems doing worse when they try to be abstractive.

Looking into the robustness of systems, we found that they seem to be robust to more recent articles in the sense that they do not get worse. In fact, their editing style is remarkably stable, indicating that the exact content of the article might not even be that important to how systems summarize. If all that matters is the layout of the article, we would expect systems to keep on working until changes to this aspect occur. More experiments should look into this aspect, e.g. by manipulating only certain parts of an article and inspecting the implications on the resulting summaries. Another aspect we investigated was the effect of topical novelty. We presented an initial operationalization of novelty, but found it to be lacking in distinctiveness. More research should focus on identifying articles from novel topics and how current summarization systems perform for them.

These findings demonstrate that simply deploying a system into the wild is currently not an option for an outlet conscious of its reputation. Instead, it might be necessary to introduce human oversight, but even then, some errors might be difficult to spot if the editor does not employ close attention. If such resources are not available, but summarization is desired, it might be preferable to employ a purely extractive system: Current abstractive systems are not very abstractive anyway, but they are more likely to make errors.

There is ample potential for future research. As we established above, there has been some criticism of the CNN/DM dataset, to which this thesis has added by pointing out the troublesome implications of the fact that all headlines are omitted from the dataset, even though the reference summaries are often written in such a way as to rely on their presence. How are automatic summaries to learn to write coherent summaries that are free of dangling anaphora and unclear expressions if their training input does not conform to this requirement? As all the original web scraping code is available and the original researchers archived all the news articles by means of the Web Archive, there is nothing that would bar us from simply re-running the original experiments, including the headline along the reference summary and additionally performing some additional data cleaning such as deleting duplicate articles or removing image captions, hoping to avoid repetitions and improving generalization to articles that lack image captions in their body. These relatively minor improvements can be predicted to have a beneficial effect on system performance and might help to substantially reduce the number of malformed sentences and to possibly improve factual correctness. The community has invested a fair amount of effort into engineering ever more complex neural pipelines, but so far neglected to pay more attention to data quality. In this area, a little could already help to go a long way towards better summarization.

Another promising research direction is the automatic detection of summary errors. This idea builds on the intuition that errors are governed by clear linguistic patterns. Though observed agreement between annotators was not high, some error types were less ambiguous and should be easy to stably identify. It seems feasible to train a system for detecting these errors on a feature representation that captures linguistic patterns related to the errors. For example, improper deletions that disrespect the boundaries of the phrase structure of a sentence could be found by means of a parsing step. In a similar way, we could flag up the insertion of completely novel words not semantically related to the article content (such as fabricated names). Other resources could also be leveraged where available, e.g. information-extraction based methods that generate

semantic triplets. If trained on a sufficient number of annotated summary sentences, the system might be able to accurately estimate the chance of a given summary or sentence containing an error. It could then be used as an automatic warning system, alerting human editors of possible problems with summaries to reduce the chance of them getting published. Alternatively, one might also use it as an evaluation component in a summarization ensemble. Given the low correlation between systems in terms of error prevalence we established above, it is likely that for any given article, there is one system that generates an error-free summary. This chance increases as more sufficiently diverse systems are included. In summarization ensembles, multiple systems or additionally multiple versions of the same system with varying training inputs or readout parameters would be used to generate summaries and each of the resulting summaries would be classified as to whether it contains an error. A selection component would then select a summary based on the estimate of the classifier and possibly other factors such as abstractiveness. This component would also give the user more control, allowing them to explicitly specify a trade-off between the risk of factual errors and the abstractiveness of summaries, e.g. by formulating rules of the sort *"pick the most abstractive summary whose estimated chance of containing an error is below 5 %"*.

Finally, a detailed experimental study should investigate how subtle manipulations of the input article affect summarization systems. This thesis has laid the groundwork for these follow-up studies by devising the typology, the annotation tool as well as pipelines for loading articles and performing inference with various systems.

References

[1] Maya Sappelli, Dung Manh Chu, Bahadir Cambel, David Graus, and Philippe Bressers, "SMART Journalism: Personalizing, Summarizing, and Recommending Financial Economic News," in *The Algorithmic Personalization and News (APEN18) Workshop at ICWSM*, 2018, vol. 18.

[2] Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher, "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)*, Hong Kong, China, Nov. 2019, pp. 540–551, Association for Computational Linguistics.

[3] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych, "Ranking generated summaries by correctness: An interesting but challenging application for natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019, pp. 2214–2220, Association for Computational Linguistics.

[4] Alec Radford, "Improving Language Understanding by Generative Pre-Training," in *OpenAI research publication*, 2018.

[5] Günes Erkan and Dragomir R Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[6] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou, "SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017, pp. 3075–3081.

[7] Elena Lloret, Laura Plaza, and Ahmet Aker, "The challenging task of summary evaluation: an overview," *Language Resources and Evaluation*, vol. 52, no. 1, pp. 101–148, Mar. 2018.

[8] Liana Ermakova, Jean Valère Cossu, and Josiane Mothe, "A survey on evaluation of summarization methods," *Information Processing & Management*, vol. 56, no. 5, pp. 1794–1814, Sept. 2019.

[9] Feifan Liu and Yang Liu, "Correlation between rouge and human evaluation of extractive meeting summaries," in *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*. Association for Computational Linguistics, 2008, pp. 201–204.

[10] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.

[11] Ibrahim F. Moawad and Mostafa Aref, "Semantic graph reduction approach for abstractive Text Summarization," in *2012 Seventh International Conference on Computer Engineering & Systems (ICCES)*. 2012, pp. 132–138, IEEE.

[12] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li, "Faithful to the original: Fact aware neural abstractive summarization," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018, pp. 4784–4791.

[13] Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser, "Sample Efficient Text Summarization Using a Single Pre-Trained Transformer," *arXiv:1905.08836 [cs]*, May 2019, arXiv: 1905.08836.

[14] Luke de Oliveira and Alfredo Láinez Rodrigo, "Repurposing Decoder-Transformer Language Models for Abstractive Summarization," *arXiv:1909.00325 [cs]*, Sept. 2019, arXiv: 1909.00325.

[15] Yang Liu and Mirella Lapata, "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)*, 2019, pp. 3721–3731.

[16] Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy, "Neural Abstractive Text Summarization with Sequence-to-Sequence Models," *arXiv:1812.02303 [cs, stat]*, Dec. 2018, arXiv: 1812.02303.

[17] Yue Dong, "A Survey on Neural Network-Based Summarization Methods," *arXiv:1804.04589 [cs]*, Mar. 2018, arXiv: 1804.04589.

[18] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang, "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond," in *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016, pp. 280–290.

[19] Abigail See, Peter J. Liu, and Christopher D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1073–1083.

[20] Yen-Chun Chen and Mohit Bansal, "Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 675–686.

[21] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, pp. 9, 2019.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 5998–6008. Curran Associates, Inc., 2017.

[23] Andrew Pau Hoang, Antoine Bosselut, Asli Çelikyilmaz, and Yejin Choi, "Efficient adaptation of pretrained transformers for abstractive summarization," *ArXiv*, vol. abs/1906.00138, 2019.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[25] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov, "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 5797–5808, Association for Computational Linguistics.

[26] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[27] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom, "Teaching machines to read and comprehend," in *Advances in neural information processing systems*, 2015, pp. 1693–1701.

[28] Donna Spencer, *Card sorting: Designing usable categories*, Rosenfeld Media, 2009.

[29] Paul Kroeger, *Analyzing meaning: An introduction to semantics and pragmatics*, 2018.

[30] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[31] Michael Röder, Andreas Both, and Alexander Hinneburg, "Exploring the Space of Topic Coherence Measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai, China, Feb. 2015, WSDM '15, pp. 399–408, Association for Computing Machinery.

9 Appendix

9.1 Source code

The source code for this thesis is available on GitHub. The SummaryInspector interface and source code that was used for the analysis of the IAA and the final results can be found here: `https://github.com/CreateRandom/summary-inspector-gui`. The code that was used for ingesting generated summaries is here: `https://github.com/CreateRandom/summary-inspector-ingestion`.

9.2 Annotation specification

The following is the specification as used during the annotation. Note that the terminology differs. The meaning dimension was originally referred to as *error effect*, and the mapping dimension was referred to as *error source*.

Annotation specification

Annotation procedure

The aim of the annotation is to understand the nature of linguistic errors present in short summaries of news articles generated by various automatic systems. Annotators are presented with articles and associated summaries. Annotation proceeds as follows:

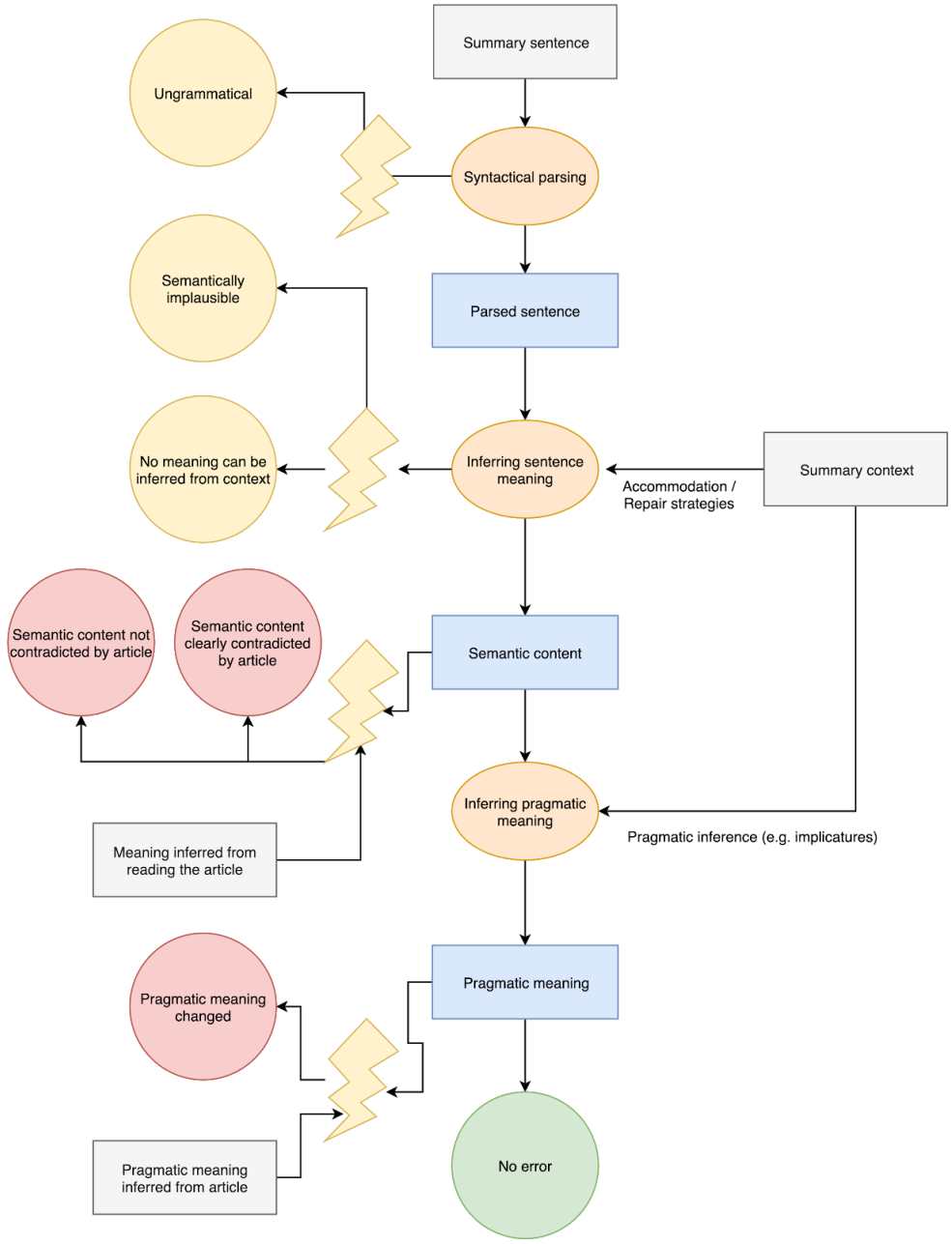
1. Only read the summaries in isolation. Check for ungrammatical sentences, semantically implausible sentences and sentences for which no meaning can be inferred and indicate them.[1]
2. Read the article.[2] If there is a mismatch between the content of the article and the summary, indicate this using the checkbox at the bottom. For sentences that pass the first step, check whether they reflect a change of semantic content given the article or a change in pragmatic meaning.
3. Add comments where appropriate.
4. For every sentence that contains an annotation on the effect level, compare it to the article to be able to annotate the error source.

[1] In the annotation software, sentences that are copied verbatim from the article are distinguished from sentences that have at least one edit. The latter are indicated by a star. This can help you at this stage, as sentences that are copied verbatim are unlikely to be ungrammatical or non-sense, though they still might not make sense in isolation.

[2] If the article as depicted in the article window does not make sense or seems unclear, you can use the 'Google' button. It will search for the first couple of words of the article on the internet and open the best match in your browser.

Error effect

We distinguish a total of six error effects. They are grounded in linguistic theory, c.f. the following flow chart.



For all errors, we assume the reader has only access to the summary text, i.e. they cannot refer back to the original article. We assume readers will make an effort to linguistically accomodate.

The following sections describe the error types in more detail and provide examples and edge cases.

Ungrammatical

Definition: A sentence that violates the rules of English syntax.

Edge cases

- Telegraphic sentences?
    - “Queen present at race track” etc. → fine in a human summary → **don’t label**
- Sentences that would only be okay as image captions?
    - “Donald Allen, pointing a gun at an officer”
    - In an image caption, the sentence would implicitly be read as “This is Donald Allen... ” → however, here, this is not possible → **label**

Examples

- united's manager felt following the worst day he has experienced in the premier league.
    - Using a transitive verb as if it was intransitive
- comres survey of 4,000 undecided voters found a slim majority have been impressed with the prime minister than that of ed miliband during the election campaign
    - Missing *more*, *that of* cannot stand on its own
- analysts employed at us intelligence agencies like the cia.
    - Passive phrase cannot stand on its own

Semantically implausible

Definition: A sentence that is grammatically correct, but which has no plausible interpretation based on world knowledge.

Examples

- the anniversary of ve day on may 8 , 1945 , will be a woman.
    - Grammatically correct, but meaningless

No meaning can be inferred

Definition: A sentence that is grammatically correct, but to which no meaning can be assigned, even after accommodating.

- *Unresolvable anaphoric expressions would end up here*

Edge cases

- Dangling anaphora that can be resolved by *following* context
    - “She was blindsided by his decision. Hamilton filed for divorce.”
    - The anaphora can get resolved correctly here. It's a bit unusual / more literary, but probably does not qualify as a violation → **don’t label**

Examples

- there has n't been any evidence she was involved with the scheme , and her supporters say the position is merely a figurehead .
    - No prior mention of any person the expression could refer to

Semantic content changed

Semantic content changed, not entailed

Definition: When read in the context of the surrounding summary, the semantic content assigned to a sentence is not entailed by the original article.

Examples

- X is nine-months pregnant → X is nine-months pregnant with her first child
    - There is no information about how many children X has in the article. It's possible this is true, but we cannot say for certain.

Semantic content changed, contradicted

Definition: When read in the context of the surrounding summary, the semantic content assigned to a sentence is in contradiction to what is said in the article.

Examples

- australians send 30 per cent more lgbt-related emoji than the average
    - The original article states that this applies to the US, not Australia
- christian trousedale became an internet hit after a picture of him carrying a 95-year-old pensioner.
    - The article states he carried the pensioner's shopping

Pragmatic meaning changed

Definition: When read in the context of the surrounding summary, the sentence gains a pragmatic meaning that was not present in the original article. Alternatively, a pragmatic meaning present in the original article is not faithfully retained in the summary.

Edge cases

- Person: 'Some statement' → Some statement
    - Summary sentence missing the attribution? Should this be annotated?

Examples

- He could even miss → He could miss

    - Here, the word even gets deleted - the journalist used this to hedge, implying this to be unlikely, the deletion makes it seem more likely
- Something happened, she alleges → something happened
    - The journalist uses the phrase to distance himself, implying he does not endorse this as true → the deletion turns it into a fact

Error source

By comparing the summary sentence to sentences in the article, we can judge how an erroneous sentence came about. We distinguish between four causes.

Lack of re-writing

Definition: Removing an article sentence from its surrounding context and then failing to adequately rewrite it in order to compensate for the missing context.

Examples

- Dilma Rousseff is in trouble. She has lost the support of Brazilians. → She has lost the support of Brazilians.
    - The second sentence contains the anaphoric expression ‘she’. By copying the sentence and not re-writing it, the system creates a sentence that the reader cannot understand.

Omission

Definition: Omitting words or phrases from an article sentence.

Examples

Article context: [...] she suffered from the rare disease progeria which ages the body at eight times the normal rate. [...]
Summary sentence: she suffered from rare disease progeria which ages the body at eight times.

- The omission of the phrase ‘the normal rate’ causes the sentence to become ungrammatical.

Article context: [...] sandra bullock was horrified when she discovered that joshua corbett had broken into her house last summer. [...]
Summary sentence: sandra bullock had broken into her house last summer.

- The omission of large parts of the main clause causes a change in meaning.

Wrong combination

Definition: Combining words or phrases from article sentences improperly. More than two sentences can be involved.

Edge cases

- Sentences parts separated by dashes
    - Should not be considered separate sentences
    - Thus, if they get combined and words in between are deleted, it's an omission rather than WC
- Sentence parts whose order is changed compared to the summary
    - It's still a wrong combination.

Examples

Article context: [...] if the player misses the girl, she starts to lose weight until she eventually dies.. [...] gamers have to throw food at the girl who appears in one of nine holes before she disappears again. [...]
Summary sentence: gamers have to throw food at the girl who appears in one of nine holes before she dies

- Information from two different sentences is combined improperly

Fabrication

Definition: Inserting new words or phrases not present in the original article.

Examples

Article context: [...] yet for decades, many have stood by the belief that such programs, known as syringe exchange or syringe services programs, promote and encourage drug use..
Summary sentence: john avlon: syringe exchange programs have been used by many to promote drug use.

- The phrase 'john avlon' appears nowhere in the article

Error in article

Definition: The error was present in the article as well.

9.3 Inter-annotator agreement

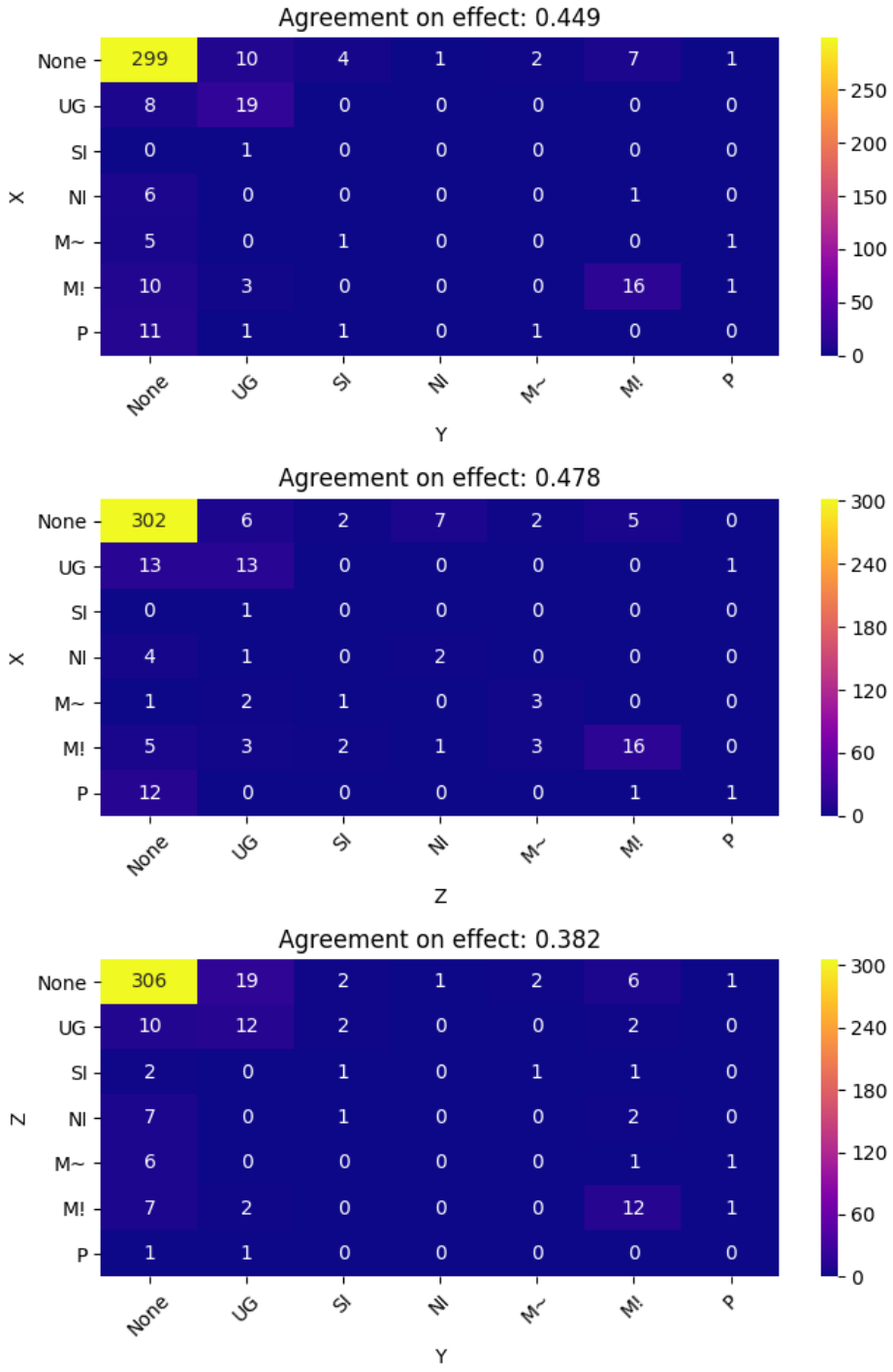


Figure 21: Rating comparison for three raters, meaning dimension.

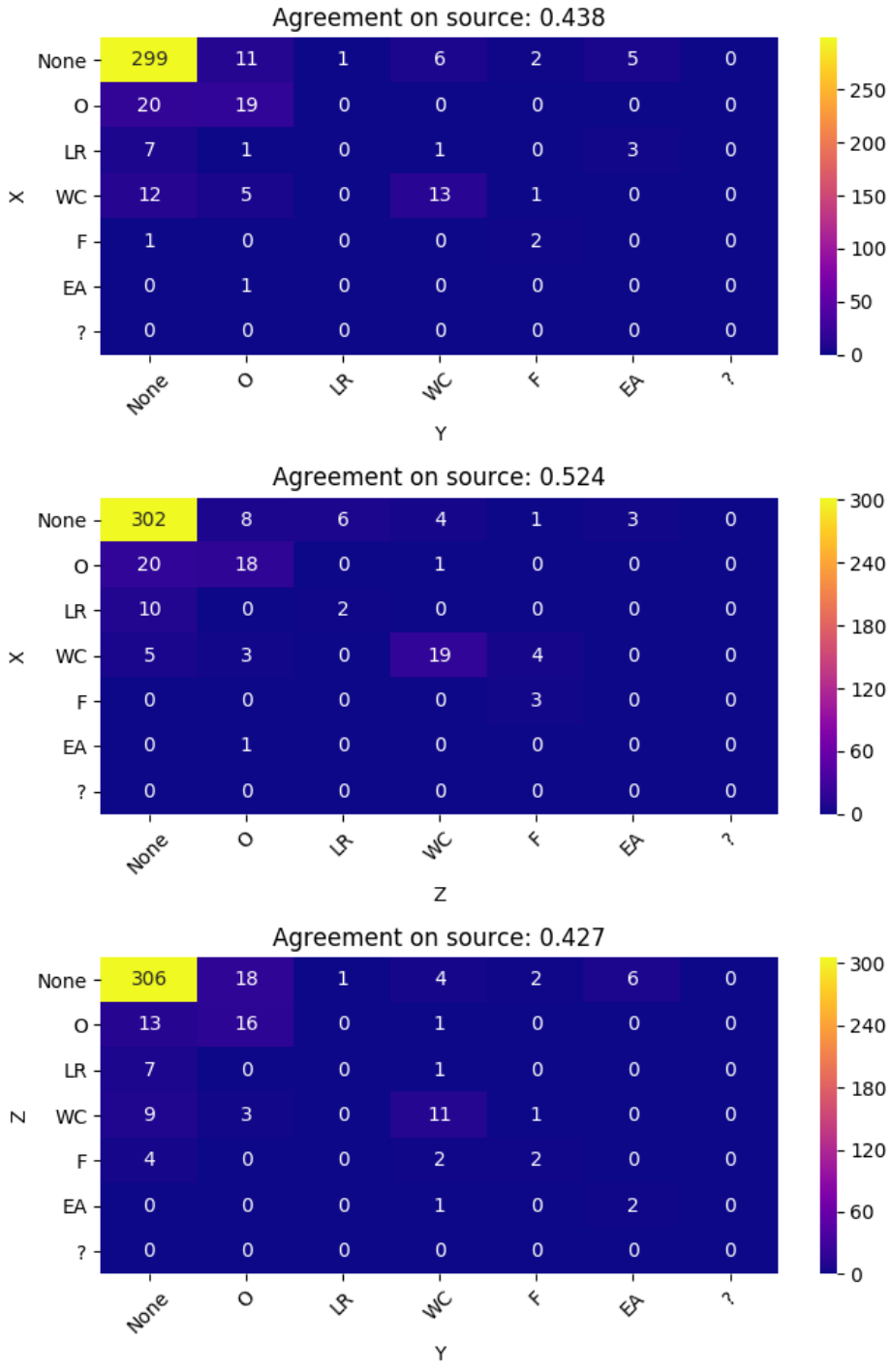


Figure 22: Rating comparison for three raters, mapping dimension.

9.4 Modelling topical novelty

This section contains some information on how we attempted to model topical novelty and some intermediate results.

9.4.1 Methods

Individual articles were represented by unigram and bigram counts. Rather than extracting all bigrams, we relied on the Phrases helper in Gensim, using its default parameters as of version 3.8.1. The *scoring* parameter was set to 100 to reduce occurrences of spurious phrases. We performed stopword removal, using the NLTK stopword list. Words were lemmatized and only certain lemmata were retained, namely nouns, adjectives and adverbs. We used the LDA implementation from Gensim, setting $alpha = auto$, $eta = auto$, $passes = 10$, $iterations = 150$. We performed grid search over the number of topics, using $[25, 30, 40, 50, 75, 100]$ and the chunksize, using $[500, 2000, 4000]$. For each model, two random seeds were tried. For both datasets, chunksize of 500 performed best. For DM, 30 topics showed the best coherence, for CNN 25 topics did.

9.4.2 Results

Tables 9 and 10 show the topics for the best models. The ratio is between the mass of the topic for all new articles and the mass for all old articles. In both cases, the expected ratio if no difference exists for a topic is simply the ratio between old and new articles, this is $3600/12000 = 0.3$. The total mass shows how prevalent the topic is. Finally, the associated words are the words that feature most prominently in the distribution over words for each topic, giving an indication what the topic is about. We can see that only a small number of topics differ markedly in distribution between new and old articles. There are also a number of topics with fairly general words that cover a large proportion of documents, e.g. Topic 7 and 23 for CNN and Topic 14 and 17 for DM. These results indicate that a different methodology might need to be employed to better capture finer topical variety.

Topic id	Ratio	Total mass	Associated words
10	3.34	185.28	.059 romney,.056 trial,.034 witness,.028 ipad,.023 defense,.017 trump,.017 testimony,.016 question,.014 evidence,.013 president
24	.61	474.43	.036 election,.030 political,.028 campaign,.025 vote,.020 president,.019 state,.019 candidate,.017 party,.016 republican,.016 democratic
13	.42	708.04	.028 year,.016 percent,.014 money,.014 oil,.013 government,.013 economy,.013 company,.011 job,.011 country,.011 business
9	.42	135.97	.217 official,.068 security,.057 source,.047 intelligence,.035 information,.031 senior,.021 investigation,.020 strike,.020 operation,.020 threat
8	.41	424.63	.047 company,.023 new,.019 user,.017 technology,.016 product,.014 site,.014 online,.014 internet,.013 apple,.013 computer
4	.38	91.7	.199 school,.192 student,.067 teacher,.063 pilot,.046 education,.043 crash,.029 helicopter,.026 high,.019 campus,.014 classroom
6	.36	2294.91	.012 also,.012 public,.012 statement,.010 law,.009 decision,.008 government,.008 issue,.008 member,.008 right,.008 report
18	.34	97.53	.113 black,.047 white,.033 spanish,.032 defeat,.032 ball,.030 set,.022 race,.022 color,.019 racial,.017 convention
1	.33	348.48	.029 health,.027 doctor,.025 patient,.024 drug,.022 study,.020 treatment,.018 medical,.016 hospital,.014 disease,.014 test
5	.32	30.93	.112 woman,.052 film,.045 movie,.035 man,.022 actor,.021 boy,.020 female,.020 show,.018 character,.018 marriage
14	.32	111.27	.055 hotel,.035 store,.028 agent,.027 room,.027 sale,.027 ticket,.024 theater,.024 travel,.021 fee,.021 mexican
2	.31	273.72	.045 flight,.041 plane,.038 food,.035 passenger,.031 ship,.023 airport,.021 animal,.020 aircraft,.018 airline,.015 crew
19	.3	224.9	.076 water,.034 storm,.020 foot,.018 space,.015 scientist,.015 inch,.015 temperature,.014 power,.014 weather,.013 tree
12	.29	67.27	.134 music,.109 tour,.060 singer,.048 band,.036 bus,.025 pop,.025 rock,.023 celebrity,.022 record,.020 stone
11	.28	614.53	.044 case,.043 court,.029 charge,.028 year,.020 crime,.020 attorney,.020 prosecutor,.017 lawyer,.017 man,.016 authority
15	.26	2536.97	.023 time,.016 good,.015 year,.014 thing,.014 people,.013 way,.012 even,.012 really,.011 first,.011 much
7	.25	389.11	.065 country,.040 government,.027 leader,.022 nation,.022 international,.018 world,.018 power,.016 political,.016 rebel,.015 russian
3	.24	159.04	.112 team,.074 player,.063 football,.059 video,.048 fan,.030 coach,.023 game,.021 athlete,.018 stadium,.017 championship
23	.24	2852.87	.029 people,.016 day,.012 also,.011 city,.010 week,.009 many,.009 area,.009 year,.008 home,.008 hour
0	.23	253.5	.239 police,.092 officer,.060 car,.054 authority,.039 vehicle,.027 report,.026 driver,.018 statement,.018 truck,.016 body
21	.23	743.86	.071 family,.066 child,.030 old,.028 year,.020 life,.020 home,.020 mother,.018 young,.017 parent,.017 death
20	.22	36.99	.120 israeli,.078 egyptian,.062 prison,.059 prisoner,.048 tennis,.046 palestinian,.046 refugee,.033 settlement,.029 tunnel,.027 inmate
16	.19	462.78	.060 attack,.056 military,.042 force,.023 group,.023 troop,.023 government,.022 soldier,.019 security,.018 terrorist,.016 civilian
22	.15	245.52	.052 government,.049 violence,.044 people,.038 protest,.037 gun,.028 protester,.024 group,.019 street,.018 activist,.016 opposition
17	.12	775.45	.035 game,.035 year,.026 season,.025 first,.025 second,.024 last,.019 time,.019 match,.018 team,.017 final

Table 9: Result of topic modelling for the CNN dataset.

Topic id	Ratio	Total mass	Associated words
0	1.26	122	.041 good,.036 award,.029 winner,.028 actress,.024 ceremony,.023 actor,.018 prize,.017 flower,.016 exhibition,.014 category
4	.86	485.32	.034 black,.025 white,.020 hair,.017 model,.014 style,.012 dress,.012 fashion,.011 collection,.009 boot,.009 designer
5	.76	83.74	.077 music,.059 artist,.051 singer,.047 art,.033 band,.028 painting,.025 song,.018 performance,.016 good,.015 rapper
29	.62	39.16	.055 video,.039 show,.038 film,.023 tv,.021 twitter,.021 star,.016 fan,.014 picture,.014 actor,.013 photo
9	.56	327.85	.070 couple,.057 wife,.029 wedding,.029 husband,.027 friend,.026 relationship,.024 marriage,.024 former,.024 royal,.019 together
11	.54	579.76	.047 player,.037 team,.037 season,.036 club,.026 football,.019 last,.018 manager,.018 former,.018 year,.016 game
15	.42	43.93	.051 game,.026 goal,.024 first,.023 second,.020 minute,.019 side,.019 match,.019 ball,.018 race,.016 final
25	.36	176.84	.082 food,.070 dog,.034 restaurant,.021 weight,.019 meal,.017 fat,.017 owner,.016 fish,.015 healthy,.012 fruit
17	.33	3287.84	.021 time,.016 first,.014 year,.012 people,.012 even,.012 also,.010 way,.009 day,.009 good,.008 much
2	.31	232.31	.063 patient,.035 health,.034 people,.033 cancer,.033 hospital,.022 care,.021 risk,.020 disease,.020 case,.018 medical
16	.28	823.04	.022 fire,.021 water,.012 day,.012 area,.011 people,.008 high,.008 mile,.008 today,.007 last,.007 hour
3	.28	127.7	.368 woman,.118 man,.074 sex,.036 female,.027 letter,.017 sexual,.016 campaign,.013 young,.011 presidential,.011 partner
23	.28	57.55	.069 energy,.068 metal,.050 gas,.035 plant,.029 nuclear,.029 bike,.024 toilet,.023 power,.020 material,.019 cockroach
28	.28	376.96	.033 home,.022 property,.021 animal,.019 local,.018 site,.017 house,.017 tree,.016 resident,.016 area,.015 town
7	.27	21.36	.044 store,.040 image,.030 sale,.023 item,.021 shop,.017 balloon,.014 chain,.014 camera,.013 product,.012 shopping
14	.25	1797.3	.051 year,.039 old,.038 family,.026 home,.021 mother,.018 child,.017 last,.017 life,.016 day,.015 time
8	.24	321.73	.027 study,.019 scientist,.017 population,.017 research,.014 researcher,.013 human,.012 brain,.011 group,.011 also,.011 age
27	.24	137.74	.063 flight,.058 plane,.053 passenger,.052 pilot,.030 crew,.029 air,.028 aircraft,.026 ship,.025 airport,.020 boat
20	.24	987.94	.055 year,.024 last,.019 high,.018 number,.011 people,.010 figure,.010 new,.010 country,.010 job,.009 rate
21	.22	36.91	.031 official,.030 military,.027 attack,.025 security,.017 force,.013 troop,.013 american,.013 aid,.012 russian,.012 former
6	.22	512.12	.043 company,.019 service,.015 customer,.015 user,.015 firm,.014 also,.014 phone,.013 device,.012 online,.011 system
22	.21	258.34	.052 doctor,.040 hospital,.030 condition,.025 surgery,.024 treatment,.019 operation,.019 leg,.018 heart,.016 pain,.016 blood
24	.2	118.92	.059 train,.038 hotel,.031 station,.028 space,.027 speed,.025 horse,.022 track,.019 traveller,.018 mission,.013 construction
1	.2	228.29	.108 death,.035 accident,.030 crash,.025 tragedy,.024 friend,.024 hospital,.023 tragic,.022 family,.017 alcohol,.014 ambulance
12	.19	619.05	.043 government,.020 country,.019 leader,.016 people,.014 today,.014 soldier,.013 party,.012 former,.012 last,.012 member
26	.19	746.38	.128 police,.047 officer,.042 car,.037 man,.021 victim,.018 attack,.013 people,.013 driver,.013 vehicle,.013 incident
10	.19	216.2	.176 child,.107 school,.066 student,.053 parent,.038 boy,.035 young,.032 girl,.025 teacher,.017 education,.016 adult
18	.18	271.44	.082 money,.035 price,.030 pay,.027 cash,.023 bank,.022 financial,.020 cost,.019 bill,.018 payment,.017 business
19	.16	299.16	.056 murder,.034 drug,.032 death,.032 prosecutor,.027 allegedly,.019 attorney,.016 charge,.016 gun,.015 count,.013 authority
13	.13	487.06	.085 court,.044 case,.030 trial,.024 victim,.023 lawyer,.022 evidence,.021 year,.019 prison,.017 judge,.017 charge

Table 10: Result of topic modelling for the DM dataset.