AUTOMATIC PUNCTUATION RESTORATION USING LANGUAGE MODELS AND PROSODIC FEATURES

Sahil Jayaram (UNI saj2163)

COMSE6998: Fundamentals of Speech Recognition, Columbia University

ABSTRACT

High-performing automatic speech recognition (ASR) systems can transcribe speech with impressive accuracy, in some cases achieving lower word error rates than human scribes on benchmark corpora. Few of these systems, however, are capable of producing output that contains punctuation, which has been shown to significantly increase the readability of transcribed speech. State-of-the-art approaches to automatic punctuation restoration rely solely on text input, ignoring the wealth of task-relevant information latent in the original audio. In this paper, I present a solution for punctuation restoration that is based on a top-performing, transformer-based method, differing from its precursor in that it can optionally utilize prosodic features in addition to text. My model achieved a combined F1 of 84.5 on the IWSLT 2012 TED Talk evaluation set, outperforming the current state-of-the-art model by 0.6. Furthermore, my model consistently achieved its best performance when given prosodic features, demonstrating that prosody is a valuable predictor of punctuation placement and is not rendered superfluous by text input.

***Index Terms*—** automatic punctuation, punctuation restoration

1. INTRODUCTION

ASR plays a key role in many domains, from education technology [11] to medical documentation [4] to voice-user interfaces (VUIs) [1]. For many applications, ASR output is intended to be read by humans. In such cases, automatic punctuation can be highly beneficial to the user; it has been found that removing punctuation from punctuated ASR transcripts significantly reduces text comprehensibility [14] and slows down reading comprehension [7].

Prosody, the "music" of language, is fundamentally linked to punctuation; in fact, writing systems evolved punctuation largely as cues to the prosodic qualities of an utterance in its written representation [5]. In spoken English, prosody is particularly useful for conveying phrase boundaries and differentiating between statements and questions. It is therefore no surprise that many of the earliest approaches to automatic punctuation relied on prosodic features [13].

In this paper, I focus on automatic punctuation restoration (APR), the subproblem of automatic punctuation in which punctuation is predicted from unpunctuated ASR output (non-restorative approaches to automatic punctuation may involve predicting punctuation and lexical content in conjunction [2]). The most successful APR systems compute predictions from text via a windowed approach. These methods make use of deep neural architectures such as bi-directional transformers [3] and deep recurrent neural networks with attention [8], outperforming earlier, less-parameterized systems, many of which used acoustic features in addition to text. This paper introduces an augmented version of one such state-of-the-art model, capitalizing on the record-breaking success of its transformer-based approach as well as the linguistically-explicable advantages of utilizing prosodic information. Specifically, in addition to text inputs, my approach uses features based on energy (a measure of the loudness of a signal) and fundamental frequency (F0; the acoustic correlate of pitch).

2. RELATED WORK

Courtland et al. demonstrated the high capacity of pretrained language models such as BERT as core components in a punctuation restoration system [3]. Their most successful architecture consists of RoBERTa$_{\text{base}}$ and two linear layers. Their model is essentially a sequence-to-sequence model, with text tokens as inputs and punctuation labels corresponding to `Period`, `Comma`, `Question`, or `None` as outputs. For any input sequence $(T_1, ..., T_n)$, output sequence $(Y_1, \ldots, Y_n)$, and $i \in \{1, \ldots, n\}$, Y_i is the predicted punctuation (or lack thereof) following token T_i. Predictions are performed in parallel over a sliding window, and the predicted probabilities for each token are aggregated across all windows containing the token. The model achieved an overall F1 of 83.9 on the IWSLT 2012 Ted Talk dataset [6], surpassing the previous record of 68.6 [8].

Class	IWSLT	NSC
Period	6.0%	11.5%
Question	0.3%	2.1%
Comma	5.4%	12.3%
None	88.3%	74.1%

Fig. 1. Class breakdowns for the IWSLT and NSC data used in my experiments

3. METHODS

3.1. Data

Although the IWSLT 2012 dataset contains enough punctuated text data to train a system that relies solely on text, it lacks clean transcription-audio pairs for training [6]. Thus, I pre-trained my model on data from the Singapore English National Speech Corpus (NSC) [9], which contains thousands of hours of speech audio paired with punctuated transcripts. I then tuned it on a subset of the IWSLT training data, using empty prosodic sequences. Two sets of data were used for evaluation: an NSC test set and the IWSLT 2012 test set.

The NSC consists of 4 parts. I only utilized data from Part 2, which contains prompted recordings of randomly-generated sentences with punctuated transcripts. I extracted data for 31 speaker-sessions (100,570 words in total). For each utterance, I used the highest-fidelity recording available (Channel 0). Each audio-transcript pair represents a single sentence. I filtered out pairs with duplicate transcripts, then concatenated all of the remaining data for each speaker-session, effectively treating all of the utterances performed by any given speaker in any given session as segments of one continuous utterance. 25 speaker-sessions were used for training, 5 for evaluation, and 1 for validation.

As part of the data preparation phase, prosodic features were computed for all 31 NSC speaker-sessions as well as for the IWSLT 2012 test set. I used the Gentle forced-aligner, an open source tool built on Kaldi [12], to align each audio file with its corresponding transcript at the word-level. For each audio segment (word), I used Kaldi to compute the signal's energy, normalized F0, delta F0, and delta-delta F0 over 25 ms windows, resulting in a variable-length sequence of prosodic feature vectors. Due to the limited performance capabilities of the forced alignment system, computing prosodic features failed for approximately 40% of words in the NSC data and 80% of words in the IWSLT test set. In such cases, the prosodic sequence is represented as an empty list. Gentle's low alignment rate for words in the IWSLT data likely owes to the nature of TED Talk recordings, which contain non-linguistic sounds such as applause and laughter and are considerably longer than the single-sentence recordings drawn from NSC.

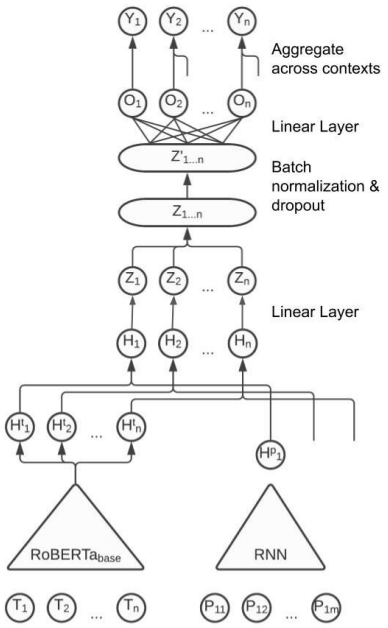

Fig. 2. The hidden representation H_i for token T_i with prosodic feature vectors $(P_{i1}, ..., P_{im})$ results from concatenating the token embedding H_i^t and the prosody embedding H_i^p. The remainder of the forward pass is copied from Courtland et al.'s design [3].

3.2. Architecture

My model architecture differs from that of Courtland et al. in that it contains an RNN component for generating fixed-length prosody embeddings from variable-length sequences of prosodic features. This RNN component consists of a single LSTM layer, where the final output of the layer is taken to be the prosodic encoding. Each prosody encoding is concatenated with the RoBERTa encoding of its corresponding token (Fig. 2). In consistency with Courtland et al.'s model, mine is trained on individual speech segments (windows) that are 100 tokens in length, and evaluated on entire speeches, with cross-window aggregation only occurring during evaluation. I used the same hyperparameter values and optimization scheme as the aforementioned authors, with the exception of batch size, which I reduced from 1,000 to 32. During training and evaluation, each prosodic sequence was left-cropped or zero-padded in order to achieve a length of 175.

3.3. Experiments

First, I trained the model for 3 epochs on the NSC training data and evaluated it on both test sets. I then tuned the model on 35,530 batches (1,136,960 windows) of IWSLT training data, and evaluated it on the IWSLT 2012 test set. Following Courtland et al., I measured the precision (P), recall (R),

Prosodic features hidden?	Period			Comma			Question			Combined		
	P	R	F	P	R	F	P	R	F	P	R	F
No	**95.4**	**94.0**	**98.5**	**94.6**	**96.4**	**97.7**	**95.0**	**95.2**	**98.1**	**96.8**	**97.0**	**96.9**
Yes	91.7	88.1	96.8	**94.6**	**96.4**	**97.7**	93.1	92.1	97.2	94.0	96.7	95.3

Fig. 3. Performance results achieved by my model on the NSC test data. Here, the model was untuned (trained only on NSC data). Row 2 shows the scores obtained when the model was evaluated with only empty prosodic inputs.

Model	Period			Comma			Question			Combined		
	P	R	F	P	R	F	P	R	F	P	R	F
Original [3]	**86.1**	**89.3**	**87.7**	76.9	**75.4**	76.2	**88.9**	**87.0**	**87.9**	**84.0**	83.9	83.9
Untuned (prosodic available)	39.8	31.8	14.3	45.9	21.3	25.8	42.6	25.5	18.4	45.5	47.4	45.7
Tuned (prosodic available)	85.8	65.0	85.2	**85.8**	73.2	**83.9**	85.8	68.9	84.6	83.7	**85.4**	**84.5**
Tuned (prosodic hidden)	85.6	64.7	85.2	85.7	73.3	**83.9**	85.7	68.8	84.6	83.7	85.3	84.4

Fig. 4. Performance results on the IWSLT 2012 test set. Shown are the scores reported by Courtland et al. (Row 1), the scores achieved by my model after being trained on only NSC data (Row 2), the scores achieved by my model after tuning it with IWSLT text data (Row 3), and the results of the same IWSLT-tuned model, evaluated with only empty prosodic inputs (Row 4).

Corpus	Unpunctuated	After punctuation restoration
NSC	`wong yoon wah kam kee yong and s r nathan lemper udang red tortoise steamed cake and mushroom tofu nothing beats having png tao in the summer chuan drive please show me the way to metta home for the disabled 364 how long will it take to walk to eu chin street what currency does timor-leste use`	`wong yoon wah, kam kee yong, and s. r. nathan. lemper udang, red tortoise steamed cake, and mushroom tofu. nothing beats having png tao in the summer. chuan drive. please show me the way to metta home for the disabled. 364. how long will it take to walk to eu chin street? what currency does timor-leste use?`
IWSLT 2012	`brings me to the beginning of my story 1996 when i gave my first tedalk rebecca was five years old and she was sitting right there in the front row i had just written a book that celebrated our life on the internet and i was about to`	`brings me to the beginning of my story. 1996, when i gave my first tedtalk. rebecca was five years old, and she was sitting right there in the front row. i had just written a book that celebrated our life on the internet, and i was about to`

Fig. 5. A demonstration of the model's performance on test data excerpts. The two examples shown were punctuated with the untuned model and with the tuned model, respectively. The NSC example was punctuated correctly, while in the IWSLT example, the model failed to predict a comma after the word 'tedtalk.'

and F1 (F) of the models overall *and* for each of the three classes of interest (`Period`, `Comma`, `Question`). Evaluation was performed with *and* without prosodic information (in the latter case, nonempty prosodic sequences were replaced with empty sequences).

4. RESULTS

My model achieved a combined F1 of 96.9 on the NSC test set (Fig. 3). After tuning it on IWSLT data, it achieved a combined F1 of 84.5 on the IWSLT 2012 test set, thus outperforming the previous state of the art model (Fig. 4). Notably, on both test sets, my model performed better when prosodic inputs were made available, although this margin was much greater for the NSC evaluation data. Fig. 5 demonstrates the performance of my end-to-end APR system.

5. DISCUSSION

5.1. Conclusions

The fact that my model outperforms the previous state of the art even when its prosodic inputs are empty sequences points primarily to the value of pretraining. However, its consistently higher performance when given nonempty prosodic inputs demonstrates the utility of prosodic features for the task of punctuation restoration.

It is noteworthy that prior to being evaluated on IWSLT (TED Talk) data, the model was tuned on same-domain data that lacked prosodic features. I hypothesize that tuning the model instead on TED Talk data that *does* contain prosodic features would increase performance on the IWSLT 2012 test set for the nonempty prosodic input case (in other words, it would boost the impactfulness of prosodic features).

Furthermore, prosodic features were only available for roughly 20% of all words in the IWSLT test set and 60% of all words in the NSC data (**3.1**). Using a more successful alignment procedure likely would have increased the model's performance when prosodic features were made available.

5.2. Future Directions

It is worth repeating the experiment using TED Talk training data that contains prosodic features, either in place of all training data (NSC and IWSLT) or in place of the IWSLT training data alone. Due primarily to time constraints, I was unable to extract and align the audio for the 100,000-word TED Talk training corpus, which is not as conveniently organized for prosodic feature extraction as the NSC. However, doing so would allow for a more meaningful comparison of my method with that of Courtland et al. [3].

Another useful modification to my method would be to align the IWSLT audio in a more effective manner. There exist tried-and-true solutions for performing forced alignment on very long audio segments [10]. Employing such an approach as part of the data extraction procedure could be extremely beneficial to model performance (**5.1**).

Another area worth exploring further is the selection of prosodic features. Rhythm, like pitch contour and energy, is deeply tied to punctuation [5]. The duration of the pauses preceding and following each word, as well as the duration of the word itself, could be included as features. Additionally, incorporating more linguistically-informed prosodic features such as phone-level or syllable-level stress could enhance model performance.

6. REFERENCES

[1] Aloufi, Ranya et al. "Privacy-preserving Voice Analysis via Disentangled Representations." *ArXiv* abs/2007.15064 (2020)

[2] Chen, C.. "Speech recognition with automatic punctuation." *EUROSPEECH* (1999).

[3] Courtland, Maury, et al. "Efficient Automatic Punctuation Restoration Using Bidirectional Transformers with Robust Inference." *Proceedings of the 17th International Conference on Spoken Language Translation*, July 2020

[4] Edwards, E. et al. "Medical Speech Recognition: Reaching Parity with Humans." *SPECOM* (2017).

[5] Fahnestock, Jeanne. "Prosody and Punctuation." (2011).

[6] Federico, M. et al. "Overview of the IWSLT 2012 evaluation campaign." *IWSLT* (2012).

[7] Jones, D. et al. "Measuring the readability of automatic speech-to-text transcripts." *INTERSPEECH* (2003).

[8] Kim, Seokhwan. "Deep Recurrent Neural Networks with Layer-wise Multi-head Attentions for Punctuation Restoration." *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019): 7280-7284.

[9] Koh, Jia Xin et al. "Building the Singapore English National Speech Corpus." *INTERSPEECH* (2019).

[10] Moreno, P. et al. "A recursive algorithm for the forced alignment of very long audio segments." *ICSLP* (1998).

[11] Mostow, J. et al. "Computer-Guided Oral Reading versus Independent Practice: Comparison of Sustained Silent Reading to an Automated Reading Tutor That Listens." *Journal of Educational Computing Research* 49 (2013): 249 - 276.

[12] Povey, Daniel et al. "The Kaldi Speech Recognition Toolkit." *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (2011).

[13] Sunkara, Monica et al. “Robust Prediction of Punctuation and Truecasing for Medical ASR.” *ArXiv* abs/2007.02025 (2020).

[14] Tündik, Máté Ákos et al. “User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning.” *INTERSPEECH* (2018).