The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature

Anonymous ACL submission

Abstract

Although multi-document summarisation (MDS) of the biomedical literature is a highly valuable task that has recently attracted substantial interest, evaluation of the quality of biomedical summaries lacks consistency and transparency. In this paper, we examine the summaries generated by two current models in order to understand the deficiencies of existing evaluation approaches in the context of the challenges that arise in the MDS task. Based on this analysis, we propose a new approach to human evaluation and identify several challenges that must be overcome to develop effective biomedical MDS systems.

1 Introduction

With the number of biomedical publications doubling every two years (Cios et al., 2019), it is difficult for medical professionals to incorporate new, often contradictory, evidence into their daily work, as it would require appraising, comparing and synthesising the outcomes of multiple primary studies (Sackett and Rosenberg, 1996). Systematic reviews, which are published for this purpose, provide only a partial solution, as they are very time-consuming to write and thus can be unavailable for newer clinical questions or quickly become outdated. In this context, the ability to automatically summarise evidence from multiple studies is of high practical importance. The task, however, is more challenging than general multi-document summarisation (MDS), as the summaries must correctly draw conclusions based on often contradictory studies, and aggregate details such as groups of patients or names and doses of treatments, in addition to dealing with often-cited difficulties posed by biomedical text such as complex lexical and semantic relationships between concepts (Plaza et al., 2011). Though recent approaches to biomedical summarisation acknowledge the additional challenges of the task and try to incorporate some domain-specific knowledge to deal with them (Wallace et al., 2021; Shah et al., 2021; DeYoung et al., 2021), we still lack a solid understanding of how well the current models are able to do that, how useful the generated summaries are, or how to measure our progress.

In this paper, we propose a systematic approach to human evaluation of biomedical summaries, and apply it to analyse the summaries generated by two state-of-art systems. We examine the common errors in generated summaries and the correlation of automatic metrics such as ROUGE (Lin, 2004) with our evaluation results. We choose summarisation models proposed by DeYoung et al. (2021), as they not only demonstrate the abilities of end-to-end neural models, but also incorporate domain-specific knowledge such as entity prompts.

The contributions of this paper are as follows: (1) We propose a new approach to human evaluation of biomedical summaries based on binary categorical ratings, which ensures that the results are interpretable, reliable and easily reproducible by non-expert annotators. (2) We show that current approaches to summarisation suffer from excessive copying from the prompt and an inability to aggregate important details from primary studies. (3) We show that automatic metrics such as ROUGE cannot reliably distinguish between factual and erroneous summaries. (4) We suggest several reasons which may explain the poor summarisation performance, and show that it is necessary to redefine our approaches to biomedical MDS. Though our focus is on the biomedical field, we raise some issues common to cross-domain summarisation, and propose a consistent approach to human evaluation and error classification which can be easily transferred to other domains.

2 Related studies and motivation

Although the importance of MDS in the biomedical domain was recognised around 20 years ago with studies such as McKeown et al. (1998) and Becher

et al. (2002) defining some requirements and operations specific to biomedical summarisation (e.g. the ability to resolve contradicting statements), until recently there have been few end-to-end systems (e.g. PERSIVAL (Elhadad et al., 2005)) due to the complexity of the task. In the last few years, apart from several shared tasks and challenges dedicated to multi-answer biomedical summarisation, including MEDIQA 2021 (Ben Abacha et al., 2021) and BIOASQ (Nentidis et al., 2021), several major threads of research have emerged. Wallace et al. (2021) and DeYoung et al. (2021) incorporate entity and discourse level prompts into their end-to-end neural summarisation models. Shah et al. (2021) revived the idea of symbolic MDS (Radev and McKeown, 1998) by combining a deterministic content plan with a pre-trained language model. Here, we are particularly interested in the model by DeYoung et al. (2021) as it reflects the setting of summarisation systems "in the wild": their input is all clinical trials cited by a systematic review rather than a sample of trials which the review was based on (Wallace et al., 2021) or a curated list of trials relevant to the summary (Shah et al., 2021).

In terms of evaluation metrics, there has been a growing awareness of the inability of ROUGE to reflect the factual accuracy of summaries, so some other automatic metrics, including inference-based (Maynez et al., 2020) and question-answering-based methods (Chen et al., 2018; Wang et al., 2020) have been proposed. There have also been attempts to make the human evaluation more objective and systematic by defining linguistically grounded error categories and evaluation criteria (Huang et al., 2020; Pagnoni et al., 2021). In the biomedical domain, although there are some new automatic measures proposed, such as Aggregation Cognisance (Shah et al., 2021) — which measures the ability of the model to recognize if the input texts are in agreement or contradiction — and ΔEI (DeYoung et al., 2021) — which reflects the alignment of summaries in terms of direction of their findings — human evaluation has been primarily done using the Likert scale (Wallace et al., 2021; Shah et al., 2021), which makes it difficult to reproduce and interpret. In this work we aim to close this gap by establishing a more reliable, grounded and objective human evaluation framework and showing its application by assessing the summaries generated by the state-of-the-art MDS system of DeYoung et al. (2021).

3 Summarisation models

The models we evaluate were trained on a large-scale dataset comprising 20K systematic reviews and 470K primary studies developed by DeYoung et al. (2021). The conclusions, taken from the abstract of the review, are the target for the summarisation. The input consists of a prompt in form of the *Background* section of the systematic review, and the abstracts of up to 25 studies cited in the review. As the prompt (*Background*) describes the review's objective, the task is similar to query-based summarisation, but with an extensive prompt.

We use the two summarisation models explored in DeYoung et al. (2021): BART (Lewis et al., 2020) and LongFormer (Beltagy et al.). Both models are similar in architecture but differ in their approach to handling long input sequences: for LongFormer (LED henceforth) *Background* is concatenated with all studies and encoded together before feeding to the decoder, while for BART each study is concatenated with *Background* and encoded separately; then their encodings are concatenated together and fed to the decoder. To adapt the models to the biomedical domain, the authors decorate the inputs by adding special tags around PICO (Richardson et al., 1995) elements, namely <pop>, <int>, <out> , and also by marking the different sections such as *Background*.

4 Evaluation process and criteria

We sampled 100 reviews each from test summaries generated by BART- and LongFormer (LED) based models. To evaluate them in a more systematic manner, we define the following quality dimensions which capture both factuality and fluency:

4.1 Factuality

Though factual errors are often attributed to hallucinations (when the model generates entities not present in the source), they can also be due to other reasons, such as omission of important details, incorrect order of tokens or syntactic relations between them. Rather than classify the factuality errors by their reason, however, we treat the summaries as a combination of important biomedical entities and the relations between them, and define the quality dimensions related to them as follows:

PICO correctness

The PICO (**P**atient/problem, **I**ntervention, **C**omparison, **O**utcome) scheme captures the

most important entities for answering biomedical questions (Richardson et al., 1995), such as "Does the acupunture (*intevention*) help to decrease inter-ocular pressure (*outcome*) in patients with glaucoma (*patient*)?". We consider a generated summary to be correct from the point of view of PICO when it mentions the same patient population, intervention and outcome (in the same lexical form or paraphrased) as the original summary.[1] When doing so, we apply strict restrictions regarding the semantic hierarchy of PICO concepts in the generated and target summaries: if one of the concepts is a hypernym of another (for example, *acetaminophen* and *analgetics*), we consider it to be a factual error, as the findings of clinical trials should not be generalized or narrowed to other intervention types, patient groups, or outcomes. It should be noted that though the PICO schema is more applicable to treatment trials, we apply these categories more broadly, as there are also clinical trials related to diagnostics, risk factors, biomarkers etc.[2]

Direction correctness

Lehman et al. (2019) defined three directions of the intervention's effect with regards to the outcome: *significantly increases*, *significantly decreases* and *no significant difference*. We keep this three-way classification, but redefine it as *positive effect*, *negative effect*, or *no effect*, which allows us to judge based on the semantics and sentiment orientation of expression rather than the surface form. As an example, consider the following:

- **Generated**: NIV is associated with an *improvement* in mortality.
- **Target**: NIV had great advantage... in *reducing* mortality.

If we follow the classification proposed by Lehman et al. (2019), these summaries have different directions in relation to "mortality" ("improvement" shows the direction of *increases*, while "reducing" has the direction of *decreases*), thus the generated summary would be erroneously considered wrong. The proposed classification of *positive/negative/no effect* avoids that, capturing the semantic orientation rather than literal meaning, similar to aspect-based sentiment analysis (Liu, 2012). It also more naturally extends to situations where the intervention does not directly affect the outcomes (so that no *increase* or *decrease* is possible), such as when we talk about the effectiveness of a diagnostics method, and to other clinical question types. For example, we assign the *positive* label if the review identifies the optimal intervention (*Which intervention works best?*), *negative* if it shows the most undesired intervention (*What are the most important risk factors?*), and *no effect* if such interventions cannot be identified.

Modality

As a linguistic category, modality reflects the possibility of a proposition (i.e. *X might increase Y* vs. *X increases Y*), but here we define it in a more pragmatic way to denote how certain we are of available evidence and thus how strong our claim is. In particular, we define we following levels of certainty: *strong claim*, *moderate claim*, *weak claim*. There are also two labels for statements where there author cannot draw any conclusions based on the evidence available to them (*no evidence*) or when the statement is descriptive and does not contain any claims regarding the direction of effect (*no claim*). Below we briefly describe the ways the modality is expressed:

Strong claim: these claims are modified by strengthening expressions such as *remarkably* or *considerably*: *MSC infiltrations... [lead] to an overall remarkable improvement*. The author can also directly appeal to the quality of available evidence: *High-quality evidence indicates that diet... can reduce the risk of excessive GWG*.

Moderate claim: this is usually an unmodified proposition, such as *Warming-up before an operative procedure improves a trainee's... performance*.

Weak claim: such statement can be hedged in multiple ways including modal verbs (e.g. *may*), introductory clauses (*It appears that...*), or adverbs (*likely*). However, the author can directly comment on the reliability of evidence (*There is **initial evidence** supporting the effectiveness*), discrepancy of the results (*denosumab... has shown a **positive but variable histological response***), or the limited applicability of findings (*HBMS programmes... have **short-term** beneficial effect*).

No evidence: there is either no primary evidence regarding the clinical question, or no conclusions can be drawn from it on account of its low quality

[1] Following Nye et al. (2018), we omit the *C*omparison (alternative intervention), as it is often a no-treatment control which is implied rather then mentioned explicitly.

[2] For example, in a study examining risk factors influencing poor response to a treatment, such risk factors as *young age*, rather than the treatment itself, are *interventions*, while the therapy response is the *outcome*.

or conflicting results. These statements are usually introduced by such clauses as *There is **insufficient evidence** to support...*.

No claim: a summary can mention the clinical question, but make no statements regarding the effect of the intervention: *[This] is the first systematic review to assess the effect of inhaled steroids on growth in children with asthma.*.

It should be noted that *modality* is different from statistical significance of an intervention's effect, which is captured by *direction*. For example, even if a clinical trial has a statistically significant effect, we can be uncertain of its results due to bias in the cohort, e.g. a small sample size. In the case of MDS, even if each of the underlying studies has shown a significant effect, their direction can be contradictory, which results in the *no evidence* judgement. On the other hand, we can be very certain that an intervention does not have any effect (*There is ... strong evidence of no significant difference between acupuncture and sham acupuncture*). Probably the most important distinction to make here is between cases where we have *no evidence* (*There is insufficient evidence to determine whether... LCPUFA improves... growth of preterm infants*) vs where we have enough evidence to state that there is *no effect* (*no clear long-term benefits or harms were demonstrated for preterm infants receiving LCPUFA*).[3]

The reason we include modality as a separate evaluation aspect is that it reflects the quality of the evidence and its potential usefulness to the medical professionals; thus, if primary studies report that a treatment *may* work, we do not want their summary to assert that the treatment *works*. Likewise, if it is impossible to aggregate the evidence with any certainty, the summary must state that the current evidence is insufficient rather than draw a particular conclusion. In this respect, modality is related to the newly-introduced category of scientific *ignorance* (Boguslav et al., 2021)) as it helps to evaluate the state of our knowledge regarding a particular clinical question.

4.2 Fluency

Errors in this category can make it difficult to read and understand the summary, but do not affect its meaning.

Grammatical correctness

This category includes morphology and syntax mistakes, such as incorrect verb form or clause structure, but also lexical mistakes (incorrect word choice) leading to grammar errors. For example, a phrase *the is* instead of *there is* would be classified as a grammar rather than lexical error.

Lexical correctness

This category is for spelling mistakes which do not affect grammar and meaning.

Absence of repetition

Neural summarisation systems commonly generate repetitive content, which can affect fluency to the point of unintelligibility. Here, repetitions are regarded as a fluency mistake only when they do not make the sentence factually or grammatically incorrect.

4.3 Evaluation process and reliability

We judged each pair of target and generated summaries as correct or wrong based on the categories outlined above.[4] To be considered valid, the summary must be correct across all these dimensions; to be considered useful or factually correct, it must be aligned with the target summary in the first three dimensions.

Although it might seem that some errors are "worse" than others (e.g. completely mixing up the interventions can seem to be a more severe mistake than mentioning a more generic concept), we treat the errors as binary. The reason behind this is twofold: first, it allows us to decompose the complex task of human evaluation into a series of pairwise yes/no decisions and thus make it easier and more objective (similar to what is already a standard practice in human evaluation of biomedical machine translation (Jimeno Yepes et al., 2017)); second, we argue that the "minor" errors are more dangerous in practice: while a completely irrelevant answer is likely to be spotted as incorrect by a medical professional, a tiny mistake in the summary can go unnoticed and thus the conclusions can be applied to a different situation than intended or with a different degree of certainty.

[3] One simple test to distinguish them is that we can add a *modality*-modifying expression on top of the *no effect* statement (*Long-chain omega-3 **probably** has... **no effect** on new neurocognitive outcomes*), while it is impossible to do this for *no evidence* or *no claim* propositions which already express the modality.

[4] In cases where the target review contained several statements, while the generated summary had only one proposition, we matched it to the closest statement in the target summary; if we required a perfect multi-proposition to multi-proposition match, the results would have been much poorer.

	PICO	Direction	Modality	Grammar	Lexical	Non-redundancy	Factually correct	Overall
Agreement	87%	83%	84%	86%	98%	95%	94%	89%
Gwet's AC1	0.80	0.70	0.77	0.75	0.97	0.95	0.93	0.82
Fleiss' κ	0.66	0.62	0.67	0.60	0.93	0.88	0.86	0.73

Table 1: Inter-annotator agreement by category.

To assess the robustness of our evaluation criteria, we asked five additional annotators, only one of whom was a medical professional, to evaluate the quality of 40 generated summaries. The details of evaluation process together with the annotation instructions and metrics used can be found in A. Table 1 presents the average agreement between each of five annotators and the expert (in terms of percentage of agreement and Gwet's AC1), as well as Fleiss' κ for all six annotators. In general, we found high agreement of annotators with the expert, and substantial agreement between all annotators, which is remarkable considering the difficulty of the task and the size of the raters group. Most of the mistakes were not systematic, though some annotators struggled to differentiate between *no evidence* and *no effect* statements. Despite some discrepancy in the category-level annotation, when we aggregate the scores across the first three categories to determine if a summary is factual, the results are highly reliable, with almost perfect agreement with the expert and strong agreement among annotators, which shows that our method can be used to robustly evaluate the usability of summaries.

5 Results

5.1 Correctness by category

As shown in Table 2, less than 5% of generated summaries did not have any errors; even if we disregard the fluency errors, only around 10% of summaries are factually correct and thus usable. Overall, the generated summaries are quite fluent, with surprisingly low redundancy; it is the factual accuracy, especially in terms of PICO and modality, that is problematic.

5.1.1 PICO

Among the PICO categories, *Intervention* is the most problematic, while *Patient* is usually generated correctly (Table 3). Below we outline some typical PICO errors:

More narrow concepts in the generated summary, usually copied from the primary studies: *women with pre-eclampsia* instead of *women* as Patient, *robocat* instead of *companion-type robots* as Intervention, *preventing HPV 16/18* instead of *preventing HPV* as Outcome.

More generic concepts in the generated summary, usually copied from the *Background*. For example, the generated summary mentions *topical agents*, while the review deals specifically with their *innovative reformulation*; the review is about a particular drug (*nedocromil sodium*) while the generated summary mentions the drug category (*inhaled corticosteroids*).

Incorrect elements copied as Intervention and Outcome: the generated summary is about the effect of *laxatives* on *constipation*, while the review examines the effect of *constipation* on *physical and mental well-being*. In some cases, the elements are correct, but the relation between them is reversed: a review studies whether *depressive symptoms* lead to *sleep disturbances*, while the generated summary is about the effect of *insomnia* on *depression*.

Hallucinated elements: surprisingly, some incorrect PICO elements have the same stem as the correct ones: *developing countries* instead of *developed countries* and *congenital hypothyroxinaemia* instead of *congenital hypothyroidism* , which seems to be due to generating a more prominent candidate continuation in a multi-token entity.

5.1.2 Direction

We calculate the direction accuracy only for the samples where the consistency of direction can be reliably determined, that is, where none of the two summaries have *no evidence* or *no claim* modality. Remarkably, if we keep the direction separate from modality, the performance for this category is quite good, which shows that getting the semantic orientation of the proposition right is relatively easy if the model is certain enough to make a statement. However, the confusion matrix for this category (Figure 1) shows that both high accuracy of this category and the highest number of mistakes can be attributed to the overwhelming presence of findings with the **positive** direction in the data. Therefore, the "easiness" of this dimension is not because the models learns to correctly capture the direction

	PICO	Direction	Modality	Grammar	Lexical	Non-redundancy	Factually correct	Fully correct
BART	45%	77%	45%	75%	69%	85%	9%	3%
LED	40%	75%	44%	63%	73%	89%	8%	4%

Table 2: Correctness by category.

	Patient	Intervention	Outcome	Fully correct
BART	83%	66%	79%	45%
LED	86%	63%	68%	40%

Table 3: Correctness by PICO element type.

of primary studies, but rather because the default **positive** direction is most often correct due to the specifics of clinical questions.

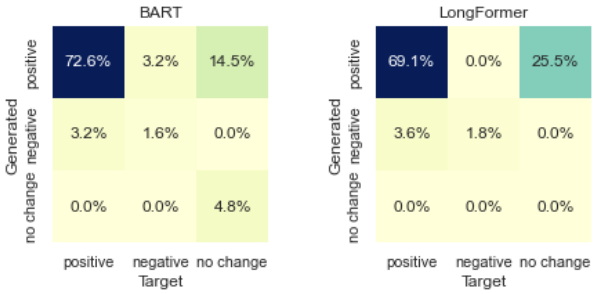

Figure 1: Direction of the generated vs target summaries.

5.1.3 Modality

In contrast to the previous category, the models produce more varied content in terms of *Modality*, which reflects a less skewed distribution in the data (see Figure 2). Though there is still a clear "majority" category (*moderate claim*), most of the errors are not due to generating too many moderate claims. In fact, for both BART and LED the most common problem is generating *no evidence* sentences instead of moderate and weak claims; for LED, there is also a good proportion of errors due to not making any claim at all. Interestingly, the number of times when the adjacent categories were mixed up (weak ↔ moderate, moderate ↔ strong) is lower than number of mistakes due to confusing quite distinct categories of *no evidence/no claim* and *moderate evidence*. Thus, even though the models sometimes correctly pick up cues showing weakness of evidence or its moderate quality, they often "give up" on trying to make any conclusion. This is especially true for LED, which generates substantially more *no claim* summaries than BART.

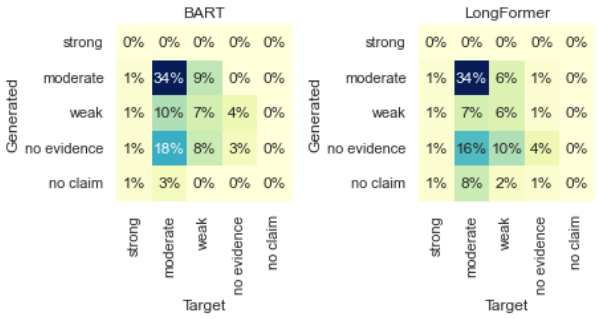

Figure 2: Modality of generated vs. target summaries.

5.1.4 Grammatical and spelling errors

The mistakes in these categories are quite uniform in the sense that they seem to be an artefact of tokenization and decoding. For example, the vast majority of spelling errors are due to incorrect merging of subwords including the article *The* at the beginning of a sentence, for example *TheCLUSIONS* instead of *The CONCLUSIONS*. The grammar mistakes are also usually caused by incorrect token *The* at the initial position: *The is insufficient evidence*, though some other errors occur at this position: *There systematic review of strategies*.

5.1.5 Repetitions

Contrary to our expectations, the amount of repetitions was small, so it is difficult to make conclusions regarding their patterns. However, there was a tendency to include prominent tokens, often paraphrased, both in the outcome and patient 'slots', which sometimes led to redundancy: *acupuncture for LBP in patients with chronic low back pain*

5.2 A closer look at the output

How much is copied from the *Background*?

As the evaluation results in the previous section were discouraging, we found it necessary to examine the way summary were generated. Upon further analysis, the majority (91% for BART and 85% for LED%) of the generated summaries are very similar in content to the *Background* section of the systematic review, which is supposed to contain a prompt for the model rather the content to be actually summarised. More specifically, they

Target	Partial replacement using both classes of scaffolds achieves significant and encouraging improved clinical results when compared with baseline values or with controls when present
Background	We systematically review the literature on clinical outcomes following partial meniscal replacement using different scaffolds.
BART	This is the first *systematic* **review** of **the literature on** the **clinical outcomes following** *meniscectomy* **using different scaffolds**.
LED	The is the first *systematic* **review** to evaluate the **clinical outcomes following** *meniscectomy* **using different scaffolds**.

Table 4: Copying from the objectives statement. Directly copied words are in bold, while paraphrases are in italic.

	Background			Target		
	R-1	R-2	R-L	R-1	R-2	R-L
BART	37.36	23.18	30.62	27.34	9.23	20.64
LED	36.61	21.93	30.05	26.98	8.84	20.39

Table 5: ROUGE scores of generated summaries against the *Background* section and the correct *Target* summary.

copy the objectives or hypothesis sentence with various degree of paraphrasing. A typical example of such copying is provided in table 4; though some paraphrasing is present, the generated summaries do not contain any information which cannot be inferred from the objectives sentence. Worse of all, they do not answer the question but rather restate it (*no claim*). To check whether this tendency is present in generated summaries in general, we calculated the unigram overlap (ROUGE-1), bigram overlap (ROUGE-2) and the longest n-gram overlap (ROUGE-L) between them and two "golden" summaries: the target summaries and *Background* text for all samples in the test set. As can be seen from Table 5, the generated summaries are much closer to the *Background* section than to the *Target* summaries; high ROUGE-2 and ROUGE-L scores against the *Background* also reflect the tendency to copy longer sequences literally.

How much is copied from studies?

Only a third of examined summaries (34% for BART and 30% for LED) included any details taken from primary studies that were meant to be summarised rather than from the prompt (*Background*). Though this in itself is concerning, it is even more striking that only in 4/2 of BART/LED summaries the model managed to copy some useful information from the studies, whereas in the majority of cases copying from studies actually caused mistakes. These mistakes can be divided into two roughly equal groups: (1) the entity copied from the studies was too narrow, which means that there was no aggregation of entities across studies which examined different groups of patients, interventions or outcomes.[5]; (2) an entity unrelated to the clinical question but frequently mentioned in the studies is copied.[6]

How much is hallucinated?

Though hallucinations are a widely known issue with neural abstractive summarisation, in the data we analysed less than 4% of summaries had incorrect details which could not be attributed to either the prompt or the included studies.

Do the summaries follow the usual discourse patterns?

Around 68% of the analysed summaries are prepended by standard phrases such as *This systematic review suggests...*. To check how widespread such phrases are in generated summaries in general, we also calculate their frequency in the whole test set: *There is insufficient evidence to support...* occurs in 25%/19% of BART/LED summaries; and *The results of this systematic review suggest...* 15%/14% for BART/LED. As was shown above in Section 5.1.3, LED makes more *no claim* statements than BART: 12% of LED summaries begin with *The is the first systematic review*, while only 2% of BART summaries do so. Overall, at least 55% of all summaries have the canned phrases we identified.

Do our metrics correlate with ROUGE scores?

Though we used ROUGE to determine the amount of lexical overlap and copying in Section 5.2 above, we do not consider it to be a reliable metric for quality estimation, especially in terms of factuality, as it does not correlate with any factuality dimensions we examined or factual accuracy in general. To determine whether the factually correct summaries had higher ROUGE scores than incorrect ones we performed a series of Student t-tests

[5]More specifically, this can be due to adding an adjective modifier (*primiparous women* instead of *women*) or copying one of the concept's hyponyms (*robocat* instead of *companion-type robots*)

[6]For example, a purpose of one review was to identify dry eye symptoms rated as most uncomfortable, but as the majority of primary studies mentioned *artificial tears* for treating this condition, this concept was included in the generated summaries.

comparing summaries with correct and incorrect PICO, direction and modality, as well as summaries with no mistakes in any of these categories versus summaries with at least one mistake. There was no statistically significant difference in terms of ROUGE-1, ROUGE-2 and ROUGE-L scores between correct and incorrect summaries in all of these tests for both BART and LED.[7] As an example, the distribution of ROUGE-1 scores for correct and incorrect BART summaries is shown in Appendix B.

6 Discussion

In this section we point out some issues which could explain the poor performance of the summarisation systems, and show how they relate to the principles underlying the aggregation of medical evidence. We present these as challenges to be tackled in MDS system development.

Perform multi-aspect summarisation

A large number of reviews (around 40%) had multiple propositions, that is, sets of PICO elements and relationships between them. For example, a review can study effects of a drug in terms of different outcomes, and each of these outcomes can have a different direction and modality. As the result, we are dealing with multi-aspect summarisation, and it can be difficult for the model to correctly identify and reproduce several sets of prominent entities and relationships.

Aggregate, don't just summarise

Primary studies are rarely, if ever, conducted for all possible groups of patients, drugs in a particular class, or outcomes. Thus to answer a clinical question, we need to aggregate across such entities. For example, if a systematic review studies the effects of counselling on breastfeeding rates across the globe, and the majority of underlying studies mention "developing countries" while other refer to specific locations such as "Baltimore", the generated summary can have a narrower *Patient* group than it should. Similarly, if primary studies examine the effects of different types of HPV vaccine (HPV-6, 11, 18 etc) for different groups of patients, we would need to aggregate across them to be able to make conclusions about the effectiveness of HPV vaccines at large.

Find answers even when they are not obvious

In many cases, the primary studies are not considering exactly the same question that the review needs to answer. For example, the review may be about the effects of depression on sleep quality, while the underlying studies examine the effects of disrupted sleep on depression. Sometimes the answer needs to be inferred based on prior knowledge. One of the reviews, for example, explored the risks of mortality due to salmeterol, while the studies included in it did not even mention mortality but rather examined potentially lethal side effects.

Learn to answer more complex questions

While the majority of clinical questions are in the yes/no form ("Does the intervention A have an effect on the outcome B?"), and the model can answer them by rephrasing the question, some questions require more difficult operations. For example, a clinical question might ask *which* strategy is more effective for preventing asthma (which requires to compare interventions), *what* education methods exist to manage hyperphosphatemia (which requires listing different interventions), or even *why* behavioral interventions work (which requires reasoning about various aspects of interventions).

7 Conclusions

In this research, we attempted to bring the importance of factuality in biomedical MDS into attention, and demonstrated that the current models are still unreliable in this respect. Moreover, we showed that they fail to pick up and aggregate important details from multiple documents, excessively relying on the prompt. To support our analysis, we established a simple and reproducible human evaluation benchmark which reflects aspects of quality important for biomedical MDS but can be translated into other domains. Finally, we showed that the progress in biomedical MDS will be limited unless we acknowledge the domain-specific challenges of the task and work towards overcoming them. Though we focused our efforts on a particular domain, we hope that this work prompts taking a closer look at the summarisation results in other areas, as only objective evaluation of what the models are capable of and prone to do will allow us to improve them.

[7] We performed the same experiments with BERTScore (Zhang et al., 2020), and though it was marginally able to differentiate between the summaries with correct and incorrect PICO, it could not capture the direction or the modality of the claim, so overall the results were statistically insignificant.

8 Ethical considerations

Done right, biomedical MDS can significantly facilitate the practice of evidence-based medicine; done wrong, however, it creates risk of misinterpretation of evidence and subsequent malpractice. For this reason, we argue that the factual accuracy of biomedical summaries should be decided on a rigid yes/no scale, and only the summaries matching in all details and intents should be considered factually correct and thus useful. In this paper we show that we still have a long way to go before biomedical summarisation systems can be reliably used and trusted, and highlight the importance of robust human evaluation in this domain.

References

Margit Becher, Brigitte Endres-Niggemeyer, and Gerrit Fichtner. 2002. Scenario forms for web information seeking and summarizing in bone marrow transplantation. In *COLING-02: Multilingual Summarization and Question Answering*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Mayla R Boguslav, Nourah M Salem, Elizabeth K White, Sonia M Leach, and Lawrence E Hunter. 2021. Identifying and classifying goals for scientific knowledge. *Bioinformatics Advances*, 1(1).

Ping Chen, Fei Wu, Tong Wang, and Wei Ding. 2018. A semantic QA-based approach for text summarization evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Krzysztof J Cios, Bartosz Krawczyk, Jacquelyne Cios, and Kevin J Staley. 2019. Uniqueness of medical data mining: How the new technologies and data they generate are transforming medicine. *arXiv preprint arXiv:1905.09203*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MS^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513. Association for Computational Linguistics.

Noemie Elhadad, Min-Yen Kan, Judith Klavans, and Kathleen McKeown. 2005. Customization in a unified framework for summarizing medical literature. *Artificial Intelligence in Medicine*, 33(2):179–198.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. Association for Computational Linguistics.

Eric Lehman, Jay B DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of NAACL-HLT*, pages 3705–3717.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Kathleen McKeown, Desmond Jordan, and Vasileios Hatzivassiloglou. 1998. Generating patient-specific summaries of online literature. In *Proceedings of of Intelligent Text Summarization, AAAI Spring Symposium*.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2021. Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 239–263. Springer.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Laura Plaza, Alberto Díaz, and Pablo Gervás. 2011. A semantic graph-based approach to biomedical summarisation. *Artificial intelligence in medicine*, 53(1):1–14.

Dragomir Radev and Kathleen McKeown. 1998. Generating natural language summaries from multiple online sources. *Computational Linguistics*, 24(3):469–500.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12–A13.

David L Sackett and William MC Rosenberg. 1996. Evidence based medicine: What it is and what it isn't. *BMJ: British Medical Journal: International Edition*, 312(7023):71–72.

Darsh Shah, Lili Yu, Tao Lei, and Regina Barzilay. 2021. Nutri-bullets hybrid: Consensual multi-document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5213–5222, Online. Association for Computational Linguistics.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. In *AMIA Annual Symposium Proceedings*, volume 2021, page 605. American Medical Informatics Association.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Evaluation

We recruited 5 volunteer annotators to evaluate the correctness of generated summaries in terms of the criteria we specified. Before the evaluation we did a pilot round where we presented the instructions and asked the annotators to judge 6 randomly selected summaries. An excerpt from the instruction and the form provided to annotators are shown in Figures 3 and 5, respectively. On average, annotators spent 30 minutes reading the instructions and evaluating the pilot summaries. After providing feedback, we asked them to evaluate 40 other randomly selected summaries (20 for each of BART and LED). The average reported speed of evaluation was 2 minutes per summary. We report the inter-annotator agreement for each of the evaluated categories (see Table 1) using the following metrics: average accuracy-type percentage of agreement with the expert annotator (author of the paper) and the average Gwet's AC1 score (Gwet, 2014) against the expert annotator, to show how accurate is the evaluation produced by annotators with minimal training, and Fleiss' κ to show the amount of disagreement between all six annotators. We choose Gwet's AC1 score rather than Coppen's κ as it is a more reliable metric for data with a strong majority class as in our case, where, for instance, almost all summaries have correct spelling.

B Distribution of ROUGE scores for correct and incorrect summaries

The distribution of ROUGE-1 scores for generated BART summaries with correct vs incorrect PICO elements, direction and modality, as well as for factually correct and wrong summaries, is shown in Figure 6.

Task description

You will be given a one-sentence summary of primary studies generated by a model, and a correct human written summary of the same studies written by a human. You will need to evaluate the generated summary (GS) against the target (human-written) summary (TS). In particular, you will need to determine if:

1) The **PICO** (**P**atient/**p**roblem, **I**ntervention, **O**utcome) components are the same in GS and TS.
2) The summaries have the same **modality**, which is how certain we are about the conclusions. Is the author making a **strong** claim, a **moderate** (usual) claim, a **weak** claim, **no claim** at all or do they say there is **not enough evidence** to make a conclusion?
3) The **direction** of the results is the same, that is, does Intervention has a desired/**positive** effect on the outcomes, a **negative**/undesired effect, or **no effect** at all?
4) The **grammar** of the GS is correct.
5) The **spelling** of the GS is correct.
6) There are no unnatural **repetitions** in the GS.

Please read below for more detailed criteria for these 6 categories.

Categories

PICO
P (Patient/problem) is a disease, condition, or a description of a patient group, such as "newborn children" or "diabetes". Sometimes P is missing at all and "all people" is implied.
I (Intervention) is usually a drug ("Panadol"), surgery ("resection") or other treatment/procedure ("mechanical ventilation"), but can be anything that influences outcome, for example a risk factor such as "smoking" or "age".
O (Outcome) is what we are measuring and what is influenced by Intervention, for example "mortality", "weight" or "blood pressure".
These elements can be expressed differently in TS and GS, for example using synonyms and paraphrases. On the other hand, sometimes the GS contains a more specific or generic group of patients, drug, or outcome: "analgesics" instead of "Panadol". These are mistakes and thus the PICO should be considered **different**.

Modality

Modality shows how sure the author is of the conclusions:

Strong claim: these claims are modified by such strengthening expressions as *remarkably* or *considerably*, but the author can also directly claim that the evidence is very reliable (*high-quality evidence)*.
Moderate claim: these statements usually do not have any modifying adverbs: *improves, increases, reduces* etc.
Weak claim: these statements can have such modal verbs as *may*, phrases such as *has potential to*, or adverbs such as *likely*. The author can also refer to lower quality of evidence or other limitations (*initial evidence, variation in quality*, *only short-term effect*).
No evidence: the summary says that there is not enough reliable evidence to make any conclusions: *insufficient evidence*, *no evidence to support or refute.*
No claim: these summaries can mention PICO, but they make no statements regarding the effect of the intervention on the outcome: *This review examines the effect of fish oil on eye health.*

Direction

If the TS or GS (or both) have *no evidence* or *no claim* modality (which means that they do not contain a conclusion), it is impossible to determine if their direction is the same or different. In such cases, please mark the direction as **N/A**.
For strong, moderate and weak claims, the described effect of the intervention on the outcome can be **positive** (the one we desired, for example, *increase the life expectancy, reduce the mortality*), **negative** (undesired outcome: *increase the risk, reduce the treatment effectiveness*), or the intervention can have **no**

Figure 3: Annotation instructions.

effect on the outcome (*fish oil has neither benefits nor harms, there was no statistically significant effect of the supplement on weight loss, Panadol is as effective as ibuprofen).*
As the direction can be expressed in different ways in TS and GS, pay attention to their sentiment rather than specific words; for example, *improve blood pressure* can mean the same for the patient with hypertension as *reduce blood pressure*, so the direction in this case is **same**.

Grammar mistakes

This and the following categories should be judged only for the GS. Do not spend time looking for grammar, spelling and repetition mistakes in the TS.

In addition to usual grammar mistakes, this category includes incorrect word choice errors, for example, when "the" is used instead of "there": "the is no evidence". Please do not pay attention to punctuation and capitalization (lower/upper case) mistakes.

Spelling mistakes

This category is for spelling mistakes, such as when two words are incorrectly merged together by a model. If you are unsure about some medical terms' spelling, please do a Google search.

No repetitions

Neural models sometimes produce the same output twice (*effects of Panadol and Panadol),* such cases should be marked as a repetition mistake. Please note, that if repetitions result in a grammar error, they should be marked as grammar mistakes.

Dealing with different degrees of content mismatch

Intuitively, it might feel that a summary which is very different in content from the target review is "worse" than the one different only in some points. Please note that our aim is not to judge the output on some scale; instead, we need to make binary decisions regarding the criteria outlined above and see which categories are problematic for the models. In addition to this, sometimes it can be hard to judge some aspects of the summary if it is very different from what is expected. Please note that you should judge all categories independent from each other, so that if, for example, PICO is completely wrong, you still can determine if the direction and modality are the same:

A. Panadol helps to reduce headache.
B. Ibuprofen might improve the muscle pain in female athletes.

Imagine that you need to compare these (highly unlikely) summaries, which are very different in P, I and O. You should be still able to determine that their modality is different (moderate in A vs. weak in B), but the direction of their findings in the same (positive).

Figure 4: Annotation instructions (cont.).

Target (correct) summary	**Generated summary**
Omega-3 supplementation during pregnancy does not reduce the incidence of preterm birth or improve neonatal outcome.	The : omega-3 supplementation reduces the incidence of preterm birth in women with a history of preterm birth.

Please select the answer for each category by clicking a checkbox. The direction should be marked as N/A if any of TS or GS do not contain a conclusion (*no evidence* or *no claim).*

Is PICO the same?	Same ☐	Different ☒		
Is modality the same?	Same ☒	Different ☐		
Is direction the same?	Same ☐	Different ☒	N/A ☐	
Is the grammar correct?	Correct ☐	Not correct ☒		
Is the spelling correct?	Correct ☒	Not correct ☐		
Are there no repetitions?	No repetitions ☒	Repetitions ☐		

Figure 5: One of summaries provided for annotation.

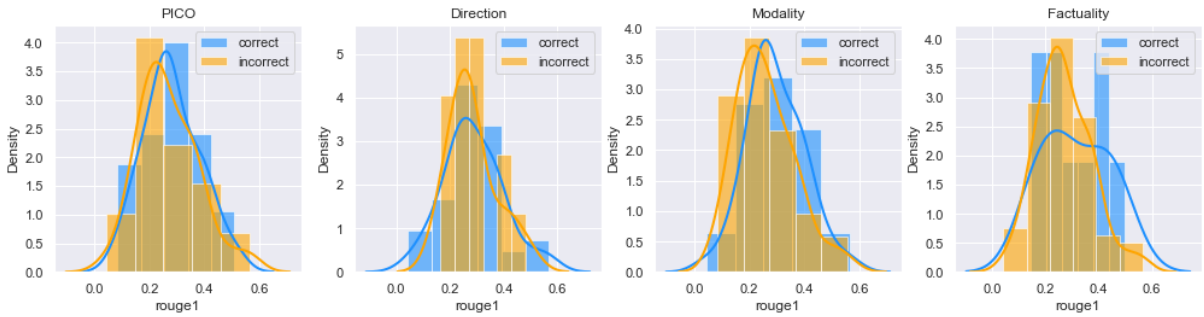

Figure 6: Distribution of ROUGE-1 scores for correct and incorrect summaries in different categories.