NANAYAKKARA, G., WIRATUNGA, N., CORSAR, D., MARTIN, K. and WIJEKOON, A. 2023. Clinical dialogue transcription error correction with self-supervision. In *Bramer, M. and Stahl, F. (eds.) Artificial intelligence XL: proceedings of the 43rd SGAI international conference on artificial intelligence (AI-2023), 12-14 December 2023, Cambridge, UK*. Lecture notes in computer science, 14381. Cham: Springer [online], pages 33-46. Available from: https://doi.org/10.1007/978-3-031-47994-6_3

Clinical dialogue transcription error correction with self-supervision.

NANAYAKKARA, G., WIRATUNGA, N., CORSAR, D., MARTIN, K. and WIJEKOON, A.

2023

Clinical Dialogue Transcription Error Correction with Self-supervision

Gayani Nanayakkara(✉), Nirmalie Wiratunga, David Corsar, Kyle Martin, and Anjana Wijekoon

School of Computing, Robert Gordon University, Aberdeen, Scotland
{g.nanayakkara,n.wiratunga,d.corsar1,k.martin3,a.wijekoon1}@rgu.ac.uk

Abstract. A clinical dialogue is a conversation between a clinician and a patient to share medical information, which is critical in clinical decision-making. The reliance on manual note-taking is highly inefficient and leads to transcription errors when digitising notes. Speech-to-text applications designed using Automatic Speech Recognition (ASR) can potentially overcome these errors using post-ASR error correction. Pre-trained language models are increasingly used in this area. However, the performance suffers from the lack of domain-specific vocabulary and the mismatch between error correction and pre-training objectives. This research explores these challenges in gastrointestinal specialism by introducing self-supervision strategies to fine-tune pre-trained language models for clinical dialogue error correction. We show that our mask-filling objective specialised for the medical domain (med-mask-filling) outperforms the best performing commercial ASR system by 10.27%.

Keywords: Automatic speech recognition; Error correction; Language models

1 Introduction

In the traditional clinical setting, healthcare providers manually take notes during conversations and patient interactions. This involves physically writing down relevant information, observations, and essential details the patient shares. The process typically entails using pen and paper or a digital device to record the information. This manual note-taking process requires clinicians to quickly process and capture information while focusing on the patient's needs.

The main drawback to this approach is the time burden of record-keeping of clinical communications [14], and it is associated with clinician burnout, increased cognitive load, information loss, and distractions [17]. One of the most promising avenues of automating clinical documentation with digital scribes is to use Automatic Speech Recognition (ASR) [18], where the audio data is converted to textual data.

Given the critical nature of the medical field, ASR systems for clinical applications must demonstrate high performance levels. However, the effectiveness of ASR systems depends on three key factors: speaker variabilities, spoken language variabilities, and other mismatch factors [2]. These factors contribute to the occurrence of errors in the textual outputs. Therefore, it is crucial to explore strategies that can mitigate the likelihood of transcription errors.

In this work, we propose a post-ASR error correction method that uses the advancements in transformer-based pre-trained language models. Our work aims to leverage the strengths of pre-trained language models and adapt them to the clinical domain for error correction. Rather than designing new architectures or fine-tuning models on specialized datasets, we aim to use publicly available clinical domain data to fine-tune these models. For that, we introduce a newly curated PubMed[1] dataset to address the challenge of the lack of clinical dialogue data for fine-tuning language models. The dataset scraped from PubMed alleviates the need for large-scale real-world transcription data for self-supervision. Our method is evaluated using the Gastrointestinal Clinical Dialogue (GCD) Dataset, which is a role-playing dataset collected in partnership with the National Health Service (NHS) Grampian Inflammatory Bowel Dis-ease (IBD) Clinic which emulates a real-world clinical setting. Results from our self-supervision strategy applied to two pre-trained language models, T5-small and BART, demonstrate that it can reduce transcription errors compared to commercial ASR systems. Accordingly, our contributions are:

1. a self-supervision strategy to fine-tune pre-trained language models for clinical dialogue error correction;
2. novel masked and med-masked PubMed datasets to fine-tune pre-trained language models using self-supervision; and
3. an empirical evaluation that compares our method with commercial ASR systems.

The rest of the paper is organised as follows. Section 2 presents related work in the ASR error correction research domain. Our approach is presented in Sect. 3 followed by evaluation and results in Sect. 4. Section 5 concludes the paper with a review of contributions and an outline of future directions.

2 Related Work

The performance of an Automatic Speech Recognition (ASR) model is influenced by several factors: speaker variabilities, spoken language variabilities, and other mismatch factors [2]. Speaker variabilities encompass changes in voice due to ageing, illness, emotions, and tiredness. Spoken language variabilities arise from variations in speech patterns, accents, and dialects. Other mismatch factors include variations in communication channels and the devices used during speech recognition. These factors contribute to transcription errors, making it challenging to extract meaningful insights from the generated transcripts [2].

When recognising the importance of error correction, there are two primary approaches to address ASR errors: incorporating an error correction algorithm

[1] https://www.ncbi.nlm.nih.gov/pubmed/.

within the ASR model itself or applying post-processing techniques to refine the ASR outputs. In the past, researchers explored the integration of error correction methods within ASR models, utilizing techniques like Hidden Markov Models (HMMs) [4,6] and more recently, deep neural architectures [5]. These approaches aimed to enhance the accuracy of ASR outputs by directly correcting errors during the recognition process.

Alternatively, the post-ASR error correction approach has gained popularity. This method involves applying error correction techniques as a subsequent step to refine the ASR outputs. Initially, unsupervised methods were employed, such as lexical co-occurrence analysis on large ASR transcription corpora [20] and statistical error correction methods [1]. These methods aimed to identify and rectify errors based on linguistic patterns and statistical analysis. More recently, transformer-based [21] language models have emerged as a promising approach for post-ASR error correction. These models, known for their robust contextual understanding, have been leveraged to improve the accuracy of ASR outputs. By fine-tuning transformer-based language models on domain-specific data, they can learn to correct errors present in the ASR transcriptions more effectively.

There are two prominent approaches to leveraging transformer-based language models for post-ASR error correction. One approach is exemplified by FastCorrect [7,8], which introduces modifications to a transformer-based encoder-decoder architecture. This architecture incorporates an error correction module that utilizes the edit distance metric [22] to guide the error correction process [8] FastCorrect models are trained on large-scale datasets and subsequently fine-tuned specifically for error correction using extensive ASR datasets [8,10,24]. In these approach, the models undergo a training process where they learn to correct errors by considering the edit distance between the ASR-generated text and the ground truth text. The models are trained to minimize this edit distance, improving their error correction capabilities.

Alternatively, pre-trained language models can be effectively fine-tuned using self-supervision for error correction, with the self-supervision objective being Machine Translation [12,13,15]. This approach involves training the models to correct errors by treating the ASR output as a source language and the ground truth transcription as a target language for translation. By fine-tuning the models using self-supervised learning, they can learn to align and correct errors in the ASR-generated text. It is worth noting that the fine-tuning process in these methods often relies on a significant portion of the ASR transcription data, typically using it as a training set for self-supervision [13]. Consequently, these approaches are particularly effective when large quantities of ASR transcriptions from the target domain are readily available.

In this paper, our approach is based on post-ASR error correction utilizing transformer-based architectures. However, instead of adopting custom-designed architectures [8,10,24] or fine-tuning specifically for error correction using a large-scale dataset [12,13], we explore how to effectively fine-tune a pre-trained model using publicly available clinical domain data when the domain-specific data is limited.

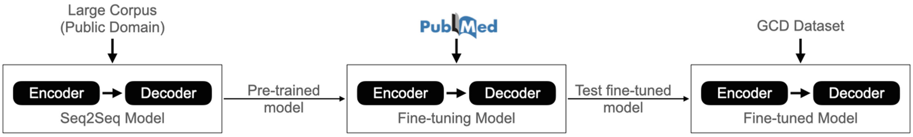

Fig. 1. Self-supervision for clinical dialogue error correction

3 Methodology

We view error correction as a seq2seq task performed using an Encoder-Decoder (ED) architecture-based language model to perform error correction, treating it as a sequence-to-sequence task. However, before this model can be effectively used for error correction, it needs to undergo a process of fine-tuning. This is necessary to address the following:

Vocabulary Gap the pre-trained language models are general-purpose and not initially tailored to handle domain-specific vocabulary (i.e., medical jargon and terms).

Objective Gap the general-purpose models are also not initially fine-tuned to perform specific downstream tasks (i.e., error correction).

To resolve these gaps, we introduce self-supervision strategies, which involve fine-tuning the pre-trained model on specific downstream datasets and tasks, specifically in the gastrointestinal domain.

3.1 Self-supervision

The approach of using the same unsupervised data to create multiple training objectives is known as self-supervision. When fine-tuning base language models, we need to create self-supervision tasks with the general structure of an input-output text pair (Fig. 1). A self-supervision dataset for fine-tuning a language model consists of input-output text pairs. And in this work, we looked at three approaches to forming a self-supervision strategy best suited for error correction: (i) standard objective approaches, (ii) standard hybrid approaches and (iii) domain-specific approaches.

3.2 Standard Objective Approaches

Here we explore three self-supervision objectives from the literature that are best suited to bridge the vocabulary gap and error correction. Examples from the gastrointestinal domain for each objective are presented in Fig. 2, where coloured boxes refer to standard objective approaches: summarization, paraphrasing and mask-filling respectively.

Summarisation task generates a summary for given text input. The goal is to capture key points of the input and present them in a concise manner.

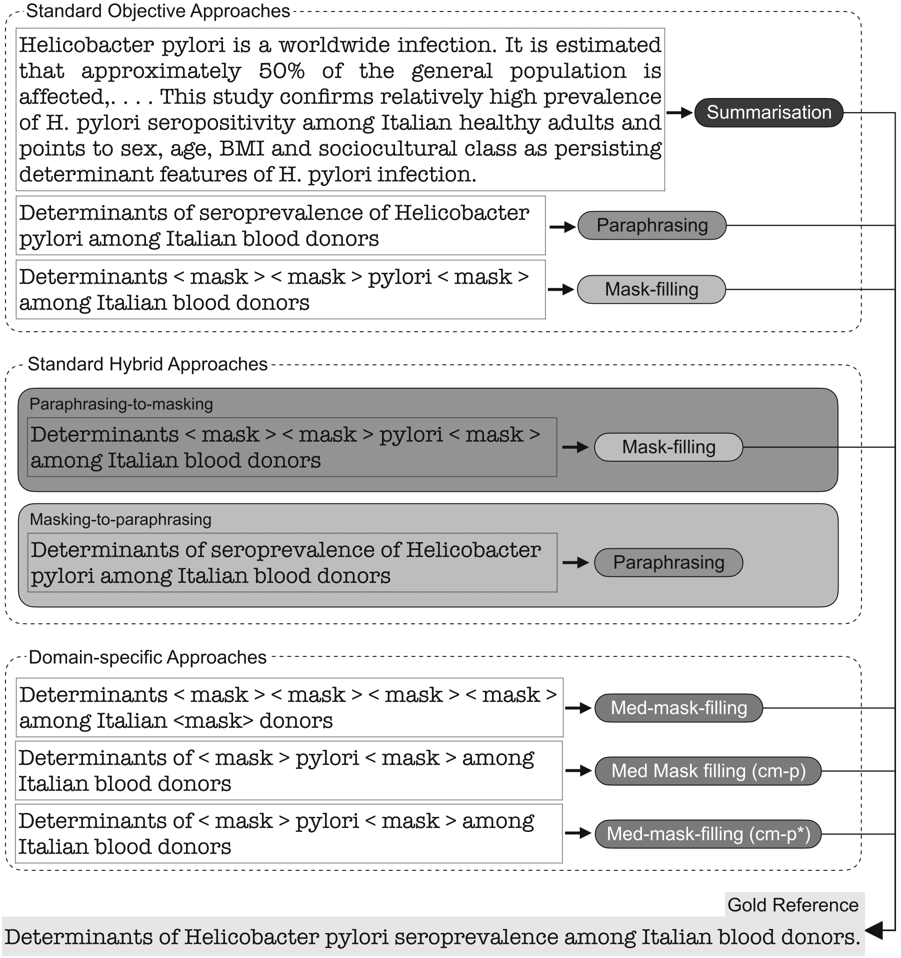
Standard Objective Approaches
Helicobacter pylori is a worldwide infection. It is estimated that approximately 50% of the general population is affected,. . . . This study confirms relatively high prevalence of H. pylori seropositivity among Italian healthy adults and points to sex, age, BMI and sociocultural class as persisting determinant features of H. pylori infection.
Summarisation
Determinants of seroprevalence of Helicobacter pylori among Italian blood donors
Paraphrasing
Determinants < mask > < mask > pylori < mask > among Italian blood donors
Mask-filling
Standard Hybrid Approaches
Paraphrasing-to-masking
Determinants < mask > < mask > pylori < mask > among Italian blood donors
Mask-filling
Masking-to-paraphrasing
Determinants of seroprevalence of Helicobacter pylori among Italian blood donors
Paraphrasing
Domain-specific Approaches
Determinants < mask > < mask > < mask > < mask > among Italian <mask> donors
Med-mask-filling
Determinants of < mask > pylori < mask > among Italian blood donors
Med Mask filling (cm-p)
Determinants of < mask > pylori < mask > among Italian blood donors
Med-mask-filling (cm-p*)
Gold Reference
Determinants of Helicobacter pylori seroprevalence among Italian blood donors.

Fig. 2. Self-supervision strategies

Paraphrasing task generates a rephrased text for a given text input. This aims to rephrase the input text while preserving semantic meaning using synonyms or by re-arranging words.

Mask-filling is the task of predicting missing words when indicated by a masked token in the input text. A percentage of the input text is replaced with a <*mask*> token, and the goal is to predict the masked words based on semantic relations.

Algorithm 1. Med-mask-filling

```
Require: $D = [S_1, S_2, ..., S_N]$: reference text document
Require: $M = [m_1, m_2, ..., m_K]$: medical vocabulary
Require: $p$: masking percentage
for all $S \in D$ do
    $S = [w_1, w_2, ..., w_n]$
    $I^M = \{i \mid w_i \in M, w_i \in S\}$
    $|I^M| = k$
    $words\_to\_mask = n \times p$
    if $words\_to\_mask = k$ then
        $(S, S') \leftarrow mask(S, I^M)$
    else if $words\_to\_mask > k$ then
        $temp \leftarrow mask(S, I^M)$
        $\hat{I} = \{j \mid w_j \in S, w_j \notin M\}$
        $q = p - \frac{k}{n}$
        $\hat{I} \leftarrow random\_select(q(n-k), \hat{I})$
        $(S, S') \leftarrow mask(temp, \hat{I})$
    else if $words\_to\_mask < k$ then
        $I^{\hat{M}} \leftarrow random\_select(n \times p, I^M)$
        $(S, S') \leftarrow mask(S, I^{\hat{M}})$
    end if
    $\mathcal{X} \leftarrow (S, S')$
end for
return $\mathcal{X}$
```

3.3 Standard Hybrid Approaches

In hybrid approaches, we explore multiple standard self-supervision tasks in an ordered manner and evaluate their impact on the model fine-tuning. Paraphrasing and mask-filling are used here as they are the most similar to error correction and also being informed by initial empirical evaluations.

Paraphrasing-to-masking is a hybrid approach where we perform paraphrasing followed by mask-filling. As shown in Fig. 2, first, the pre-trained language model is fine-tuned for paraphrasing followed by a second objective of mask-filling. Intuitively, the masking-only approach is limited to the context, but by introducing paraphrasing-to-masking, we focus on expanding the contexts for the words in which they appear.

Masking-to-paraphrasing is a hybrid approach where mask-filling is the first fine-tuning objective, followed by paraphrasing.

3.4 Domain-Specific Approaches

The goal of domain-specific self-supervision approaches are to further influence the model to reduce the vocabulary gap. Our approach to domain-specific self-supervision using conditional masking is presented in Algorithm 1. Here the inputs to the conditional masking are the reference text document D and the medical vocabulary M which is specific to the medical domain of interest and consists of a list of specialist terms. Dataset compilation is described in Sect. 4.1.

Med-mask-filling objective is derived from standard mask-filling where we randomly replaced a percentage (p) of the words in each sentence with the token $<mask>$. However, in med-mask-filling, instead of random masking, we are masking all the medical words in the sentence identified using the medical vocabulary (M). This objective ignores masking percentage p but satisfies the condition on Line 6.

Med-mask-filling (cm-p) where cm-p stands for conditional masking percentage consider two cases; (1) if the sentence contains at least one medical word ($k > 0$) we ensure they are prioritised before masking non-medical words, this may satisfy one of the three conditions in the Algorithm 1 lines 6, 8 or 14 based on k and p; and (2) in the absence of any medical words a random mask is used which satisfies condition in Line 8.

Med-mask-filling (cm-p*) is similar to the previous task; except that sentences with no medical words are not included in the Document D. This will enable us to evaluate the impact of including and excluding random masking as part of med-mask-filling.

4 Evaluation

In this section, we evaluate self-supervision strategies for clinical dialogue error correction using two pre-trained language models and compare them against commercial ASR systems. The language models are fine-tuned using the PubMed dataset and evaluated using the GCD dataset.

4.1 Datasets

Gastrointestinal Disease Dataset (GCD) consists of a set of role-playing clinical dialogues that took place at the NHS IBD Clinic. The data collection included clinical dialogues recorded with 7 participants with Scottish accents transcribed using commercial ASR systems. Here, the accent can be viewed as a form of noise in addition to common noise factors such as background noise, interruptions and repetitions. Each audio clip contains around 47 utterances by two persons engaged in a clinical conversation that is about 4–5 min long. Statistics of the GCD dataset can be found in Table 1 and some examples are presented in Table 2.

PubMed Dataset for Self-supervision the PubMed dataset consists of abstract and title pairs scraped from articles related to gastrointestinal conditions. Following variants of the PubMed dataset were created for evaluating self-supervision strategies. An example for each self-supervision task is presented in Table 3.

- **Summarisation** considers abstract as the input and title as the expected output.
- **Paraphrasing** considers a paraphrased version of the title as the input and the title as the expected output. The paraphrased title is obtained using the T5 model fine-tuned for paraphrasing using the Google PAWS Dataset [23].
- **Mask-filling** apply <*mask*> token to 25% of the words in the title to create the input and use the title as the expected output.
- **Hybrid approaches** use the above datasets created for paraphrasing and mask-filling for fine-tuning.
- **Med-mask-filling (cm-p)** strategies use datasets created using Algorithm 1.

The dataset was curated from PubMed articles with the primary goal of introducing domain-specific medical vocabulary to language models pre-trained on public domain data. The lack of availability of a larger spoken corpus in medical conversations has led us to use a written corpus, although we acknowledge the differences between written and spoken language in specialist domains. After pre-processing, we obtain a dataset with title and abstract pairs (see Table 3). This extraction method can be generalised to any medical domain by using domain-specific search queries in the PubMed search engine.

Medical Vocabulary (M) is a set of domain-specific medical terms extracted from the PubMed articles using the ScispaCy [16] models. This medical dictionary for masking contains 4231 medical terms related to the gastrointestinal area.

4.2 Experiment Setup

To compare the different self-supervision strategies we experiment with two pre-trained language models, T5 (T5-small) [19] and BART (BART-base) [9]. Two additional variants of the PubMed dataset were created to support self-supervision strategies: masking and paraphrasing. The hyper-parameters for fine-tuning were kept constant across all strategies as: optimiser is AdamW [11]; loss is cross-entropy; learning rate is 2e−5; and batch size is 16. The PubMed dataset was split 90/10 as training and validation sets and the fine-tuning was early-stopped between 10–40 epochs based on minimal validation loss. All strategies were evaluated across four commercial ASR transcriptions of the GCD dataset

Table 1. Summary of the GCD dataset

Feature	Value
No. of audio files	**7**
Mean length of an audio file	**4 min 49 s**
Mean no. of utterances in a file	**47**
Mean no. of words in an utterance	**93**

Table 2. Examples from the GCD dataset

Gold Reference	Transcription Output
So do you have any ideas as to what might be the cause of your symptoms at the moment?	So do you have any ideas as to what might be the cause of your symptoms at the moment?
Have you noticed any changes in your weight?	Do you noticed any changes in your *wit*?
Okay have you noticed any mucus in your bowel motions?	Okay have you noticed any mucus in your *bible Moshe*?

Table 3. PubMed gastrointestinal dataset pre-processed for self-supervision strategies

Task	Input	Output
Summarisation	Helicobacter pylori is a worldwide infection. It is estimated that approximately 50% of the general population is affected, but this percentage varies considerably between countries. . . . This study confirms relatively high prevalence of H. pylori seropositivity among Italian healthy adults and points to sex, age, BMI and sociocultural class as persisting determinant features of H. pylori infection.	Determinants of Helicobacter pylori seroprevalence among Italian blood donors.
Paraphrasing	Determinants of seroprevalence of Helicobacter pylori among Italian blood donors.	
Mask-filling	Determinants <*mask*> <*mask*> pylori <*mask*> among Italian blood donors.	
Paraphrasing-to-masking	Determinants <*mask*> <*mask*> pylori <*mask*> among Italian blood donors.	
Masking-to-paraphrasing	Determinants of seroprevalence of Helicobacter pylori among Italian blood donors.	
Med-mask-filling	Determinants <*mask*> <*mask*> <*mask*> <*mask*> among Italian <*mask*> donors.	
Med-mask-filling (cm-25)	Determinants of <*mask*> pylori <*mask*> among Italian blood donors.	
Med-mask-filling (cm-25*)	Determinants of <*mask*> pylori <*mask*> among Italian blood donors	

generated using Amazon Web Services (AWS) Transcribe, Google Speech-to-text, Microsoft Speech-to-text, and IBM Watson. For med-mask-filling masking percentage (p) is considered as 25% denoted by $cm - 25$.

Language models were implemented using Python Hugging Face and PyTorch frameworks while maintaining all default hyper-parameters from the base models. For the summarisation task, the encoder input and decoder output sequence lengths were set to 1024 and 128, respectively; for paraphrasing and mask-filling tasks, both encoder input and decoder sequence lengths were set to 512. Our model implementation and the reproducible code are available in GitHub[2] and the fine-tuned model variants and PubMed datasets are publicly available in Huggingface[3].

4.3 Performance Metric

Word Error Rate (WER) was selected to measure the performance of clinical dialogue error correction. WER has been used as a performance metric in ASR systems [2,3] and in post-ASR error correction [8,12]. Given a language model output and a reference text, where N refers to the number of words in the reference text and SUB, DEL and INS refer to the number of substitutions, deletions, and insertions operations needed to transform the reference text to the language model output, WER is calculated as in Eq. 1.

$$WER = \frac{SUB + DEL + INS}{N} \tag{1}$$

Lower WER scores are desirable as they indicate higher accuracy in speech recognition systems, reflecting a smaller number of word-level errors in the transcriptions.

4.4 Results

Table 4 presents a comparison of WER scores from commercial ASR systems and language models where we applied different self-supervision strategies. The WER of commercial ASR transcription against reference text is considered the baseline against which we want to improve. The best performing self-supervision strategy for each language model is highlighted in bold text. Overall, the best performing strategy is med-mask-filling with the BART model, and it has reduced WER by 10.27% of Microsoft, 12.13 % of IBM and 16.01% of Google transcriptions. In the case of AWS, while both med-mask-filling and mask-filling did lead to improvements, the degree of enhancement was not as substantial when compared to the other ASR systems.

Apart from the missed transcription of clinical terms, different ASR systems introduce different types of errors when generating transcripts from the audio. For example, Google ASR drops many words from the ASR output based on low

[2] https://github.com/gayaninan/clinical-error-correction.

[3] https://huggingface.co/gayanin.

Table 4. Comparison of self-supervision strategies for clinical dialogue error correction

Model	Self-supervision	GCD Dataset WER (%)			
	Strategy	AWS Transcribe	Microsoft	IBM Watson	Google
Baseline (Commercial ASR)		33.02	29.03	44.28	47.78
T5	Summarisation	63.39	66.89	69.44	73.80
	Paraphrasing	48.87	47.24	54.52	57.97
	Mask-filling	38.83	35.86	45.16	46.87
	Masking-to-paraphrasing	43.28	50.89	40.41	47.56
	Paraphrasing-to-masking	39.79	46.36	34.42	45.89
	Med-mask-filling	56.39	51.23	58.79	61.61
	Med-mask-filling (cm-25)	48.08	47.08	55.09	57.89
	Med-mask-filling (cm-25*)	46.10	43.03	52.23	53.06
BART	Summarisation	76.61	77.03	78.10	75.56
	Paraphrasing	43.31	37.46	47.51	49.48
	Mask-filling	**32.38**	26.38	38.92	40.43
	Masking-to-paraphrasing	46.43	41.45	49.24	51.15
	Paraphrasing-to-masking	32.77	26.82	39.45	40.97
	Med-mask-filling	32.53	**26.05**	**38.91**	**40.13**
	Med-mask-filling (cm-25)	32.61	26.48	39.28	40.69
	Med-mask-filling (cm-25*)	32.75	26.61	39.43	40.90

transcription confidence. Accordingly, we observe increased deletion operations required to convert the reference to the ASR output (in WER calculation).

We also observed that various ASR systems introduced distinct types of errors while generating transcripts. To investigate this further, in Fig. 3, we analysed the INS, DEL, and SUB operation counts for the language models that achieved the best performance for each self-supervision strategy. Google ASR tends to drop several words from the ASR output due to low transcription confidence. As a result, we noticed that there is an increased number of deletion operations required to transform the reference text into the ASR output during the calculation of WER.

In addition to that, summarisation resulted in the highest WER scores due to the length difference between generated and reference text, which makes it an unsuitable strategy for error correction. Intuitively, mask filling is more similar to error correction than paraphrasing, which is evidenced in the results. Accordingly, we use mask-filling in a domain-specific approach with additional emphasis given to correcting clinical terms (by masking clinical terms).

Comparing BART and T5, our results showed that BART is more suitable for clinical dialogue error correction. BART and T5 are both language models pre-trained for de-noising. To create noise, T5 masked 15% of words in a sequence, each word replaced by a mask [19] whereas BART used text infilling where zero or more consecutive words are replaced by a single mask [9]. Accordingly, BART had learned to predict the number of words masked in addition to predicting

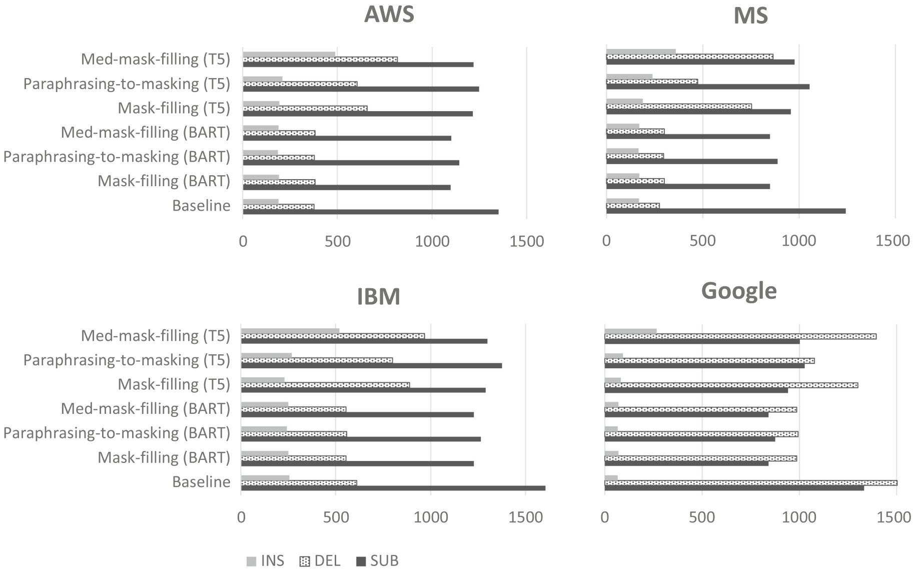

Fig. 3. INS, DEL, SUB operation counts for the best performing models for each self-supervision strategy

masked words. This is advantageous when performing clinical dialogue error correction where clinical terms can be erroneously transcribed into one or more commonly occurring words.

5 Conclusion

In this paper, we introduce a novel strategy of self-supervision for the task of clinical dialogue error correction utilizing language models. Our method addresses the challenge of sparse real-world clinical dialogue data by incorporating clinical data from the public domain. Our findings reveal that the proposed med-mask-filling strategy effectively reduces transcription errors when benchmarked against prevalent commercial ASR systems. The results underline the criticality of not only choosing the right self-supervision strategy but also understanding the impacts of varying error types generated by ASR systems. Moving forward, our focus will be on refining the GCD dataset and researching ways to continue the reduction of transcription errors. An in-depth analysis of language model outputs indicates an opportunity for further narrowing the domain-specific vocabulary gap, suggesting that the integration of knowledge graph representations is a promising path to explore.

Acknowledgements. We would like to thank the NHS Grampian, including Amy Kousha, John Thomson and Jill Ferbrache, who helped to curate the IBD clinic role-playing dialogues.

References

1. Cucu, H., Buzo, A., Besacier, L., Burileanu, C.: Statistical error correction methods for domain-specific ASR systems. In: Dediu, A.-H., Martín-Vide, C., Mitkov, R., Truthe, B. (eds.) SLSP 2013. LNCS (LNAI), vol. 7978, pp. 83–92. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39593-2
2. Errattahi, R., El Hannani, A., Ouahmane, H.: Automatic speech recognition errors detection and correction: a review. Procedia Comput. Sci. **128**, 32–37 (2018). 1st International Conference on Natural Language and Speech Processing
3. Filippidou, F., Moussiades, L.: A benchmarking of IBM, Google and wit automatic speech recognition systems. In: Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds.) AIAI 2020. IAICT, vol. 583, pp. 73–82. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49161-1
4. Humphries, J.J., Woodland, P.C., Pearce, D.J.B.: Using accent-specific pronunciation modelling for robust speech recognition. Proceeding of Fourth International Conference on Spoken Language Processing, ICSLP 199, vol.6 4, pp. 2324–2327 (1996)
5. Jain, A., Upreti, M., Jyothi, P.: Improved accented speech recognition using accent embeddings and multi-task learning. In: INTERSPEECH (2018)
6. Kamper, H., Niesler, T.: Multi-accent speech recognition of Afrikaans, black and white varieties of south African English. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3189–3192 (2011)
7. Leng, Y., et al.: FastCorrect 2: fast error correction on multiple candidates for automatic speech recognition. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 4328–4337. Association for Computational Linguistics, Punta Cana (2021)
8. Leng, Y., et al..: FastCorrect: fast error correction with edit alignment for automatic speech recognition. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 21708–21719. Curran Associates, Inc. (2021)
9. Lewis, M., et al.: 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, pp. 7871–7880. Association for Computational Linguistics, Online (2020)
10. Li, W., Di, H., Wang, L., Ouchi, K., Lu, J.: Boost transformer with BERT and copying mechanism for ASR error correction. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
12. Mani, A., Palaskar, S., Konam, S.: Towards understanding ASR error correction for medical conversations. In: NLPMC (2020)
13. Mani, A., Palaskar, S., Meripo, N.V., Konam, S., Metze, F.: ASR error correction and domain adaptation using machine translation. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6344–6348. IEEE (2020)
14. McDonald, A., Sherlock, J.: A long and winding road - improving communication with patients in the NHS (2016)
15. Nanayakkara, G., Wiratunga, N., Corsar, D., Martin, K., Wijekoon, A.: Clinical dialogue transcription error correction using Seq2Seq models. In: Shaban-Nejad,

A., Michalowski, M., Bianco, S. (eds.) Multimodal AI in Healthcare. Studies in Computational Intelligence, vol. 1060, pp. 41–57. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-14771-5_4D

16. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327. Association for Computational Linguistics, Florence (2019)
17. Quiroz, J., Laranjo, L., Kocaballi, A.B., Berkovsky, S., Rezazadegan, D., Coiera, E.: Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ Digit. Med. **2**, 114 (2019)
18. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
19. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
20. Sarma, A., Palmer, D.D.: Context-based speech recognition error detection and correction. In: Proceedings of HLT-NAACL 2004: Short Papers, pp. 85–88. Association for Computational Linguistics, Boston (2004)
21. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
22. Wagner, R.A., Fischer, M.J.: The string-to-string correction problem. J. ACM (JACM) **21**(1), 168–173 (1974)
23. Zhang, Y., Baldridge, J., He, L.: PAWS: paraphrase adversaries from word scrambling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1298–1308. Association for Computational Linguistics, Minneapolis (2019)
24. Zhao, Y., Yang, X., Wang, J., Gao, Y., Yan, C., Zhou, Y.: BART based semantic correction for mandarin automatic speech recognition system. In: Proceedings of the Interspeech 2021, pp. 2017–2021 (2021)