INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Neural Text Denormalization for Speech Transcripts

Benjamin Suter, Josef Novak

Spitch, Switzerland
benjamin.suter@spitch.ch, joe@spitch.ch

Abstract

This paper presents a simple sequence-to-sequence approach to restore standard orthography in raw, normalized speech transcripts, including insertion of punctuation marks, prediction of capitalization, restoration of numeric forms, formatting of dates and times, and other, fully data-driven adjustments. We further describe our method to generate synthetic parallel training data, and explore suitable performance metrics, which we align with human judgment through subjective MOS-like evaluations.

Our models for English, Russian, and German have a word error rate of 6.36%, 4.88%, and 5.23%, respectively. We focus on simplicity and reproducibility, make our framework available under a BSD license, and share our base models for English and Russian.

Index Terms: text denormalization, orthography restoration, punctuation prediction, inverse text normalization

1. Introduction

Automatic speech recognition (ASR) systems typically generate a sequence of lowercase words without punctuation. In order to both improve human readability and facilitate downstream processing of the output with NLP tools, it is thus typically desirable to fully restore the text to standard orthographic conventions for the target language [1].

This denormalization task includes context-aware prediction of punctuation and capitalization, but also transformation of spelled out numbers into digits, and correct formatting of alphanumeric sequences like dates, times, monetary amounts, fractions, and phone numbers. Furthermore, it involves abbreviating specific words in certain contexts (e.g., *two kilometers* $\rightarrow$ *2 km*), replacement of others with special symbols (e.g., *ten percent* $\rightarrow$ *10%*), and transformation of spelled out names, e-mail and web addresses into their correct written forms.

While limited punctuation restoration has been considered for some time [2, 3, 4], full text denormalization has only recently begun to enjoy a resurgence of concentrated attention in the research community [5, 6]. Despite the proliferation of appropriate raw resources like [7], there is still a dearth of appropriate shared test sets in the speech community, and a lack of agreement on optimal metrics for evaluation. In this work we take some small steps to address these issues by leveraging open data and processing methods, releasing our framework for open use, and investigating objective metric relevance which we align with human judgment through an MOS-like evaluation.

On the technical front, we frame the task as a standard neural machine translation problem as in [1], and leverage a sequence-to-sequence model based on a generic transformer architecture. We show that with modern methods, this simple implementation can provide a competitive balance between accuracy, speed, and training complexity.

Our main contributions are the following: 1) we describe a simple and scalable method to obtain parallel training data from any corpus in standard orthography; 2) we propose suitable metrics and in particular analyze how they align with human judgment; 3) we make our models and basic fairseq-based framework [8] publicly available.[1]

The remainder of this paper is structured as follows: Section 2 provides a brief literature review. Section 3 describes the training data generation, the model architecture and the proposed evaluation metrics. Section 4 describes the experimental results and the results from our MOS-like evaluation. Section 5 concludes the paper with a summary of our approach and an outlook on future research.

2. Related Work

While full text denormalization has received comparatively less attention so far, its counterpart, text normalization, has been studied more thoroughly. Text normalization has typically relied on weighted finite-state transducers (WFST) and handwritten rules [9] until recently, but neural models have begun to successfully take over this space in recent years, [10] and [11].

Most research in text denormalization has focused on punctuation restoration. There are two different approaches to the task: sequence labeling models, which annotate each token with its respective punctuation information, and sequence-to-sequence translation models, which generate the punctuated sequence given the unpunctuated one. As an early example of the latter approach, [1] and [12] use phrase-based machine translation models for punctuation prediction. On the other hand, [2] predict commas and periods with an LSTM sequence labeling model. Their follow-up work [3] uses a bidirectional RNN with an attention mechanism while [4] additionally predicts the position of question marks. A similar approach is presented in [13] for the prediction of commas, periods, and colons in medical texts.

More recently, [14] use an attention-based sequence-to-sequence translation model to predict not only commas, periods, and question marks, but also capitalization of words. In [15], use a bidirectional RNN to jointly learn to label sequences with punctuation marks and truecasing. In [5], the authors use variants of the BERT language model [16] to label sequences with punctuation marks and truecasing, focusing again on medical texts.

An early approach to full text denormalization is [17] which used WFSTs in combination with a language model, including punctuation, capitalization and properly written numeric entities such as times and dates. More recently, [18] proposed an LSTM model to hierarchically label the input sequence with edit operations and the specific target, combined with an FST which would then perform the edit operations. This year, [6] also proposed a transformer model for full text denormalization. Here the authors generate a parallel corpus using an WFST-

[1] https://github.com/spitch-oss/denormalizer

http://dx.doi.org/10.21437/Interspeech.2021-1814

driven text normalization model from a proprietary commercial TTS service. They also experimented with incorporating BERT [16] and observed small performance gains from it. Recently, [19] used language models for the joint task of disfluency correction and denormalization in ASR transcripts.

Some recent approaches [3] furhter incorporate prosodic features from the audio signal into the model. Speech pauses and intonation in particular have been shown to be useful to improve prediction of punctuation marks. For most other orthographic phenomena such as capitalization and number formatting however, prosodic features cannot be expected to provide relevant information. Furthermore, a possible improvement in performance by using prosodic features is offset by increased complexity and a loss of generalisation to text-only input. In this work we focus on a text-in text-out approach as this keeps the model more lightweight and affords a certain flexibility in terms of additional use cases where acoustic information may be unavailable. Also, many orthographic conventions, e.g. the formatting of dates and times, do not reflect how a phrase was verbalized. A separate denormalization component has some advantages over a full end-to-end ASR model which directly outputs denormalized text, as it allows for granular control and customization of the desired orthography, provided the training data is prepared accordingly.

3. Method

3.1. Data

The first challenge, as touched on earlier, is that there are currently no standard community corpora available for full text denormalization. In [11] however, the authors have published English, Russian, and Polish corpora for the closely related task of text normalization[2]. The parallel corpora contain web-scrawled text from Wikipedia which was run through a rule-based TTS component to generate verbalizations. Here we utilize the English and Russian corpora as a starting point, and after applying some basic text-sanitization rules, simply swap the source and target data. These English and Russian training sets contain $\sim$ 78.9 M and $\sim$ 16.4 M samples, respectively.

To the best of our knowledge, no such public, curated corpora are available for other languages. Synthetic parallel data however, can be generated easily from any corpus represented in standard orthography by removing all punctuation and capitalization, and by applying a few further transformations, including expanding abbreviations, replacing symbols (*+, -, %, &, @,* etc.) with words, and most importantly, converting digits to words. For the latter, we used `num2words`[3] in combination with regular expressions to verbalize cardinal and ordinal numbers, dates, times, and digit sequences like phone numbers. Our procedure for corpus preparation is not deterministic. Instead, it probabilistically produces different verbalizations like *two thousand twenty* and *twenty twenty* for *2020* in order to account for the fact that different verbalizations may map to the same orthography. For our German models, we use a sanitized version of the ParaCrawl corpus [7] as raw data. We further clean the corpus by harmonizing formats for dates, times and other entities. The final German training set consists of 36.5M samples. The process described here can be applied to any monolingual corpus of any size, and very few adaptations are required to apply it to other languages.

[2] `https://github.com/rwsproat/text-normalization-data`

[3] `https://github.com/savoirfairelinux/num2words`

Table 1: *Sample question from the survey. The full survey contains 50 questions like this.*

	Which variant do you prefer?
a)	*Freudig wird der Wahlsieg des ehemaligen grünen Chefs auch von estnischen Politikern kommentiert.*
b)	*Freudig wird der Wahlsieg des ehemaligen Grünen-Chefs auch von estnischen Politikern kommentiert.*
	What do you prefer in the selected variant?
a)	Punctuation
b)	Capitalization
c)	Digits or number words
d)	Something else
e)	Nothing

3.2. Architecture and training details

Our system architecture closely follows standard approaches for neural machine translation. We use BPE [20] with 10 000 merge operations for subword segmentation. We train small and large variants of the transformer model [21] implemented with fairseq [8]. Both models have equal encoders and decoders and shared embeddings. The small model has 3 layers in the encoder and decoder. It has an embedding dimension of 256, a feed-forward network embedding dimension of 1024, and 4 attention heads. It has 8 152 064 trainable parameters overall. The large model is very similar to the base transformer provided by fairseq. It has 6 layers each, an embedding dimension of 512, a feed-forward network embedding dimension of 2048, and 8 attention heads. It has 49 383 424 trainable parameters.

Both models employ pre-layer normalization, a dropout rate of 0.3 and an attention dropout rate of 0.1, and an activation dropout rate of 0.1. The maximum number per batch is 3 584. We train for 10 epochs with a learning rate of 0.001, inverse square root as a learning rate scheduler, and parameter updates after every 16 batches.

3.3. Evaluation metrics

Full text denormalization is still developing as a topic, and there are as yet no universally agreed upon standard evaluation metrics. Most works that follow the sequence labeling approach use precision and recall based metrics (e.g., [3]). In our sequence-to-sequence setting, we instead start with the well-established word error rate (WER) metric from ASR.

In addition to word error rate, we calculate error rates for specific subsets of output tokens, in particular 1) copy words (cWER), 2) punctuation marks (pWER), 3) digit sequences (dWER), and 4) capitalized words (uWER). We do this by removing all non-matching tokens from both the reference sequence and the hypothesis sequence and then calculating the word error rate as usual on the remaining token sequences. For instance, in a sentence pair like:

- *Und auch wer nur hier im Land operiert , zahlt nur zehn Prozent Steuern .* (reference)
- *Und auch wer nur hier im Land operiert zahlt nur 10 % Steuern .* (hypothesis)

the uWER is calculated on the sequences [*Und Land Prozent Steuern*] and [*Und Land Steuern*], and the pWER is calculated on [, .] and [.].

We define as punctuation marks all characters from the set {.,:;!?/-}. Capitalized words are defined as tokens which contain at least one uppercase character, and not necessarily the first

Table 2: *Experimental results for small and large English, German, and Russian models. Best values across languages in bold.*

	BLEU	**words (WER)**			**copy words (cWER)**			**punctuation (pWER)**			**digits (dWER)**			**capitalization (uWER)**		
		errors	tokens	rate	errors	tokens	rate	errors	tokens	rate	errors	tokens	rate	errors	tokens	rate
EN-S	82.0	10 425	117 171	8.90	1 944	56 216	3.46	3 659	17 658	20.72	542	7 742	7.00	5 430	34 556	15.71
EN-L	87.2	7 452	117 171	6.36	1 304	56 216	2.32	2 602	17 658	14.74	390	7 742	5.04	3 948	34 556	11.42
DE-S	84.8	16 778	214 097	7.84	1 707	108 745	1.57	5 776	24 435	23.64	556	4 830	11.51	8 661	74 288	11.66
DE-L	89.6	11 198	214 097	5.23	1 215	108 745	**1.12**	4 030	24 435	16.49	311	4 830	6.44	5 706	74 288	**7.68**
RU-S	84.8	9 938	134 842	7.37	1 918	75 205	2.55	4 471	22 917	19.51	520	8 649	6.01	3 511	27 391	12.82
RU-L	**90.0**	6 586	134 842	**4.88**	1 172	75 205	1.56	2 984	22 917	**13.02**	343	8 649	**3.97**	2 258	27 391	8.24

one. Digit tokens are all tokens which contain digits, and not necessarily exclusively digits. Finally, copy words are defined as all tokens which need only to be copied by the denormalizer. They are identified by extracting all tokens from the reference which contain only characters from the input character range.

Note that due to the removal of non-matching words from the sequence, these metrics can only evaluate whether the expected tokens appear in the correct sequence relative to each other. Whether two matching tokens actually appear in the exact same position in the full sequence is ignored in this case. These metrics thus primarily provide insight into a gross concept of punctuation completeness, but we rely on the overall WER to provide information about full sequence alignment.

Not all errors are captured by the specialized WERs. For instance, abbreviation expansion and replacement of specific words with symbols like % are only captured by the main WER. Nevertheless, Table 3 shows that the suggested error types indeed capture the most common phenomena.

Finally for completeness, we also report BLEU scores (calculated with `sacrebleu` [22]), however we note that, as an n-gram based metric this is not ideally suited to the denormalization task with its large number of simple copy operations; BLEU scores are high, but as a result provide less granular insight into actual performance.

In order to validate the proposed word error rates and evaluate the relative importance of different error types for human readers, we conduct a MOS-like survey involving 20 native German speaking participants who were asked to subjectively evaluate parts of the denormalizer output. In particular each participant was shown 50 German sentences where the denormalizer output contained one or more of the previously defined error types. Two variants of the same sentence were provided in random order: one of them a hypothesis from the denormalizer, and one an original reference orthography. Participants were asked to indicate which variant they preferred, and further to indicate the main reason that they preferred the variant that they selected. An example question is shown in Table 1. The survey focused on capitalization errors, punctuation errors, and number format errors. In order to try and tease apart the impact of multiple versus isolated errors, 20 sentence pairs were selected to contain only one of the three error types in question, while the others were permitted to contain a mixture of multiple error types.

Table 3: *Corpus statistics. Total number of tokens in the training set and percentage of main token types.*

	EN	**DE**	**RU**
# tokens	933M	783M	221M
copy tokens	48.07%	50.73%	55.79%
punctuation tokens	14.96%	11.38%	17.09%
digit tokens	6.65%	2.31%	6.36%
capitalized tokens	29.45%	34.78%	20.29%
remaining tokens	0.87%	0.80%	0.47%

4. Results

4.1. Experimental results

Detailed results for all models and WER types are reported in Table 2. The large models have consistently lower error rates for all languages and all token types compared to the corresponding small models. The word error rate is in the range of 4.88% (RU-L) to 6.36% (EN-L) for large models, and in the range of 7.37% (RU-S) to 8.90% (EN-S) for small models. Furthermore, the performance on German, 5.23% (DE-L), 7.84% (DE-S) suggests that our synthetic parallel data generation is competitive with the more carefully curated corpora from [11].

Roughly half of all tokens in the corpora are words which need only to be copied (see Table 3). Copy errors are rare, between 1.12% (DE-L) and 2.55% (RU-S). Copy errors are typically introduced when a token is wrongly capitalized, or when a number word is transformed into a digit yet the reference has a number word.

Punctuation tokens make up between 11.38% (German) and 17.09% (Russian) of all tokens in the corpus. Punctuation error rates are considerably higher than the overall word error rate, being in the range between 13.02% (RU-L) and 23.64% (DE-S).

On the contrary, digit tokens, which account for only 2.31% (German) to 6.65% (English) of all tokens, have error rates in a similar range like the overall word error rate, ranging from 3.97% (RU-L) up to 11.51% (DE-S).

Capitalized words are more prevalent in German (34.78%) than in English (29.45%) and Russian (20.29%). This is because all nouns are regularly capitalized in German. The respective error rates vary considerably from 7.68% (DE-L) up to 15.71% (EN-S). The good performance of the German models can be explained by the strict rules for capitalization in this language. On the contrary, the high error rates of the English model are partially caused by the fact that the training data con-

Table 4: *Selected examples of non-matching predictions (EN-L)*

Input	*ninth planet may exist beyond pluto scientists report*
Ref	*Ninth Planet May Exist Beyond Pluto, Scientists Report.*
Pred	*9th planet may exist beyond Pluto scientists report.*
Input	*you want fame*
Ref	*You want fame?*
Pred	*You want fame.*
Input	*archived from the original on the ninth of november two thousand nine*
Ref	*Archived from the original on 9 November 2009.*
Pred	*Archived from the original on 2009-11-09.*

tains samples with title case. In a future iteration, title casing should be normalized to normal casing in order to improve performance in English.

If all copy tokens as well as punctuation, capitalization and digit tokens are subtracted from the total number of tokens, only a small number of tokens is left (e.g., 0.87% of all tokens in English). This shows that the proposed token groups cover the vast majority of tokens. A particular subgroup of the remaining tokens are symbols like +, -, %, § etc. In a future iteration, these may be accounted for by including a specific word error rate for symbols.

Across languages, RU-L has the lowest error rates for WER, pWER, and dWER, while DE-L has the lowest error rates for cWER and uWER. We hypothesize that the English models have higher error rates despite larger amounts of generic training data due to the well known orthographic complexity of written English.

Although some error rates, particularly those for punctuation and capitalization, are relatively high, we observed in practice that many differences between the denormalizer output and the reference are either due to stylistic choice (e.g., `thirteen` vs. `13`) or caused by ambiguity (e.g., period vs. question mark in some cases). Selected samples of non-matching predictions which exhibit these phenomena are shown in Table 4. We conclude from this observation that it will be worth investing further time in curation of a clean test set which is consistent in itself and reflects the desired stylistic choices. On the topic of speed, we note that while the smaller models have slightly lower performance, they trade this for significantly faster inference speed, typically a factor 2x speedup compared to their larger variants.

4.2. Survey results

There were 20 participants rating 50 sentence pairs each. Of all 1000 answers, the reference was preferred in 60.0% of the cases, and there was no preference in 8.9% cases. Interestingly, the hypothesis was the preferred choice in 31.1% cases. This may indicate that, as discussed earlier, some denormalization 'errors' are actually interpreted as stylistic variants, e.g., a digit instead of a spelled out number, and shows the potential importance of aligning training data format with the target audience.

We focus on two different questions in our analysis. For the 20 sentence pairs with a single error type, we analyze how much human readers care about each of these error types. For the remaining 30 sentence pairs containing multiple error types, we analyze which error type human readers identify as the main issue when asked to choose a single option.

Table 5 reports the survey results for sentences with a single error type. For all three error types, participants rarely chose to rate both variants as being of equal quality. This is particularly clear in the case of punctuation and capitalization differences

Table 5: *Survey results for sentences with a single error type. 20 sentence pairs rated by 20 participants.*

	Issue noted in:					
hyp/ref differ in:	**hypothesis**		**reference**		**none/both**	
punctuation	141	78.3%	32	17.8%	7	3.9%
capitalization	108	60.0%	44	24.4%	28	3.3%
number formatting	19	47.5%	15	37.5%	6	15.0%
total	**268**	**67.0%**	**91**	**22.8%**	**41**	**10.3%**

Table 6: *Survey results for sentences with multiple error types. 30 sentence pairs rated by 20 participants.*

	Identified main issue in:			
hyp/ref differ in:	**punct**	**cap**	**digits**	**none/all**
punct + cap	53.3%	40.0%	0.0%	6.7%
punct + digits	32.5%	0.0%	50.0%	17.5%
cap + digits	4.0%	4.5%	78.5%	13.0%
all three	28.0%	3.5%	52.0%	16.5%

where only 3.9% and 3.3% of all answers report equal quality. This is a strong indication that human readers notice and do care about differences in punctuation and capitalization in particular. With punctuation and capitalization, the issue is located in the hypothesis in 78.3% and 60.0% of the cases which indicates that these are true errors in the sense that the denormalizer does not produce the preferred orthography.

The result is less clear for number formatting issues. While in 47.5% of the cases, the issue is located in the hypothesis, in as much as 37.5%, it is the reference which is seen as worse. Indeed, many mismatches of number formatting in the survey samples can be viewed as equally correct and more a matter of style. This shows that it is important to define the desired output formats.

Table 6 reports survey results for sentences with multiple error types. In sentences with all three error types, the respondents note the main issues in the following order of importance: digits (52.0%) > punctuation (28.0%) > capitalization (3.5%). The same relative importance is also reflected consistently for all combinations of only two error types. Interestingly, this inverts the token-type prevalence found in the German corpus: digits (2.3%) < punctuation (11.4%) < capitalization (34.7%). Furthermore, this relative prevalence of token types holds true across the other languages as well, suggesting that indeed, not all errors are created equal.

5. Conclusions

In this paper we have presented a simple sequence-to-sequence approach to fully restore orthography in raw, normalized speech transcripts, including prediction of punctuation, capitalization, and formatting of numbers, times, and dates. The desired transformations are implicitly derived from the training data.

Our models have a word error rate of 6.36% for English, 4.88% for Russian, and 5.23% for German. We further proposed more fine-grained WER-based metrics to evaluate different aspects of orthography restoration models, and help focus future work. Because the notion of metrics for full text denormalization in still unsettled, we also conducted a series of subjective human-driven evaluations to help identify which error types actually matter. We described a simple and scalable approach to quickly generate parallel synthetic data from any corpus in standard orthography and showed that it leads to models with a comparable quality as more carefully curated corpora. We make our English and Russian models publicly available along with the framework.

In future we plan to investigate more efficient customized architectures, focusing particularly on numerical forms and generic punctuation in response to our survey findings. We are also interested in incorporating turn-taking information from dialogues as a means to better model and predict sentence-end punctuation independent of acoustic features. Finally we would like to expand our subjective evaluations to continue the metric development discussion in this area.

6. References

[1] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling punctuation prediction as machine translation," in *International Workshop on Spoken Language Translation (IWSLT) 2011*, 2011.

[2] O. Tilk and T. Alumäe, "LSTM for punctuation restoration in speech transcripts," in *Proceedings INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, 2015.

[3] O. Tilk and T. Alumäe, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration," in *Proceedings Interspeech 2016 – 17th Annual Conference of the International Speech Communication Association*, 2016.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ICLR 2015*, 2014.

[5] M. Sunkara, S. Ronanki, K. Dixit, S. Bodapati, and K. Kirchhoff, "Robust prediction of punctuation and truecasing for medical ASR," in *ACL Workshop on NLP for Medical Conversations 2020*, 2020.

[6] M. Sunkara, C. Shivade, S. Bodapati, and K. Kirchhoff, "Neural inverse text normalization," 2021.

[7] M. Esplà, M. Forcada, G. Ramírez-Sánchez, and H. Hoang, "ParaCrawl: Web-scale parallel corpora for the languages of the EU," in *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation, Aug. 2019, pp. 118–119. [Online]. Available: https://www.aclweb.org/anthology/W19-6721

[8] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[9] P. Ebden and R. Sproat, "The kestrel TTS text normalization system," *Natural Language Engineering*, vol. 21, no. 3, 2015.

[10] R. Sproat and N. Jaitly, "RNN approaches to text normalization: A challenge," *ArXiv*, vol. abs/1611.00068, 2016.

[11] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, "Neural models of text normalization for speech applications," *Computational Linguistics*, vol. 45, no. 2, p. 293–337, 2019.

[12] E. Cho, J. Niehues, and A. H. Waibel, "Segmentation and punctuation prediction in speech language translation using a monolingual translation system," in *International Workshop on Spoken Language Translation (IWSLT) 2012*, 2012.

[13] W. Salloum, G. Finley, E. Edwards, M. Miller, and D. Suendermann-Oeft, "Deep learning for punctuation restoration in medical reports," in *BioNLP 2017*, 2017.

[14] B. Nguyen, V. B. H. Nguyen, H. Nguyen, P. N. Phuong, T. Nguyen, Q. T. Do, and L. C. Mai, "Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging," in *22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) 2019*, 2019.

[15] V. Pahuja, A. Laha, S. Mirkin, V. C. Raykar, L. Kotlerman, and G. Lev, "Joint learning of correlated sequence labeling tasks using bidirectional recurrent neural networks," in *Interspeech 2017*, 2017.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[17] M. Shugrina, "Formatting time-aligned ASR transcripts for readability," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[18] E. Pusateri, B. R. Ambati, E. Brooks, O. Platek, D. McAllaster, and V. Nagesha, "A mostly data-driven approach to inverse text normalization," in *Proc. Interspeech 2017*, 2017, pp. 2784–2788.

[19] J. Liao, S. E. Eskimez, L. Lu, Y. Shi, M. Gong, L. Shou, H. Qu, and M. Zeng, "Improving readability for automatic speech recognition transcription," *arXiv preprint*, 2020.

[20] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.

[22] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://www.aclweb.org/anthology/W18-6319