# LED down the rabbit hole: exploring the potential of global attention for biomedical multi-document summarisation

**Yulia Otmakhova**[1,*], **Hung Thinh Truong**[1,*], **Timothy Baldwin**[1,3],
**Trevor Cohn**[1], **Karin Verspoor**[2,1], **Jey Han Lau**[1]
[1]The University of Melbourne, [2]RMIT University, [3]MBZUAI
{yotmakhova,hungthinht}@student.unimelb.edu.au, tb@ldwin.net,
trevor.cohn@unimelb.edu.au, karin.verspoor@rmit.edu.au, jeyhan.lau@gmail.com

## Abstract

In this paper we report on our submission to the Multidocument Summarisation for Literature Review (MSLR) shared task. Specifically, we adapt PRIMERA (Xiao et al., 2022) to the biomedical domain by placing global attention on important biomedical entities in several ways. We analyse the outputs of the 23 resulting models, and report patterns in the results related to the presence of additional global attention, number of training steps, and the input configuration.

## 1 Introduction

In this paper we describe our experiments and results on the Multidocument Summarisation for Literature Review (MSLR) shared task.[1] In particular, we attempt to improve on previous multi-document summarisation models in the biomedical domain, which have tried to integrate domain knowledge by marking important biomedical entities (Wallace et al., 2021; DeYoung et al., 2021). We hypothesise that highlighting such entities by placing global attention on them will enable better aggregation and normalisation of related entities across documents, and thus improve the factuality of the generated summaries. To explore this idea, we experiment with four different ways of modifying the global attention mechanism of PRIMERA (Xiao et al., 2022), a recent state-of-the-art model designed for multi-document summarisation (MDS). In particular, while by default the global attention tokens in Primera are used to separate documents in the input and capture their relationships, we assign global attention to important biomedical entities in input documents to create links between them. Moreover, to examine the effect of content selection on the quality of summaries produced by this underlying model, we compare results where we use the whole abstract as input vs. only the concluding sentences (which we expect to be more informative). We train and analyse models in zero-shot, few-shot (10 and 100), as well as fully fine-tuned scenarios. Overall we evaluate (using both automatic metrics and human evaluation) a total of 23 models, two of which formed our official submissions to the leaderboard.[2] Both submitted models substantially outperform the baseline approaches (DeYoung et al., 2021) in terms of automatic metrics, and one achieves the best performance in terms of BERTScore and ROUGE-2 among all submissions. Overall, our contributions in comparison to the previously published domain-specific models for MDS are the following:

- We explore the potential of using global attention as a means to highlight important biomedical entities, in order to improve aggregation across input documents.
- We examine how the amount of training data influences the quality of generated summaries, and propose several scenarios where the performance of few-shot and even zero-shot models is on par with that of fully fine-tuned ones.
- We show that in the fine-tuned scenario, the model is able to select important content without additional marking.

## 2 Dataset

We use the Cochrane dataset as provided in the shared task without any additional data. See Table 5 in Appendix A for dataset statistics.

### 2.1 Pre-processing

As the trials are collected automatically from the Cochrane library, they contain redundant metadata

[*]Equal contribution
[1]https://github.com/allenai/mslr-shared-task
[2]Additional results and code for all models is provided at https://github.com/joey234/PRIMER-pico-attn.



*Proceedings of the Third Workshop on Scholarly Document Processing*, pages 181–187
October 12–17, 2022.

such as hyperlinks, trial identifiers, funding information, copyright statements, and publication records. We perform string matching using regular expressions to remove this content. Following Wallace et al. (2021), for each review, we concatenate all corresponding documents and add a separator token to denote the end of each document.

### 2.2 Entity marking

The PICO framework describes several essential components of the central question in a clinical trial, including Populations (e.g. *diabetics*), Interventions (e.g. *animal insulin*), Comparators (e.g. *human insulin*), and Outcomes (e.g. *glycaemic control*) (Huang et al., 2006). We tag PICO spans in input and target documents to make the summarisation models explicitly attend to them. We train a tagger on the EBM-NLP dataset (Nye et al., 2018), which contains annotations for the P, I, and O classes[3] on abstracts of randomized controlled trials. Using this dataset, we fine-tune the BioLinkBERT model (Yasunaga et al., 2022), a BERT variant that leverages links between documents that achieve state-of-the-art results on various biomedical NLP tasks, including the PI(C)O tagging task. We adopt the same hyperparameters as in Yasunaga et al. (2022) using the $\text{BioLinkBERT}_{\text{base}}$ model, and achieve 74.06 macro-$F_1$ score on the EBM-NLP test set, which is comparable to the reported results in Yasunaga et al. (2022). We run the trained PIO tagger on the Cochrane dataset for both the documents and summaries. For simplicity, we only use two new special tokens ***<ent>*** and ***</ent>*** to mark the beginning and the end of each PICO span (e.g. ***<ent>*** *Magnesium sulfate* ***</ent>*** *does not have a major impact on disease progression in* ***<ent>*** *women with mild preeclampsia* ***</ent>***.).

Table 5 presents basic statistics of the Cochrane dataset used in this challenge. The average number of PIO spans in the summary and input documents is based on the output of the trained PIO tagger. Note that target summaries for the test set are not provided to participants.

## 3 Evaluation

For the automatic evaluation, in addition to ROUGE scores (Lin, 2004) and BERTScore[4] (Zhang et al., 2019), we report the metrics introduced in DeYoung et al. (2021), namely $\Delta$EI which measures the distance in predicted direction of the conclusions (*increases*, *decreases*, or *no change*) in the target and generated summaries. For this metric, we report the average distance across samples and also macro-F1 score, in which the predicted direction for the target summary is treated as the correct label ($\Delta$EI-$F_1$).

To estimate quality of the generated summaries, especially in terms of their factuality, we also perform human evaluation, for which we adopt the binary decision method proposed in Otmakhova et al. (2022). As we need to assess results from a large number of models, we simplify the evaluation, focusing only on factual errors and collapsing the categories of *modality* and *polarity* into a single category with five potential values (*positive*, *negative*, *no effect*, *no evidence*, *no claim*), similar to how it was done by DeYoung et al. (2021). Thus, we report if **PICO** elements used in the correct and generated summaries are aligned, if the **direction** of the findings is the same, and if the summaries are **factual**, that is, correct in these two aspects. In addition, to analyse common errors, we annotate generations as **contradictory** (i.e. containing statements with the same set of PICO elements but different polarity), **malformed** (i.e. including lexical and grammatical errors or repetitions), and **not evidential** (i.e. claiming that there is not enough evidence to determine the effect of intervention). We list some examples of contradictory, malformed and non-evidential summaries in Appendix B.

As the vast majority of the target summaries were multi-aspect — that is, contained statements regarding several groups of patients, interventions or outcomes — one of the difficulties we experienced during the evaluation was comparing them to generated summaries which were either single-aspect or contained different sets of PICO elements. We adopted a precision-based approach when evaluating such pairs of summaries: while it is not necessary for the generated summary to contain all PICO elements included in the target to be considered correct, it must not include any extra PICO elements. In the case of extra PICO elements in the generated summaries, we compared them against the *Objectives* section of the review's abstract to determine if they were truly erroneous or if the target conclusion underreported some of the elements. Moreover, in the case of multi-aspect summaries

[3] Comparators are grouped with Interventions in the dataset due to the difficulty in distinguishing them.

[4] Hash code: `roberta-large_L17_no-idf_version=0.3.11(hug_trans=3.1.0)`

| Setting | Description |
|---|---|
| DocSep | The global attention is only set on the document separation token (***<doc-sep>***) as in the original PRIMERA model. The attention on ***<doc-sep>*** is used across the board in all settings described below. |
| EntMarkers | In addition to the ***<doc-sep>*** global attention, we set global attention on tokens which mark the beginning and end of entities (i.e. ***<ent>***, ***</ent>***). |
| EntMarkersSpans | In addition to the ***<ent>*** and ***</ent>*** tags, global attention is set on the tokens between them, that is, the entities themselves. |
| EntSpans | We only assign global attention to the entity spans. The ***<ent>*** and ***</ent>*** tokens are replaced by the padding mask token to mask them in inputs and thus do not get either global or local attention. |
| EntOnly | We additionally mask out all tokens outside the entity spans so they do not get either global or local attention; thus we only pass entities with global attention on them to the decoder. We test this scenario to see how well the summaries can be recovered from only the essential entities plus information collected by ***<doc-sep>*** tokens. |

Table 1: Global attention settings

we consider direction to be correct only if it is correct for the corresponding set of PICO elements.

Thus though our evaluation approach is less detailed than the one proposed in Otmakhova et al. (2022), it is more strict in terms of alignment of multi-aspect summaries.

## 4 Experiments

### 4.1 Model

We base our experiments on PRIMERA (Xiao et al., 2022), which was designed for multi-document summarisation, and experiment with zero-, 10-, 100-shot, and fine-tuning scenarios with the same hyperparameters reported by the authors of the paper. We use the same random seed for all models to ensure consistency. For the baseline model (*No entity*) we use documents and summaries without any entity marking; all other models use documents with entity tags.

### 4.2 Entity marking and global attention

PRIMERA is based on Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), which uses sparse attention (global attention) in addition to fixed-sized window attention (local attention). Here, we experiment with employing the global attention mechanism to highlight PICO elements and aggregate them across the documents. Specifically, for the scenario with entity spans in input and target texts, we use the five settings for global attention listed in Table 1.

### 4.3 Manipulating inputs

As dealing with lengthy inputs is a well-known issue for multi-document summarisation, especially in scientific and biomedical domains, we experiment with several settings to control the length of individual input documents:

- *Default*: The default PRIMERA setting where LED's token budget of 4096 tokens is distributed evenly across all input documents and they are truncated to the corresponding length.
- *Last 3*: In the biomedical domain the most important information appears in conclusions at the end of the paper, so we include only the last three sentences, based on NLTK's sentence tokenizer.[5]

## 5 Results

Tables 2 and 3 report the results of automatic and human evaluation, correspondingly.

### 5.1 Models with and without global attention on entities

Though we do not see major improvements in ROUGE scores between the model without PICO entity marking (*No entity*) and the models with global attention on PICO entities (with the exception of *EntMarkers* and *EntSpans*) and even observe some decrease in factuality scores, on closer inspection the summaries generated by those systems prove to be qualitatively different. In particular, the *No entity* model is more extractive and more extensively copies the input studies, while the results of models with global attention on entities are more abstractive. For example, for review CD005963 (Table 7 in Appendix C), the *No entity* model copies the term *Mental Health Act* often mentioned in source documents but absent in target conclusions, while the other models do not.

Table 8 in Appendix D shows how the overlap with source documents decreases when the entity marking with global attention is used, thus making the summaries more abstractive. This, however comes at a cost: we notice that the models

[5]https://github.com/nltk/nltk

| | | R-1↑ | R-2↑ | R-L↑ | BERTScore↑ | ΔEI↓ | ΔEI-$F_1$ ↓ |
|---|---|---|---|---|---|---|---|
| ***Zero*** | Default | 0.215 | 0.032 | 0.132 | 0.834 | 0.580 | 0.321 |
| | Last 3 | 0.245 | 0.063 | 0.179 | 0.871 | 0.260 | 0.385 |
| ***10-shot*** | No entity | 0.229 | 0.037 | 0.147 | 0.857 | 0.269 | 0.328 |
| | DocSep | 0.234 | 0.041 | 0.155 | 0.864 | 0.267 | 0.367 |
| | EntOnly | 0.197 | 0.024 | 0.139 | 0.834 | 0.297 | 0.330 |
| | EntMarkers | 0.208 | 0.035 | 0.143 | 0.859 | 0.286 | 0.327 |
| | EntSpans | 0.235 | 0.036 | 0.155 | 0.854 | 0.307 | **0.295** |
| | EntMarkersSpans | 0.187 | 0.266 | 0.122 | 0.831 | 0.322 | 0.319 |
| ***100-shot*** | No entity | **0.259** | 0.052 | 0.171 | 0.864 | 0.302 | 0.376 |
| | DocSep | 0.251 | 0.048 | 0.164 | 0.862 | 0.339 | 0.452 |
| | EntOnly | 0.237 | 0.038 | 0.157 | 0.851 | 0.308 | 0.389 |
| | EntMarkers | 0.244 | 0.048 | 0.164 | 0.864 | 0.284 | 0.369 |
| | EntSpans | **0.259** | 0.049 | 0.170 | 0.863 | 0.273 | 0.314 |
| | EntMarkersSpans | 0.251 | 0.048 | 0.166 | 0.863 | 0.301 | 0.315 |
| ***Full*** | No entity | 0.256 | 0.064 | **0.182** | 0.871 | 0.308 | 0.409 |
| | DocSep | 0.234 | 0.060 | 0.170 | 0.869 | 0.337 | 0.373 |
| | EntOnly | 0.236 | 0.060 | 0.174 | 0.872 | 0.256 | 0.310 |
| | EntMarkers | 0.244 | **0.066** | 0.179 | **0.874** | 0.246 | 0.312 |
| | EntSpans | 0.237 | 0.061 | 0.174 | **0.874** | 0.251 | 0.302 |
| | EntMarkersSpans | 0.230 | 0.059 | 0.168 | 0.873 | **0.244** | 0.321 |

Table 2: Results of automatic evaluation; ↑: higher is better, ↓: lower is better

| | | PICO↑ | Direction↑ | Factual↑ | Contradict.↓ | Malformed↓ | No evid.↓ |
|---|---|---|---|---|---|---|---|
| ***Zero*** | Default | 50 | 15 | 5 | 0 | 0 | 0 |
| | Last 3 | 50 | 50 | 30 | 0 | 5 | 70 |
| ***10-shot*** | No entity | 25 | 45 | 10 | 5 | 30 | 100 |
| | DocSep | 25 | 50 | 10 | 15 | 20 | 95 |
| | EntOnly | 10 | 30 | 0 | 10 | 75 | 35 |
| | EntMarkers | 25 | 50 | 15 | 0 | 0 | 70 |
| | Ent Spans | 30 | 35 | 5 | 5 | 30 | 65 |
| | EntMarkersSpans | 20 | 35 | 10 | 5 | 70 | 40 |
| ***100-shot*** | No entity | 50 | 50 | 20 | 5 | 5 | 60 |
| | DocSep | 50 | 50 | 20 | 10 | 15 | 65 |
| | EntOnly | 45 | 35 | 5 | 5 | 35 | 45 |
| | EntMarkers | 50 | 45 | 30 | 25 | 25 | 85 |
| | EntSpans | 35 | 40 | 15 | 20 | 10 | 100 |
| | EntMarkersSpans | 60 | 40 | 25 | 0 | 0 | 75 |
| ***Full*** | No entity | 50 | 60 | 35 | 10 | 10 | 35 |
| | DocSep | 50 | 50 | 25 | 5 | 10 | 65 |
| | EntOnly | 30 | 40 | 20 | 0 | 5 | 85 |
| | EntMarkers | 35 | 40 | 20 | 10 | 0 | 90 |
| | EntSpans | 55 | 40 | 25 | 5 | 5 | 90 |
| | EntMarkersSpans | 50 | 40 | 25 | 5 | 0 | 100 |

Table 3: Results of human evaluation; ↑: higher is better, ↓: lower is better. ***Zero*** denotes the zero-shot setting.

with additional global attention produce remarkably more *no evidence* summaries, and in the fully fine-tuned scenario the number of such summaries grows with the number of tokens on which we place global attention. This is consistent with the results of another model which extensively uses global attention (DeYoung et al., 2021) which also produces a large number of *no evidence* summaries (Otmakhova et al., 2022). Another behaviour of models with extra global attention observed both in DeYoung et al. (2021) and here is that they generate sequences which are representative of biomedical text style. For example, in addition to conclusions, the summaries generated by such models contain generic sentences such as *There is a need for more studies of high methodological quality*. Thus we hypothesise that tokens with global attention tend to accumulate and reproduce information common to a large number of documents in the training set rather than information shared by a particular set of input documents. Finally, though we expected the *EntOnly* model, which only uses only PIO enti-

| | | R-1↑ | R-2↑ | R-L↑ | BERTScore↑ | ΔEI↓ | ΔEI-$F_1$ ↓ |
|---|---|---|---|---|---|---|---|
| ***Default*** | Zero-shot | 0.215 | 0.032 | 0.132 | 0.834 | 0.580 | **0.321** |
| | 10-shot | 0.229 | 0.037 | 0.147 | 0.857 | 0.269 | 0.328 |
| | 100-shot | **0.259** | 0.052 | 0.171 | 0.864 | 0.302 | 0.376 |
| | Full | 0.256 | **0.064** | **0.182** | 0.871 | 0.308 | 0.409 |
| ***Last 3*** | Zero-shot | 0.245 | 0.063 | 0.179 | **0.871** | **0.260** | 0.385 |
| | 10-shot | 0.211 | 0.030 | 0.135 | 0.853 | 0.289 | 0.342 |
| | 100-shot | 0.250 | 0.046 | 0.164 | 0.862 | 0.341 | 0.424 |
| | Full | 0.239 | 0.061 | 0.171 | 0.870 | 0.279 | 0.382 |

Table 4: Results of automatic evaluation; ↑: higher is better, ↓: lower is better

ties as inputs and thus loses information about the relations between them, to perform much worse than the other models, it is very similar to them both in automatic metrics and *Direction* scores. We maintain that it shows that even if the models are able to attend to all tokens, they only reproduce PIO entities and are not able to consistently capture the relationships between them.

### 5.2 Zero-shot vs. few-shot vs. fully fine-tuned models

We notice that in terms of automatic metrics, zero-shot models are comparable to fine-tuned ones or even outperform them; however they perform substantially worse in terms of factuality, especially for the direction. We find that in zero-shot scenarios, PRIMERA copies spans of text from one or several of the input documents, focusing mostly on their beginnings, rather than aggregates information across documents. Thus it outputs either conclusions copied from a single document, or, more often, makes no claims at all by reporting the objectives of the review or its setup.

Another interesting finding is that the ROUGE scores tend to be the highest in the 100-shot scenario and go down for the fully fine-tuned models. We maintain that in 10-shot scenarios the models are still unable to correctly capture and reproduce important entities (which is also reflected in their low accuracy in terms of PICO), while in the fully fine-tuned models, there is a tendency to generate broader and generic entities, for example *metal-protein attenuation compounds* instead of *PBT1/PBT2* in the target summary.

Not surprisingly, the number of malformed generations decreases with increasing the number of training samples: the majority of summaries produced by *EntOnly* and *EntMarkersSpans* after 10 shots are malformed, but even 100-shot training significantly reduces this amount. On the other hand, it is surprising to see that the more the models are fine-tuned the more *no evidence* statements they produce, with some models generating only such summaries in fully fine-tuned scenario.

Lastly, we find that the 100-shot *EntMarkers* model is similar in terms of factuality to the fully fine-tuned model without entity marking (*No entity*). This is an encouraging result as high-quality multi-document summarisation data is scarce in biomedical domain, so few-shot learning is a practically important direction to explore.

### 5.3 *Default* vs. *Last3*

For few-shot and fine-tuned models we find no major improvements in quality when restricting the inputs to the last three sentences only (Table 4). This shows that after fine-tuning PRIMERA is able to detect most useful spans without relying on their explicit marking. On the other hand, for the zero-shot scenario, where the model tends to copy from the beginning of input documents, the quality dramatically improves when we force it to extract only from a more informative span at the end of documents. Interestingly, such an easy manipulation of inputs allows to achieve results comparable to the best 100-shot and fully fine-tuned models without any training on the in-domain dataset. Again, this is a promising direction for research considering the scarcity of high-quality data.

## 6 Conclusion

We tackle the problem of biomedical multi-document summarisation by incorporating PICO information into a strong summarisation model, and using global attention to enhance the representation of this information. Through automatic and human evaluations on an extensive set of experiments, we find that adding global attention to PICO spans would help in (1) generating more abstractive summaries, and (2) improving summarization quality in few-shot settings, which is especially important in the biomedical domain.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their comprehensive and constructive reviews. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. This research was conducted by the Australian Research Council Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030) and funded by the Australian Government.

## References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of PICO as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359. American Medical Informatics Association.

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.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron 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, Melbourne, Australia. Association for Computational Linguistics.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. 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.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

## A Dataset statistics

Table 5 reports some basic statistics of the Cochrane dataset used in this challenge. The Average number of PICO spans in the summary and the input documents (Avg. # PICO spans) are obtained using the trained PICO tagger. Note that target summaries for the test set are not provided.

| | Train | Valid. | Test |
|---|---|---|---|
| # samples | 3752 | 470 | 470 |
| Avg. input length | 2417 | 2389 | 2677 |
| Avg. summary length | 68 | 70 | n/a |
| Avg. # PICO spans in input | 213 | 209 | 236 |
| Avg. # PICO spans in summary | 4 | 4 | n/a |

Table 5: Cochrane dataset statistics.

## B Examples of malformed, contradictory and non-evidential summaries

To clarify the criteria we used for evaluation, Table 6 lists some examples of contradictory, malformed and non-evidential summaries. Malformed summaries are ones containing repetitions, incomplete text or corrupted tokens. The spans of text corresponding to errors are in **bold**.

## C Examples of generated summaries

Table 7 shows the examples of summaries generated for input documents for review CD005963.

| Error | Summary |
|---|---|
| Contradiction | *There is **insufficient evidence** to support the use of edaravone as a therapy for acute ischemic stroke. However, it may be useful for treating other types of ischemic stroke. The current review **provides a rationale basis** for the use of edaravone as a therapy for acute ischemic stroke.* ***In the absence of evidence to support** the use of PBT2 in patients with severe Alzheimer's disease, clinicians and patients **should recommend** the continued use of PBT2.* |
| Malformed | *There is inadequate evidence to evaluate the effect of percutaneous endoscopic gastrostomy on the incidence of percutaneous wound infections. The current evidence base is limited due to the differing methodologies employed in the trials. **The current evidence base is limited due to the differing methodologies employed in the trials. The current evidence base is limited due to the differing methodologies employed in the trials...*** *We found no clear evidence to support the use **of** There is limited evidence to suggest that the use of **apleuapleuapleuapleuapleuapleua...*** |
| No evidence | *There is insufficient evidence to support the use of metal-protein-attenuating compounds for the treatment of AD. Further trials are needed.* |

Table 6: Examples of contradictory, malformed and non-evidential summaries

| Setting | Summary |
|---|---|
| No entity | *... the results suggest that advance directives may be beneficial in reducing the number of people admitted to hospital under the Mental Health Act.* |
| DocSep | *There is insufficient evidence to support or refute the use of advance directives for people with mental illnesses.* |
| EntMarkers | *There is insufficient evidence to support or refute the use of advance directives for people with severe mental illness.* |
| EntMarkersSpans | *There is insufficient evidence to support the use of advance directives for people with severe mental illness.* |
| EntSpans | *There is insufficient evidence to support the use of advance directives for people with mental illness.* |
| EntOnly | *There is insufficient evidence to support the use of advance directives for people with severe mental illness.* |

Table 7: Examples of generated summaries

## D Lexical overlap with the input documents

Table 8 shows the amount of lexical overlap with the source documents in terms of ROUGE scores. The lower the score is, the less is copied from the source and the more abstractive the summary is.

| | | R-1↓ | R-2↓ | R-L↓ |
|---|---|---|---|---|
| ***Full*** | No entity | 0.052 | 0.022 | 0.040 |
| | DocSep | 0.042 | 0.019 | 0.034 |
| | EntOnly | 0.043 | 0.021 | 0.036 |
| | EntMarkers | 0.042 | 0.018 | 0.033 |
| | EntSpans | 0.040 | 0.017 | 0.032 |
| | EntMarkersSpans | 0.037 | 0.016 | 0.030 |

Table 8: Token overlap with the source as a measure of extractiveness; lower = more abstractive