Benchmarking ASR Systems Based on Post-Editing Effort and Error Analysis

Martha Maria Papadopoulou[1], Anna Zaretskaya[2], and Ruslan Mitkov[1]

[1] University of Wolverhampton, UK {m.m.papadopoulou,R.Mitkov}@wlv.ac.uk
[2] TransPerfect azaretskaya@translations.com

Abstract. This paper presents a comparative evaluation of four commercial ASR systems which are evaluated according to the post-editing effort required to reach "publishable" quality and according to the number of errors they produce. For the error annotation task, an original error typology for transcription errors is proposed. This study also seeks to examine whether there is a difference in the performance of these systems between native and non-native English speakers. The experimental results suggest that among the four systems, Trint and Microsoft obtain the best scores. It is also observed that most systems perform noticeably better with native speakers and that all systems are most prone to fluency errors.

Keywords: Automatic Speech Recognition· speech-to-text· post-editing · error annotation.

1 Introduction

The rapid technological progress in the field of Automatic Speech Recognition (ASR) has lead to claims that speech-to-text systems can achieve up to 90% accuracy [9,15]. The aim of this paper is to shed some light on the impact that this progress has on the productivity of end users. Until now, the evaluation of ASR systems relied exclusively on Word Error Rate (WER) and similar metrics. Calculating these metrics is usually an expensive and time-consuming task as manual transcriptions are used for reference. In addition, these traditional approaches do not provide information on the cognitive effort required to reach "publishable" quality. In this paper, the aim is to address the aforementioned issues by proposing a way to depart from the traditional methods of ASR evaluation. The key idea is to deploy the post-editing (PE) method in the evaluation process. To bridge the gap of the underrepresented aspect of cognitive effort, four ASR systems: Amazon[3], Microsoft[4], Trint[5] and Otter[6] were evaluated based on

[3] https://aws.amazon.com/transcribe/
[4] https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/
[5] https://trint.com/
[6] https://otter.ai/

Translation and Interpreting Technology Online, pages 199–207,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_023

post-editing effort. To this end, the PET tool [1] was employed to compute the post-editing (PE) effort in terms of PE time and PE distance at a sentence level. The objective of the PE process was to rank all systems based on their overall score.

In the attempt to provide a qualitative analysis, a secondary objective seeks to investigate the types of errors that these systems produce. To accomplish this, a new error typology for transcription errors was developed following the TAUS DQF-MQM [17] main error categories. In this novel typology, which is essential for this study, the subcategories are tailored to suit transcription errors. To the best of our knowledge, this is the first study that seeks to investigate which types of errors the ASR systems are most prone to. The comparison between the error annotation results and the post-editing results will lead to new insights of their correlation.

Another goal of this study is to examine the role of the speaker's accent. It investigates whether the performance of the systems is affected by the speaker's accent. To address this question, results from native and non-native English speakers are compared.

The rest of the paper is structured as follows: Section 2 contextualises the current study by discussing related work. Section 3 outlines the data used, Section 4 presents the experimental setup. Section 5 discusses the results of the experiments conducted. Finally, Section 6 provides the conclusions of this study.

2 Related Work

The rapid development of state-of-the-art Automatic Speech Recognition systems led to the need for these systems to be evaluated. A recent study [2] benchmarked commercial ASR systems by comparing their results against human quality and evaluating them using the WER metric. This research pays specific attention to named entity recognition. Related research includes [6], where a tool was designed to perform comparisons between commercial and open-source ASR systems using the WER metric. A recent systematic review [14] discusses the problems of benchmarking ASR systems, which were presented in various studies and expresses skepticism for the very low WER results reported. They demonstrated that the WER rate was considerably higher than the best results reported in those studies. A further study [4] also benchmarked three commercial ASR systems, but they reported results using three metrics: WER, Hper and Rper. A qualitative analysis [10] on ASR systems was performed aiming to evaluate the accuracy of the Language Model adaptation; in order to do so, the WER metric was applied only to relevant words.

It is worth noting that none of the existing approaches appears to overcome the limitations of the WER metric. There is therefore a need for a new evaluation approach. Two new performance metrics: MER (match error rate) and WIL (word information lost) were proposed in [12]. Furthermore, with the aim to represent human perception of ASR accuracy, HPA (Human Perceived Accuracy) was developed [11]. Another metric was introduced in [3] seeking to achieve a

better correlation to human evaluation. Finally, an extension of the WER metric was proposed in [8], where weighted penalties were applied by implementing word embeddings.

To the best of our knowledge, studies that employ post-editing in order to evaluate ASR performance are scarce. Post-editing was explored in [7], where users browsed and corrected automatic transcriptions of lectures in a web-based interactive interface. This study aimed to compare WER rates with comprehensibility improvements after transcripts were post-edited. As detailed in [16], an ASR system was developed for Polish, which introduced the novel idea of applying automatic post-editing in the ASR output. Finally, two crowdsourcing studies were compared in [5] with the objective to investigate whether it is preferable to transcribe from scratch or to perform post-editing on ASR output. They concluded that post-editing is preferable only when WER accuracy is lower than 30%. However, effort indicators of the post-editing task were not examined in this study.

The above discussion provides compelling evidence that there is a pressing need for an alternative approach to account for the cognitive effort required to post-edit raw ASR outputs. To the best of our knowledge, this study constitutes the first analysis of the evaluation of post-editing effort in this field. The added value of this paper is also highlighted by the qualitative analysis on the transcription errors, which remains unexplored in the literature. With this aim in mind, this paper puts forward an error typology for ASR transcription errors. The suggested error typology is the first of its kind to be specifically designed for the use case where the ASR output is post-edited by humans to reach "publishable quality".

3 Data Description

For the purpose of this study, the video data were obtained from the research seminar series "Specialised Seminar: Technologies for Translation and Interpreting: Challenges and Latest Developments"[18], hosted by Prof R Mitkov at the University of Wolverhampton. More specifically, the videos were recordings of talks given by invited speakers on topics related to Translation and Interpreting Technologies, which were held online via Zoom. Thus, all data have the same register and belong to the same domain. It should be mentioned that for Microsoft and Amazon ASR systems the video files were converted to audio files (.wav), as these systems operate exclusively on this file format. Two videos were used as input data: one with a native American English speaker and one with a non-native American English speaker. The mother tongue of the non-native English speaker is Russian. The videos were trimmed in order to have the same length—approximately 15 minutes per video. Each ASR system produced a transcription of approximately 2,000 words per video, thus the size of the post-editing and error annotation tasks for all four systems consisted of approximately 16,000 words.

4 Experimental Setup

4.1 Post-editing

The transcriptions produced by the ASR systems were exported in simple text format and tokenised into sentences in order to be imported into the post-editing tool. The tokenisation task was performed using the Punkt Sentence Tokenizer module from the NLTK Python library. The post-editing process was carried out using the PET tool [1], an open-source post-editing tool, which served a double purpose both to facilitate the post-editing task and to collect sentence-level information. Along with the post-editing process, this tool gathered information related to the post-editing effort such as editing time and number of edits per segment. These results were exported to calculate the post-editing effort.

The character-based Levenshtein distance was used in this experiment. It was calculated on the basis of the number of characters that were changed (insertions, deletions and substitutions) out of the total number of characters in the segment.

The PE task was performed by a single post-editor with intermediate experience in the field. As the desired outcome was a verbatim transcription, the post-editor was instructed to perform light post-editing. For this reason, speech disfluencies and repetitions were not corrected.

4.2 Error Annotation

For the error annotation task, the BLAST tool [13] was used, which is an open-source tool. For the purposes of this task an error typology was designed following the DQF-MQM TAUS Error Typology format, which was customised to correspond to transcription-related errors only (see Table 1). The DQF-MQM TAUS Error Typology was selected as a basis since its main error categories correspond to transcription errors and customisation was only required for the sub-categories. The error annotation task was performed by the post-editor. The results of each annotation task were automatically generated by the BLAST tool.

5 Results

As seen in Table 2, Microsoft obtained the best score for total PE time. It is also worth noting that all systems required more PE time for the non-native speaker, with the exception of Otter. It was also noted that Otter required the most PE time for the native speaker transcription. This is mainly caused by the increased average segment length of 144.32 characters compared to the rest of the systems, whose average segment length range between 78.13–97.88 characters. In particular, Otter's average segment length reached a peak of 5,082 characters in a single segment. It is worth highlighting the significant difference in PE time between the native and non-native Amazon transcriptions. This is also represented in the PE distance and will be discussed further as part of the error analysis.

Table 1. Error Typology

Accuracy	Omission	Prefix Suffix Article Preposition
	Addition	Prefix Suffix Article Preposition
	Mistranscription	Proper noun Number Single to multiple words Multiple to single word Single to single word Multiple to multiple words
	Homophone	
Fluency	Segmentation	
	Punctuation	Additional punctuation mark Missing punctuation mark Wrong punctuation mark
	Spacing Capitalisation Filler word	
Grammar	Grammatical number Grammatical tense	
Style	Inconsistent style Abbreviated form Spelled out form	
Terminology	Term Abbreviation	

Table 2. Total PE time(s)

	Native Speaker	Non-Native Speaker	Total
Microsoft	2,087.93	2,426.79	4,514.72
Trint	2,165.87	2,442.37	4,608.25
Amazon	1,520.41	4,855.12	6,375.53
Otter	5,550.37	3,039.74	8,590.11

In terms of average and overall PE distance, Trint produced the best score (see Table 3). The aforementioned differences between native and non-native speakers for Amazon and Otter are also reflected in the PE distance results.

Table 3. Average PE distance per segment

	Native Speaker	Non-Native Speaker	Overall Average
Trint	4.14%	7.41%	5.69%
Otter	8.95%	4.50%	6.43%
Microsoft	5.34%	8.65%	7.10%
Amazon	4.37%	15.66%	9.08%

According to the PE results, Trint performed best in terms of post-editing effort, taking into consideration both PE distance, where it scored first, and in terms of PE time, where it delivered the second best results but with small differences from the first system.

According to the error annotation results, Trint performed the best with the lowest total number of errors for both speakers (see Table 4).

Table 4. Total number of errors

	Native Speaker	Non-Native Speaker	Total
Trint	109	185	294
Microsoft	141	210	351
Otter	213	250	463
Amazon	163	464	627

As for the results related to the different error categories, a general tendency towards fluency errors was observed (see Table 5). The percentage of fluency errors ranges between 48.55%–71.12% of the total errors. The tendencies towards the second and the third most frequent error categories are also consistent through all systems, with accuracy ranging between 21.57%–37.08% in second place, and terminology ranging between 4.20%–7.86% in third place.

Table 5. Percent of errors per error category

	Fluency	Accuracy	Terminology	Grammar	Style
Trint	48.55%	37.08%	7.86%	2.72%	3.80%
Microsoft	56.20%	32.04%	5.95%	3.80%	2.01%
Otter	71.12%	21.57%	5.20%	1.03%	1.07%
Amazon	60.81%	32.48%	4.20%	2.19%	0.21%

As seen in Table 6, the systems are ranked in terms of PE time, PE distance and number of errors. It is evident that the number of errors does not always correlate with the PE effort. The results support the conclusion that systems with lower number of errors do not necessarily have the best score in terms of PE time and PE distance.

Table 6. Ranking systems based on PE time, PE distance and number of errors

PE time	PE distance	Number of errors
Microsoft	Trint	Trint
Trint	Otter	Microsoft
Amazon	Microsoft	Otter
Otter	Amazon	Amazon

A closer look at the error annotation results suggests further observations regarding the correlation of PE time and error categories. Firstly, there is a strong correlation between fluency errors and PE time: the higher the rate of fluency errors the more PE time is required. For example, Otter has the highest fluency rate and is the system that required almost twice as much PE time as the systems that ranked first and second (see Table 3). The most frequent fluency errors in this case were punctuation and segmentation. These two categories also justify the longer segment rate for Otter and the correlation with the increased PE time.

Secondly, a weak correlation between accuracy errors and PE time is noted. A high rate of accuracy errors, contrary to the popular belief, does not require extra PE time. For instance, Trint reported the highest accuracy rate; however, it was ranked second based on PE time. In this case, the low correlation could be justified by the high number of omission and addition errors, which are easily detectable and require less cognitive effort, combined with the low number of mistranscription errors, which require more cognitive effort.

Finally, it should be highlighted that there is a big performance difference in PE time between native and non-native speakers for Amazon. This difference can be explained by the high number of filler word, mistranscription, segmentation and terminology errors of the non-native speaker transcription.

6 Conclusions

In this study, outputs from commercial ASR systems were post-edited and then the errors were annotated. The ASR systems were ranked based on the post-editing effort required to reach "publishable" quality and the number of errors they produced. In accordance with the PE and error results presented above, it can be concluded that with the data used in this experiment, Trint is the best performing system in terms of PE distance and total number of errors, while Microsoft is the best performing system in terms of PE time. Moreover, the

number of errors does not always correlate with the PE effort. It is also evident that there is a general tendency towards fluency errors, which are assumed to be the most time-consuming errors. The experiments point to the conclusion that most ASR systems perform better with a native speaker.

The constraints of this study include its limited scope and the involvement of only one post-editor and annotator; larger-scale study results may be different. While the size of the data was another constraint, the results reported remain insightful. In particular, this study will pave the way for further research in the field of ASR evaluation, post-editing and error analysis. Future work could explore the correlation between the suggested approach and the traditional WER metric.

7 Acknowledgements

I would like to thank TransPerfect for offering me a placement opportunity to conduct this research. In particular, I am very grateful to my placement supervisor Anna Zaretskaya for her guidance and support. I would also like to thank my supervisor Prof R Mitkov for his constant support and encouragement. Some special thanks go to my classmate Rea Bartlett Tandon for proofreading this paper.

References

1. Aziz W, Castilho S, Specia L.: PET: a Tool for Post-editing and Assessing Machine Translation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp. 3982-3987. European Language Resources Association (ELRA) (2012)
2. Del Rio M, Delworth N, Westerman R, Huang M, Bhandari N, Palakapilly J, McNamara Q, Dong J, Zelasko P, Jette M.: Earnings-21: A Practical Benchmark for ASR in the Wild. arXiv preprint arXiv:2104.11348 (2021).
3. Favre B, Cheung K, Kazemian S, Lee A, Liu Y, Munteanu C, Nenkova A, Ochei D, Penn G, Tratz S.: Automatic human utility evaluation of ASR systems: Does WER really predict performance? In: INTERSPEECH-2013, pp 3463–3467 (2013).
4. Filippidou, F., Moussiades, L.: A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. In: International Conference on Artificial Intelligence Applications and Innovations IFIP, pp.73-82. Springer International Publishing (2020).
5. GaurY, Lasecki WS, Metze F, Bigham JP.: The effects of automatic speech recognition quality on human transcription latency. In: 13th Web for All Conference, pp.1–8. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2899475.2899478
6. Këpuska, V., Bohouta, G.: Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). Int. Journal of Engineering Research and Application 7(3), 20–24 (2017).
7. KolkhorstH, Kilgour K, Stüker S, Waibel A.: Evaluation of interactive user corrections for lecture transcription. In: International Workshop on Spoken Language Translation (IWSLT) 2012, pp. 1-8 (2012).

8. Le N-T, Servan C, Lecouteux B, Besacier L.: Better evaluation of ASR in speech translation context using word embeddings. In: Interspeech 2016, pp. 1-6 (2016).
9. LevitM, Chang S, Buntschuh B, Kibre N.: End-to-end speech recognition accuracy metric for voice-search tasks. In: International Conference on Acoustics (IEEE), Speech and Signal Processing (ICASSP), pp. 5141-5144 (2012). https://doi.org/10.1109/ICASSP.2012.6289078
10. MdhaffarS, Estève Y, Hernandez N, Laurent A, Dufour R, Quiniou S: Qualitative Evaluation of ASR Adaptation in a Lecture Context: Application to the PASTEL Corpus. In: INTERSPEECH-2019, pp.569–573 (2019).
11. MishraT, Ljolje A, Gilbert M.: Predicting human perceived accuracy of ASR systems. In: 12th Annual Conference of the International Speech Communication Association, pp.1945-1948 (2011).
12. MorrisAC, Maier V, Green P.: From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: 8th International Conference on Spoken Language Processing, pp. 2765-2768 (2004).
13. Stymne, S.: Blast: A tool for error analysis of machine translation output. In: Proceedings of the ACL-HLT 2011 System Demonstrations, pp. 56-61 (2011).
14. SzymańskiP, Żelasko P, Morzy M, Szymczak A, Żyła-Hoppe M, Banaszczak J, Augustyniak L, Mizgajski J, Carmiel Y: WER we are and WER we think we are. arXiv preprint arXiv:2010.03432 (2020).
15. WilliamsJD, Melamed ID, Alonso T, Hollister B, Wilpon J.: Crowdsourcing for difficult transcription of speech. In: IEEE Workshop on Automatic Speech Recognition & Understanding, pp.535-540 (2011). https://doi.org//10.1109/ASRU.2011.6163988.
16. Wnuk D, Wołk K.: Post-editing and Rescoring of Automatic Speech Recognition Results with OpenNMT-APE, In: Proceedings of the PolEval 2020 Workshop, pp. 33-37 (2020).
17. Harmonized DQF-MQM Error Typology, `https://www.taus.net/qt21-project#harmonized-error-typology`. Last accessed 22 May 2021
18. Technologies for Translation and Interpreting: Challenges and Latest Developments 2020/21, `https://em-tti.eu/em-tti-seminar-series/`. Last accessed 22 May 2021