Towards objectively evaluating the quality of generated medical summaries

Francesco Moramarco[*]
Babylon Health / London, UK
University of Aberdeen / Aberdeen, UK
francesco.moramarco[†]

Damir Juric[*]
Babylon Health / London, UK
damir.juric[†]

Aleksandar Savkov
Babylon Health / London, UK
sasho.savkov[†]

Ehud Reiter
University of Aberdeen / Aberdeen, UK
e.reiter@abdn.ac.uk

[†]@babylonhealth.com

Abstract

We propose a method for evaluating the quality of generated text by asking evaluators to count facts, and computing precision, recall, f-score, and accuracy from the raw counts. We believe this approach leads to a more objective and easier to reproduce evaluation. We apply this to the task of medical report summarisation, where measuring objective quality and accuracy is of paramount importance.

1 Introduction

Natural Language Generation in the medical domain is notoriously hard because of the sensitivity of the content and the potential harm of hallucinations and inaccurate statements (Kryscinski et al., 2020; Falke et al., 2019). This informs the human evaluation of NLG systems, selecting accuracy and overall quality of the generated text as the most valuable aspects to be evaluated.

In this paper we carry out a human evaluation of the quality of medical summaries of Clinical Reports generated by state of the art (SOTA) text summarisation models.

Our contributions are: (i) a re-purposed parallel dataset of medical reports and summary descriptions for training and evaluating, (ii) an approach for a more objective human evaluation using counts, and (iii) a human evaluation conducted on this dataset using the approach proposed.

2 Related Work

A recent study by Celikyilmaz et al. (2020) gives a comprehensive view on different approaches to text summary evaluation. While many of these can be wholly or partly translated between different domains, the medical domain remains particularly problematic due to the sensitive nature of its data. Moen et al. (2014) and Moen et al. (2016) try to establish if there is a correlation between automatic and human evaluations of clinical summaries. A 4-point and 2-point Likert scales are used for the human evaluation. In Goldstein et al. (2017) the authors generate free-text summary letters from the data of 31 different patients and compare them to the respective original physician-composed discharge letters, measuring relative completeness, quantifying missed data items, readability, and functional performance.

Closest to our approach is the Pyramid method by Nenkova et al. (2007), which defines semantically motivated, sub-sentential units (Summary Content Units) for annotators to extract in each reference summary. SCUs are weighed according to how often they appear in the multiple references and then compared with the SCUs extracted in the hypothesis to compute precision, recall, and f-score.

3 Data

The *MTSamples* dataset comprises 5,000 sample medical transcription reports from a wide variety of specialities uploaded to a community platform website[1]. The dataset has been used in past medical NLP research (Chen et al., 2011; Lewis et al., 2011; Soysal et al., 2017) including as a Kaggle dataset[2].

There are 40 medical specialties in the dataset, such as 'Surgery', 'Consult - History and Phy.', and 'Cardiovascular / Pulmonary'. Each specialty

[*]Equal contribution

[1]https://mtsamples.com

[2]https://www.kaggle.com/tboyle10/medicaltranscriptions

Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 56–61
Online, April 19, 2021.

contains a number of sample reports ranging from 6 to 1103.

The reports are free text with headings, which change according to the specialty. However, all reports also have a description field, which is a good approximation of a summary of the report. The length of each report varies greatly according to the specialty, with an average of 589 words for the body of the report, and 21 words for the description. Figure 1 shows an example of MTSamples reports, inclusive of description.

Medical Specialty: Diets and Nutritions
Sample Name: Dietary Consult - 2

SUBJECTIVE: The patient's assistant brings in her food diary sheets. The patient says she stays active by walking at the mall.

OBJECTIVE: Weight today is 201 pounds, which is down 3 pounds in the past month. She has lost a total of 24 pounds. I praised this and encouraged her to continue. I went over her food diary. I praised her three-meal pattern and all of her positive food choices, especially the use of sugar-free Kool-Aid, sugar-free Jell-O, sugar-free lemonade, diet pop, as well as the variety of foods she is using in her three-meal pattern. I encouraged her to continue all of this.

ASSESSMENT: The patient has been successful with weight loss due to assistance from others in keeping a food diary, picking lower-calorie items, her three-meal pattern, getting a balanced diet, and all her physical activity. She needs to continue all this.

PLAN: Followup is set for 06/13/05 to check the patient's weight, her food diary, and answer any questions.

DESCRIPTION: The patient has been successful with weight loss due to assistance from others in keeping a food diary, picking lower-calorie items, her three-meal pattern, getting a balanced diet, and all her physical activity.

Figure 1: An MTSamples clinical report of specialty 'Diets and Nutritions'. Note the reference Description at the bottom.

Given the brevity of some descriptions, we discard reports with descriptions shorter than 12 words and consider a dataset of 3242 reports. By examining the dataset, we note that descriptions are mostly extractive in nature, meaning they are phrases or entire sentences taken from the report. To quantify this we compute n-gram overlap with Rouge-1 (unigram) and Rouge-L (longest common n-gram) (Lin, 2004) precision scores, which are 0.989 and 0.939 respectively.

We split the dataset into 2 576 reports for training (80%), 323 for development (10%) and 343 for testing (10%). We perform the split separately for each medical specialty to ensure they are adequately represented and then aggregate the data.

The dataset, models, and evaluation results can be found on Github[3].

4 Experimental Setup

For our experiment, we consider one baseline and three SOTA automatic summarisation models (extractive, abstractive, and fine-tuned on our training set respectively). More specifically:

- **Lead–3** — this is our baseline. Following Zhang et al. (2018), this model selects the first three sentences of the clinical report as the description;
- **Bert–Ext** — the unsupervised extractive model by Miller (2019) [4];
- **Pegasus–CNN** — an abstractive model by Zhang et al. (2019) trained on the CNN/Daily mail dataset and used as is;
- **Bart–Med** — an abstractive model by Lewis et al. (2020), which we fine-tune on our MTSamples training set.

We generate descriptions with these 4 models using the entire clinical report text as input.

5 Human Evaluation Protocol

We select 10 clinical reports and summary descriptions from our MTSamples test set. Our subjects are three general practice physicians. They are employed at Babylon Health and have experience in AI research evaluation. The task is implemented with the Heartex Annotation Platform[5], which lets researchers define tasks in an XML language and specify the number of annotators. It then generates each individual task and collates the results.

The task involves (i) reading the clinical report, (ii) reading the reference description (supplied by the dataset, see Figure 1), (iii) then evaluating 4 generated descriptions by answering 5 questions (for a total of 40 generated descriptions). We ask the evaluators to count the "medical facts" in each generated description and to compare them against those in the reference. Initially, we considered listing the types of facts to be extracted, as done by Thomson and Reiter (2020), but the sheer diversity in the structure and content across the specialties

[3] https://github.com/babylonhealth/medical-note-summarisation

[4] https://pypi.org/project/bert-extractive-summarizer/

[5] https://www.heartex.ai/

in our dataset made this approach impractical. Instead, we give evaluators instructions containing two examples and ask them to extrapolate a process for fact extraction. Figure 2 shows the instructions we give them.

> The evaluation consists of reading a clinical report and a number of short descriptions, then quantifying how many "medical facts" were correctly reported. We understand that the definition of a "medical fact" is vague, and so it's up to your interpretation. As an example, in the following description:
>
> ```
> 2-year-old female who comes in for
> just rechecking her weight, her
> breathing status, and her diet.
> ```
>
> There are (arguably) 4 facts:
>
> - 2 year old female
> - coming to recheck her weight
> - coming to recheck her breathing status
> - coming to recheck her diet
>
> Here's another example:
>
> ```
> The patient had a syncopal episode
> last night. She did not have any
> residual deficit. She had a headache
> at that time. She denies chest pains
> or palpitations.
> ```
>
> Here there are (arguably) 5 facts:
>
> - had a syncopal episode last night
> - no residual deficit
> - headache
> - no chest pains
> - no palpitations

Figure 2: Instructions to evaluators.

The evaluators are asked to read the clinical report (as shown in Figure 1), then to analyse the reference description by reporting the number of facts. To aid them in the task, they can optionally select the facts in the text using an in-built Heartex feature. Next, they are shown four generated descriptions (one per model) and asked to count facts and answer 5 questions. Figure 3 shows the reference, generated descriptions, and questions for a given task, and gives an example annotation from one of the evaluators. When answering question 3 (How many facts in G are correct?) they refer to the clinical report as a ground truth.

Based on this set of questions, we gather the following raw counts:

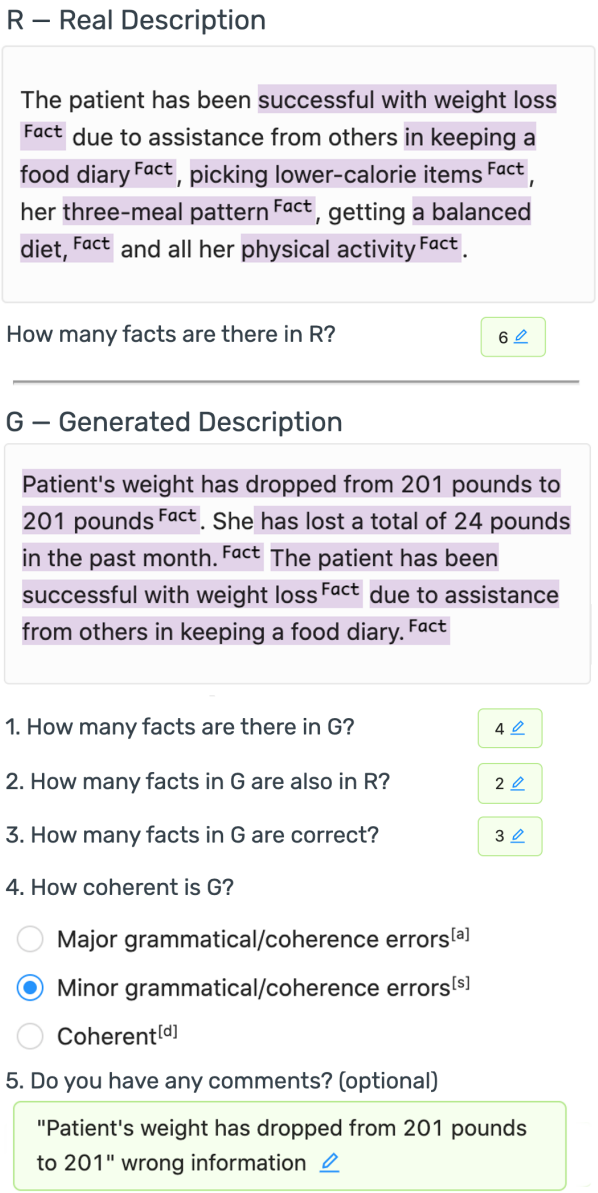

Figure 3: A completed task. *Real Description* is the reference.

- R: facts in the reference description
- G: facts in the generated description
- $R\&G$: facts in common
- C: correct facts in the generated description

We use these raw counts to compute four derived metrics:

- **Precision**, calculated as $\frac{R\&G}{G}$
- **Recall**, calculated as $\frac{R\&G}{R}$
- **F-Score**, calculated as $2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$
- **Accuracy**, calculated as $\frac{C}{G}$

For *Coherence*, we take Chen et al. (2020) and Juraska et al. (2019) definition: "*whether the generated text is grammatically correct and fluent, re-*

Model	Metric	Eval 1	Eval 2	Eval 3	Avg
Lead-3	Precision	0.42	0.43	0.46	0.44
	Recall	0.64	0.60	0.73	0.66
	F-Score	0.49	0.45	0.51	0.48
	Accuracy	1.0	1.0	1.0	**1.0**
	Coherence	0.95	0.95	0.90	0.93
Bert-Ext	Precision	0.58	0.48	0.48	0.51
	Recall	0.62	0.61	0.60	0.61
	F-Score	0.59	0.52	0.51	0.54
	Accuracy	1.0	1.0	1.0	**1.0**
	Coherence	1.0	0.95	1.0	**0.98**
Pegasus-CNN	Precision	0.29	0.36	0.31	0.32
	Recall	0.43	0.50	0.50	0.47
	F-Score	0.34	0.40	0.36	0.37
	Accuracy	0.97	0.97	0.98	0.97
	Coherence	1.0	1.0	0.95	**0.98**
Bart-Med	Precision	0.65	0.58	0.55	**0.59**
	Recall	1.0	0.96	0.97	**0.98**
	F-score	0.77	0.70	0.68	**0.72**
	Accuracy	1.0	1.0	1.0	**1.0**
	Coherence	1.0	0.75	0.95	0.90

Table 1: Derived metrics for each model and each evaluator, aggregated across tasks.

gardless of factual correctness" and ask evaluators to choose between three options (Coherent, Minor Errors, Major Errors) and convert these into continuous numbers with Coherent = 1.0, Minor Errors = 0.5, and Major Errors = 0.0.

6 Results and Discussion

Table 1 shows the results for all derived metrics, calculated on the raw counts from the evaluators. Expectedly, Bart-Med, the model trained on the MTSamples training set, scores highest in all metrics (except Coherence).

Interestingly, all four models score almost-perfect accuracy, meaning they don't hallucinate medical facts. This is not a surprise for Lead-3 and Bert-Ext, which are extractive in nature. As for Pegasus-CNN and Bart-Med, while the models are abstractive, we notice they tend to mostly select and copy phrases or entire sentences from the source report. The only hallucination the evaluators found is a numerical error, reported by Pegasus-CNN in the following generated description:

```
Patient's weight has dropped from 201
pounds to 201 pounds. She has lost a
total of 24 pounds in the past month.
```

Whereas, the source report states:

	Metric	E1–E2–E3	E1–E2	E1–E3	E2–E3
Raw Counts	R facts	0.25	0.44	0.27	0.01
	G facts	0.33	0.50	0.26	0.12
	G&R facts	0.55	0.74	0.50	0.40
	G acc facts	0.34	0.51	0.27	0.13
	Coherence	0.40	0.14	0.56	0.49
Der. Metrics	Precision	0.87	0.84	0.88	0.88
	Recall	0.90	0.93	0.89	0.89
	F-Score	0.89	0.88	0.91	0.87
	Accuracy	0.87	0.79	0.96	0.84

Table 2: Krippendorff Alpha for each metric, where *R* is reference, *G* the generated description, *G acc facts* the count of accurate facts in the generated description, *E1-E2-E3* the agreement of all three evaluators, and *Ex-Ey* the agreement between Evaluator x and Evaluator y.

```
Weight today is 201 pounds, which is
down 3 pounds in the past month. She
has lost a total of 24 pounds.
```

6.1 Agreement

To validate the human evaluation task, we compute inter-annotator agreement for each derived metric, as well as on the raw counts. We use Krippendorff Alpha (Hayes and Krippendorff, 2007) as we are dealing with continuous values. Table 2 includes overall agreement and a breakdown for each pair of evaluators.

Looking at the *E1-E2-E3* column, we note a clear divide between the low agreement on raw counts and the high agreement on the derived metrics. We investigate this by comparing the facts selected by each annotator and notice a degree of variability in the level of granularity they employed. Consider the description:

```
An 83-year-old diabetic female presents
today stating that she would like dia-
betic foot care.
```

Table 3 shows the facts selected by the three evaluators.

We compute pairwise agreement in Table 2 and notice that two of the evaluators (E1 and E2) share a similar (more granular) approach to fact selection, whereas E3 is less granular.

We also investigate the low agreement for Coherence and discover that it's due to a strong imbalance of the three classes (Coherent, Minor Errors, and Major Errors) where Coherent appears 91.67% of cases, Minor Errors 6.67% and Major Errors 1.67%. While this causes a low Krippendorff Alpha, we count the number of times all three

E	Count	Selected Facts
E1	2	- 83-year-old diabetic female - would like diabetic foot care
E2	5	- 83-year-old - diabetic - female - presents today - would like diabetic foot care
E3	3	- 83-year-old female - diabetic - would like diabetic foot care

Table 3: Example of evaluators disagreement in fact selection.

evaluators agree on a generated description being Coherent and find it to be 82.5%.

Finally, for all derived metrics the agreement scores are very high. This shows a robustness of these metrics even with different granularity in fact selection, and that the three evaluators agree on the quality of a given generated description. In other words, the evaluators agree on the quality of the generated descriptions even though they don't agree on the way of selecting medical facts.

7 Future Work

In this paper we presented an evaluation of the quality of medical summaries using fact counting. The results of this study help us to identify a number of insights to guide our future work:

- We could work on better defining a medical fact (as in Dušek and Kasner (2020)) and to prompt agreement on the level of granularity, for instance by instructing evaluators to split a description into the highest number of facts that are meaningful;
- Our evaluation focused on the quality of the generated descriptions and did not evaluate their usefulness in the medical setting. Such extrinsic evaluation would be very valuable;
- We could compare our approach of fact counting with the more common Likert scales.

References

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Elizabeth Chen, Sharad Manaktala, Indra Sarkar, and Genevieve Melton. 2011. A multi-site content analysis of social history information in clinical notes. *AMIA Annual Symposium Proceedings*, 2011:227–36.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137.

Tobias Falke, Leonardo FR 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.

Ayelet Goldstein, Yuval Shahar, Efrat Orenbuch, and Matan J Cohen. 2017. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. *Artificial intelligence in medicine*, 82:20–33.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Juraj Juraska, Kevin Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 164–172, Tokyo, Japan. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

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.

Neal Lewis, Daniel Gruhl, and Hui Yang. 2011. Extracting family history diagnosis from clinical texts. pages 128–133.

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

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Hans Moen, Juho Heimonen, Laura-Maria Murtola, Antti Airola, Tapio Pahikkala, Virpi Terävä, Riitta Danielsson-Ojala, Tapio Salakoski, and Sanna Salanterä. 2014. On evaluation of automatically generated clinical discharge summaries. In *PAHI*, pages 101–114.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25 – 37.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *TSLP*, 4.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Craig Alexander Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784.