Truth or Error? Towards systematic analysis of factual errors in abstractive summaries

Klaus-Michael Lux
Radboud University
info@klauslux.de

Maya Sappelli
HAN University of Applied Sciences
maya.sappelli@han.nl

Martha Larson
Radboud University
m.larson@cs.ru.nl

Abstract

This paper presents a typology of errors produced by automatic summarization systems. The typology was created by manually analyzing the output of four recent neural summarization systems. Our work is motivated by the growing awareness of the need for better summary evaluation methods that go beyond conventional overlap-based metrics. Our typology is structured into two dimensions. First, the Mapping Dimension describes surface-level errors and provides insight into word-sequence transformation issues. Second, the Meaning Dimension describes issues related to interpretation and provides insight into breakdowns in truth, i.e., factual faithfulness to the original text. Comparative analysis revealed that two neural summarization systems leveraging pre-trained models have an advantage in decreasing grammaticality errors, but not necessarily factual errors. We also discuss the importance of ensuring that summary length and abstractiveness do not interfere with evaluating summary quality.

1 Introduction

We are currently witnessing a sharp increase of research interest in neural abstractive text summarization. However, we have also seen growing concern that truth, as represented in the original document, becomes lost or twisted during the summarization process. The issue was raised recently by Kryscinski et al. (2019), who point out that widely used automatic metrics, which rely mostly on word overlap, fail to reflect factual faithfulness of a summary to the original text. Until now, work on summarization has not provided systematic analysis of factual faithfulness. Instead, the trend has been for papers to provide a few examples or general descriptions of frequent errors. An example is Falke et al. (2019), 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."*, but stop short of providing a more rigorous analysis. Recent work that breaks the trend is Durmus et al. (2020), who propose an evaluation framework for faithfulness in abstractive summarization. The summaries used to develop the framework are annotated with different types of faithfulness errors. However, the annotation scheme does not incorporate linguistic concepts, e.g., does not differentiate between semantic and pragmatic faithfulness.

The aim of our research is to go beyond existing characterizations and provide a comprehensive typology that can be used to understand errors that neural abstractive summarization systems produce, and how they affect the factual faithfulness of summaries. The contribution of this paper is an error typology that was created by analyzing the output of four abstractive summarization systems. The systems vary in their use of pre-training, their model architecture and in the integration of extractive tasks during training. We carry out a comparative analysis that demonstrates the ability of the typology to uncover interesting differences between systems that are not revealed by conventional overlap-based metrics in current use. This paper represents the main results of Lux (2020), which contains additional examples and analysis. Further, annotations used for our analysis and more detailed statistics are publicly available[1] to support future research on faithfulness errors.

2 Related work

Over the years there have been several methods to evaluate summarization methods (Lloret et al., 2018; Ermakova et al., 2019), each with their own strengths and challenges. In this section, we first

[1] https://tinyurl.com/truth-error-2020.

Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP), pages 1–10,
November 20, 2020.

cover the ROUGE score, which is the main target of the criticism of overlap-based summarization metrics, such as from Kryscinski et al. (2019) mentioned in Section 1. We then provide a discussion on the relatively limited amount of work that has dealt with factual errors in summaries. Finally, we introduce the automatic summarization systems that we use in our study.

2.1 ROUGE

ROUGE is a set of metrics that measures textual overlap (Lin, 2004). The ROUGE score is almost exclusively used as the optimization and evaluation metric in neural summarization methods, even though it has been recognized to be difficult to interpret and does not correlate well with human judgement (van der Lee et al., 2019).

The major issue with the ROUGE score is its focus on textual overlap with a reference summary, which does not measure important aspects in summaries such as redundancy, relevance and informativeness (Peyrard, 2019a). Moreover, there is no clear optimal variant of ROUGE, and the exact choice can have a large impact on how a (neural) summarizer behaves when it is used as a training objective (Peyrard, 2019b). Sun et al. (2019) demonstrate another shortfall of ROUGE-based evaluation: Since the metric does not adjust for summary length, a comparison between systems can be misleading if one of them is inherently worse at the task, but better tuned to the summary length that increases ROUGE.

The shortcomings of ROUGE suggest that we should work towards metrics that are more focused on summary quality as perceived by readers. Unfortunately, quality is hard to measure, demonstated by an interactive summarization experiment by Gao et al. (2019), in which the authors show that users find it easier to give preference feedback on summaries. Simple preference ordering, however, does not give insight in the actual cause of preference. An important factor of perceived quality can be the errors being made by the summarizer. Grammatical errors can have an effect on the perceived quality, credibility and informativeness of news articles when there are many (Appelman and Schmierbach, 2018). Moreover with the rise in fake news and misinformation it seems important to have a better grip on factual errors that are a result of the summarization process.

2.2 Factual errors in summaries

Recent abstractive systems have a tendency to generate summaries that are factually incorrect, meaning that they fail to be factually faithful to the documents that they summarize. An analysis by Cao et al. (2018) of a neural summarization system finds that up to 30% of generated summaries contain "fabricated facts". Similarly, the authors of Falke et al. (2019) evaluate three different state-of-the-art systems and find that between 8 and 26% of the generated summaries contain at least one factual error, even though ROUGE scores indicate good performance.

Kryściński et al. (2019) propose a weakly supervised method for verifying factual consistency between document and summary by training a binary model that predicts whether or not a sentence is consistent. For this purpose they artificially generate a dataset with various types of errors, such as entity or number swapping, paraphrasing, pronoun swapping, sentence negation and noise injection. The authors claim the error patterns to be based on an error analysis of system output. However, it is not conclusively established that they constitute a good approximation of the actual errors that current summarization systems make.

Additionally, Goodrich et al. (2019) compare several models such as relation extraction, binary classification and end-to-end models (E2E) for estimating factual accuracy on a Wikipedia text summarization task. They show that their E2E model for factual correctness has the highest correlation with human judgements and suggest that the E2E models could benefit from a better labeling scheme.

In contrast, Lebanoff et al. (2019) are interested in what happens when summarization systems fuse sentences from the source. They automatically extract fused summary sentences generated by five different systems and conduct a manual annotation of faithfulness and grammaticality using crowd sourcing. Reference summaries are annotated as well. Generally, they find that fused sentences are often unfaithful to the source, especially when there is a marked imbalance in the contribution of multiple sentences. Surprisingly, the reference summaries achieve lower than the expected 100% faithfulness and grammaticality, which may have been due to low inter-annotator agreement or to presentation bias as suggested by the authors. Out of all five systems, See et al. (2017) and Chen and Bansal (2018) perform best, but are still more error-prone

than reference summaries.

We find that previous research has not established a detailed typology of summarization errors. Most work instead relies on on a binary distinction between correct and erroneous (Cao et al., 2018; Falke et al., 2019; Lebanoff et al., 2019) or faithfulness measured on a Likert scale (Goodrich et al., 2019). However, not all errors are created equal. Some errors might be less severe than others. As mentioned in Section 1, Durmus et al. (2020) is an exceptional case that looks at different kinds of errors related to faithfulness. Our work goes further, since it recognizes linguistic differences between factual errors, providing a more detailed typology.

2.3 Neural summarization systems

Here, we describe the summarization systems that generate the summaries used to create the typology (Section 3) and to carry out our comparative analysis (Section 4). We include two older approaches trained entirely from scratch on the summarization task, namely a pointer-generator architecture (See et al., 2017), henceforth referred to as **PG** and an RL-inspired rewriting paradigm (Chen and Bansal, 2018), **FAST-ABS-RL**. Additionally, two approaches using pre-trained language models are included: The first is **TRANSFORMER-LM**, proposed by (Hoang et al., 2019), a language-modeling approach leveraging GPT (a transformer-based model trained on roughly 7,000 books). The second is **BERTSUM**, an approach leveraging pre-trained BERT encoders (another transformer-based model trained on the books and the English Wikipedia), proposed by (Liu and Lapata, 2019). All four models were trained on the same split of the non-anonymized version of the CNN/Daily Mail dataset. PG and TRANSFORMER-LM directly train on the abstractive task and do not involve extraction. In contrast, BERTSUM performs initial fine-tuning on an extractive task and FAST-ABS-RL even involves an extractive sub-step directly in the pipeline.

3 Building the typology

In this section, we describe our methods and present the typology that we created.

3.1 Methodology

We collected the output of four summarization systems varying in a number of design aspects in order to capture as much linguistic diversity of generated text as possible. All systems were trained on the CNN/Daily Mail dataset (CNN/DM), a large corpus of news articles with associated abstractive summaries (Hermann et al., 2015), which has been widely used in the summarization literature. Generated summaries of test set articles as provided by the original authors were used. We conduct sentence-level annotation, allowing us to look at fine-grained differences.

Our typology was created in two steps. First, we carried out a card sort to establish an initial set of categories. For each of the four summarization systems, we randomly sampled 30 of its summaries, ensuring that each corresponded to a different article. Each summary was divided into sentences and one sentence was printed on a card, with the respective article printed above. This yielded a total of 393 sentences. Six experts in the news domain working at a news company sorted the cards (including one of the paper authors). Cards with similar errors were placed together in a pile. Then the experts iterated over the piles together, dividing and merging them until the sentences were grouped into a stable set of categories.

Second, we carried out a review of the categories in order to ensure that the boundaries of the categories were clear and to connect the categories to linguistic concepts. The review was carried out by the authors of the paper, two of whom were working at the news company. This group differed from the card sort group in that they have had training in linguistics. It was observed that some of the categories established in the card sort focused on surface nature of the error, others dealt more with the consequences of the error. This led us to establish a two dimensional typology, described in the following section.

3.2 Typology of summarization errors

The resulting error typology distinguishes two dimensions of summary error. First, 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. This dimension can help us to understand the cause of an error, potentially helping to establish how these errors can be avoided. It distinguishes the four categories in Table 1. Second, the *Meaning Dimension* describes the effect of the error on whether the sentence can be understood and how the reader interprets it. This dimension distinguishes six categories, presented in Table 2.

Omission	Copying words from an article sentence, but omitting necessary words or phrases.
Wrong combination	Copying words or phrases from multiple article sentences and combining them into an erroneous sentence.
Fabrication	Introducing one or multiple new words or phrases that cause an error.
Lack of re-writing	Failing to adequately re-write sentences, e.g., by not replacing referential expressions with their original antecedents in the text. When the antecedents are not present in the preceding summary context, this causes an error.

Table 1: The *Mapping Dimension* of summarization system errors

Malformed	
Ungrammatical	A sentence that is syntactically unnatural and would not be uttered by a competent speaker. Syntactically malformed.
Semantically implausible	A sentence that is semantically unnatural and would not be uttered by a competent speaker. 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 accommodation.

Misleading	
Meaning changed, not entailed	In the summary context, the semantic content assigned to a sentence is not entailed by the original article.
Meaning changed, contradiction	In the summary context, the semantic content assigned to a sentence is in contradiction to the article.
Pragmatic meaning changed	In the summary context, the sentence gains a pragmatic meaning not present in the original article. Or, a pragmatic meaning present in the article is lost.

Table 2: The *Meaning Dimension* of summarization system errors, seperated into errors resulting in malformed sentences and errors resulting in misleading sentences

This dimension provides insight into the interaction of linguistic concepts and factual correctness. Errors from the first three categories can be considered to be **malformed** sentences: They will cause readers to stumble and question the quality of the summary, but they do not have the potential to mislead. In contrast, the remaining three categories can be considered **misleading**: They could give rise to incorrect beliefs that would not have been produced by the article alone. Misleading errors can be equated with factual errors in traditional parlance. Examples of errors and the corresponding annotation can be found below.

To validate the typology, we computed the inter-annotator agreement of three annotators. We selected a random subset of 30 articles from the CNN/DM dataset. Three annotators (the authors) applied the typology to judge the summaries generated by all four systems for this subset of articles. The origin of the summaries was not specified and the summaries were presented in a random order for each article. Annotators could refer to the original article and no time restrictions were applied. Each sentence that contained an error was assigned both a Meaning and a Mapping category. For cases where there was no majority agreement, arbitration was used to reach agreement.

We analyzed the sentence-level inter-annotator agreement (Cohen's κ) of each dimension separately. Both showed moderate agreement (Meaning Dimension: $\kappa = 0.44$; Mapping Dimension: $\kappa = 0.46$). Further analysis of the annotations revealed that most disagreement was not between different categories in the dimension, but rather caused by raters not agreeing whether a sentence contains an error at all.

We reviewed all cases for which we disagreed on whether an error was present. There are two likely sources of lower than expected agreement. First, the annotation task is not trivial and requires maintaining close attention: A total of 14 misleading sentences were missed entirely by at least one annotator. Often, these sentences are perfectly plausible at the surface (cf. Example 1) and only a very close reading of both the article and the summary ensures they are identified. 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 annotator and acceptable by two others that reflected different personal views on whether certain edits had faithfully retained original meaning. Consider Example 2: It shows that it is plausible that prior knowledge that the annotators might have (here, about the football team in question) causes them to accept the sentence as faithful, while annotators without this knowledge might disagree.

4 Comparison of summarizers

In this section, we carry out a comparative analysis of the four summarization systems using the error typology. This analysis highlights the usefulness of the typology for achieving insight into the nature of summary errors. We made a random selection of 170 articles and one annotator annotated all four summaries for each article using the typology. These were combined with the previously annotated set of 30 articles. This yielded a total of 800 summaries with roughly 2600 annotated sentences. Sentence annotations were additionally aggregated to summary level: A summary is labeled as malformed if it contains at least one malformed sentence, but no misleading sentence. If it contains at least one misleading sentence, it is labeled as misleading.

4.1 Meaning dimension errors

Our comparative analysis focuses on the Meaning Dimension of the typology, starting with the sentence level errors. Figure 1 presents the distribution of errors at the level of malformed and misleading errors. Exact sentence and summary level rates are presented in Table 3. A larger table including the fine-grained categories is released with the annotations.

All systems produce both misleading and malformed errors, but the distribution is quite different. PG, which does not use pre-training, produces the fewest misleading sentences. Malformed sentences are much more common for PG and FAST-ABS-RL, which are trained from scratch, than for TRANSFORMER-LM and BERTSUM, which use pre-training.

Next, we look at summary-level errors. We see that around 40% summaries contain at least one error of any kind for three of the systems and FAST-ABS-RL faring worse at almost 75%. Between 1 in 3 and 1 in 10 summaries generated by our systems contain at least one misleading statement. Our observations are consistent with summary-level error estimates reported by Falke et al. (2019). Their estimates for PG (8%) and FAST-ABS-RL (26%) are both somewhat lower than our rates, but the general trend is reflected.

For all systems, the observed summary-level error rate is closely aligned with what would be expected if errors were distributed randomly across summaries. This means that longer summaries, such as produced by FAST-ABS-RL, will have a higher error-rate independently of the sentence-level error rate. This observation underlines the importance of our choice to carry out error analysis at the sentence-level.

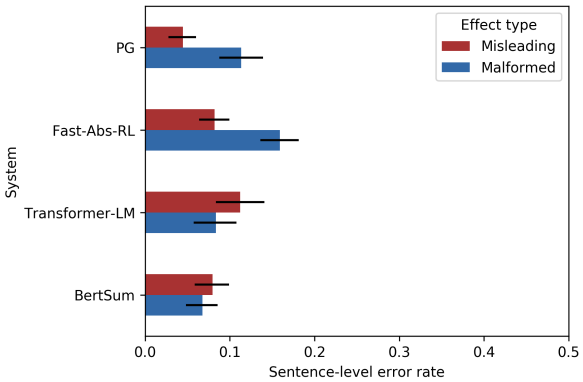

Figure 1: Sentence-level error type incidence rates by system, c.f. Table 3. 95 % CI obtained by bootstrap sampling.

4.2 Interaction of mapping and meaning

Next, we look into the interaction between the two error dimensions. Figure 2 illustrates the distribution of errors over summarization systems and the connection between the categories of the Meaning and Mapping dimensions. All systems suffer about equally from *lack of re-writing* and *wrong combinations*. However, the two pre-trained systems (TRANSFORMER-LM and BERTSUM) engage more frequently in *fabrications* and less frequently in *omissions*. FAST-ABS-RL suffers markedly from omissions.

Figure 2 also reveals that there is a correlation between the Mapping and Meaning dimension, but that essentially the dimensions are capturing two different aspects of summarization error. An important insight is that all four categories of Mapping error contribute to misleading errors, the more harmful type of Meaning error.

Headline	PICTURED: Mother-of-three who 'dropped her son in a cheetah pit' as it's revealed she is a CHILDCARE WORKER
Article excerpt	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 1: Wrong combination – Meaning changed, contradiction. Missed by two raters.

Headline	West Brom vs Leicester City: Team news, kick-off time, probable line-ups, odds and stats for the Premier League clash
Article excerpt	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 2: Omission – Pragmatic meaning changed. Two aspects of pragmatic meaning, i.e. that the fight has already started and that it was not for existence, but to avoid relegation, were resolved using background knowledge by two raters, but caused one rater to flag the sentence.

System	PG		FAST-ABS-RL		TRANSFORMER-LM		BERTSUM	
	Sent.	Sum.	Sent.	Sum.	Sent.	Sum.	Sent.	Sum.
Malformed	0.11	0.26	0.16	0.41	0.08	0.17	0.07	0.16
Misleading	0.04	0.12	0.08	0.32	0.11	0.19	0.08	0.23
Total	0.15	0.38	0.24	0.73	0.19	0.38	0.15	0.39
Avg # sentences	2.91		4.93		2.27		3.33	

Table 3: Error rates for the Meaning Dimension, sentence-level (Sent.) and summary-level (Sum.).

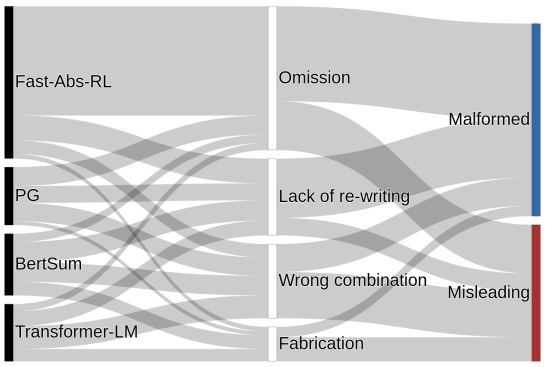

Figure 2: Sankey diagram showing the interaction of summarization systems, Mapping Dimension errors and Meaning Dimension errors.

4.3 Differences in abstractiveness

We turn now to the connection between abstractiveness and error. Improving the abstractiveness of a summary involves increasing the amount of rewriting. 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. Durmus et al. (2020) found that more abstractive systems are generally more error-prone, but did not look into the interaction of sentence-level error rates and abstractiveness. For this reason, we carry out a sentence-level analysis.

We calculate an abstractiveness score for each sentence in each summary as follows. For each sentence, we automatically select the closest document sentence in terms of word overlap. We then compute ROUGE-L. Normalizing by the length of the article sentence gives the precision of ROUGE-L and thus shows how much of the article sentence is retained. Similarly, normalizing by the summary length gives the recall of ROUGE-L, capturing how much of the summary originates from the closest document sentence. To get a combined metric, we compute ROUGE-L-F1, the harmonic mean of precision and recall for all ROUGE values. Sentences are then binned into two equal size bins, yielding a threshold of 0.705. We consider sentences about the threshold to have *high* abstractiveness and those below to have *low* abstractiveness.

Figure 3 displays the sentence-level error-rates for high and low abstractiveness summaries, separately for all four systems. Across all systems, higher abstractiveness is associated with a higher error rate. BERTSUM has a slightly lower error rate for highly abstractive sentences than the other systems with similar error rates. For

Headline	The Justice Department's questionable battle against FedEx
Article excerpt	*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 3: Lack of re-writing – No meaning can be inferred. System: PG

Headline	Prince Charles leads tributes to '100-year-old teenager' Hayley Okines as hundreds gather for her funeral.
Article excerpt	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 4: Omission – Ungrammatical. System: FAST-ABS-RL

largely extractive sentences (low abstractiveness), PG, TRANSFORMER-LM and BERTSUM perform about equally well, while FAST-ABS-RL has a higher error rate. These findings support the observation that an absolute difference in sentence error rate between systems could be explained not by one system being inherently better, but just being less likely to write more abstractively and thus more error-prone. We also observed that 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.

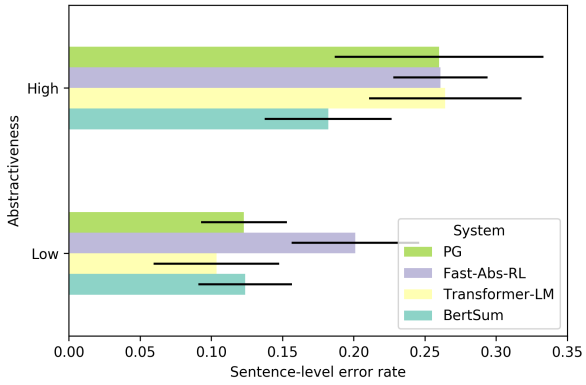

Figure 3: Binned ROUGE-F1 scores, average error rates in bins separately by system. 95 % CI obtained by bootstrap sampling.

5 Conclusion and outlook

In this section, we tie together the main contributions and insights of this paper, and discuss the avenues that it opens for future work.

5.1 Summary and discussion

In this paper, we have presented a typology of errors produced by automatic summarization systems, created by analyzing the output of four recent neural systems. The typology describes summary errors along a Mapping Dimension and a Meaning Dimensions, which are related, but are shown to capture different aspects of summary error. The Meaning Dimension is further divided into types of errors that describe malformed sentences and those that describe misleading sentences. The typology supports systematic analysis of abstractive summaries, and allows for focusing on the misleading sentences produced by automatic summarization systems. These errors are highly problematic because they impact the truth of a summary, i.e., its factual faithfulness to the original document.

Our comparative analysis has revealed the importance of using well-designed summarization metrics. With the wrong metrics, summarization systems will appear to be successful if the length of the summary or its abstractiveness has been decreased. In order to avoid these effects, and to achieve truly improved summaries, more advanced evaluation methods must be developed. The typology of errors that we have proposed here provides the basis for such methods. Metrics can become independent of length and abstractiveness if they take into account sentence-level errors and if they treat different errors differently. In particular, we recommend that misleading errors should be more important in signalling failed summaries than malformed errors.

If we consider the practical implications of improved summary evaluation, our typology makes a contribution in three related, but distinct directions: First, it can support the training of human assessors who can monitor live summarization systems in order to ensure that they do not lead to the publication of misinformation, which can have dangerous consequences. Second, it would be possible to train machine learning systems to support these human judgements. Third, it would be possible to improve automatic summarization systems in a way that

Headline	Andy Murray will jet straight from wedding with Kim Sears to run rule over prospective new assistant coach Jonas Bjorkman
Article excerpt	Mauresmo, who is to give birth some time in August, will be around eight months' pregnant during Wimbledon this summer.
Summary sentence	Mauresmo is eight months' pregnant *with her first child.*

Example 5: Fabrication – Meaning changed, not entailed. System: T-LM

Headline	Amazon removes new game that mocks anorexia sufferers by allowing players to throw food and sweets at character to fatten her up
Article excerpt	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 6: Wrong combination – Meaning changed, contradiction. System: BERTSUM

allows them to specifically avoid generating misleading sentences. The work we have presented here has set down a foundation for these directions.

In terms of improving summmarization systems, our typology has supported interesting insights: The four neural summarization systems that we studied differ considerably in their error patterns (cf. Table 3 and Figure 2). For example, we see that sentence-based rewriting such as in FAST-ABS-RL leads to omission errors, resulting in a higher risk of malformed sentences. More strikingly, the two pre-trained systems are somewhat more successful at avoiding malformed sentences, indicating that pre-training helps to improve grammaticality. This finding makes intuitive sense, as learning the statistical properties of a large corpus of text can be expected to boost the ability to generate grammatical text. However, misleading sentences and fabrication errors are more common for these pre-trained systems. Overall, we observe that if any one of these systems were to be used in a real-world scenario, readers could frequently end up confused, irritated or worst of all misled to hold incorrect beliefs. Using our typology these effects can be properly understood and quantified.

5.2 Future work

The typology presented in this paper opens several avenues for future work. First, here, we used summaries from only a single data set (CNN/DM) in a single domain (news). The typology should be validated on different data from different domains, which may allow more nuance to be added to the categories of the dimensions.

Second, further research is necessary in order to determine whether it is possible to achieve higher levels of inter-annotator agreement. Recall that we saw a relatively low agreement among annotators as to whether a sentence contains an error at all. This is in line with observations made by Lebanoff et al. (2019), who noted a relatively low inter-annotator agreement for binary faithfulness annotation. However, more investigation is needed.

We point out that the inter-annotator agreement has a possible dependency with the domain and data that is being analyzed. Specific linguistic properties of the CNN/DM dataset could have negatively affected agreement about the malformedness of sentences, namely telegraphic language style and the issue of reference summaries lacking relevant context. The lack of context issue is specific to the data set, which omits the article headline, even though summaries often rely on it for interpretability. This means that some reference summaries are hard to understand in isolation, and could potentially bias systems to imitate the style. Summary sentences that suffer from these issues are a likely source of annotator disagreement.

We hope that researchers will build on and continue to refine the typology that we have presented here. For example, more detailed study of how human judgement interacts with malformed vs. misleading errors could lead to an improvement in the category descriptions or in the divisions between the categories. A refined typology would support standardization of the judgement protocols for automatically generated summaries, which would in turn help fight the adverse effects of factual errors.

Acknowledgments: We thank FD Mediagroep for conducting the Smart Journalism project which allowed us to perform this research.

References

Alyssa Appelman and Mike Schmierbach. 2018. Make no mistake? Exploring cognitive and perceptual effects of grammatical errors in news articles. *Journalism & Mass Communication Quarterly*, 95(4):930–947.

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

Yen-Chun Chen and Mohit Bansal. 2018. 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)*, pages 675–686.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information Processing & Management*, 56(5):1794–1814.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. 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*, pages 2214–2220. Association for Computational Linguistics.

Yang Gao, Christian M Meyer, and Iryna Gurevych. 2019. Preference-based interactive multi-document summarisation. *Information Retrieval Journal*, pages 1–31.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175.

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

Andrew Hoang, Antoine Bosselut, Asli Celikyilmaz, and Yejin Choi. 2019. Efficient adaptation of pretrained transformers for abstractive summarization. *arXiv:1906.00138 [cs]*. ArXiv: 1906.00138.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551. Association for Computational Linguistics.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv:1910.12840 [cs]*. ArXiv: 1910.12840.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. *arXiv:1910.00203 [cs]*. ArXiv: 1910.00203.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3721–3731.

Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, 52(1):101–148.

Klaus-Michael Lux. 2020. On the factual correctness and robustness of deep abstractive text summarization. Master's thesis, Radboud University, Nijmegen, August.

Maxime Peyrard. 2019a. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard. 2019b. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. 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)*, pages 1073–1083.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? Pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29. Association for Computational Linguistics.