UNIFIED MULTIMODAL PUNCTUATION RESTORATION FRAMEWORK FOR MIXED-MODALITY CORPUS

Yaoming Zhu, Liwei Wu, Shanbo Cheng, Mingxuan Wang

ByteDance AI Lab
{zhuyaoming, wuliwei.000, chengshanbo,wangmingxuan.89}@bytedance.com

ABSTRACT

The punctuation restoration task aims to correctly punctuate the output transcriptions of automatic speech recognition systems. Previous punctuation models, either using text only or demanding the corresponding audio, tend to be constrained by real scenes, where unpunctuated sentences are a mixture of those with and without audio. This paper proposes a unified multimodal punctuation restoration framework, named UniPunc, to punctuate the mixed sentences with a single model. UniPunc jointly represents audio and non-audio samples in a shared latent space, based on which the model learns a hybrid representation and punctuates both kinds of samples. We validate the effectiveness of the UniPunc on real-world datasets, which outperforms various strong baselines (*e.g.* BERT, MuSe) by at least 0.8 overall F1 scores, making a new state-of-the-art. Extensive experiments show that UniPunc's design is a pervasive solution: by grafting onto previous models, UniPunc enables them to punctuate on the mixed corpus. Our code is available at `github.com/Yaoming95/UniPunc`

***Index Terms*—** Speech, Punctuation Restoration, Multimodal

1. INTRODUCTION

Automatic speech recognition (ASR) has wide application and serves multiple tasks as an upstream component, like voice assistants and speech translations. Typically, ASR output unsegmented transcripts without punctuations, which may lead to misunderstanding for people [1], and affect performance of downstream modules, such as machine translation [2] and information extraction [3].

To address the issue, researchers proposed the automatic punctuation restoration task and designed a series of models. Conventional punctuation restoration models had achieved considerable progress [4, 5, 6, 7, 8], but they were solely based on the lexical information, which gives rise to some problems. One sentence may have varied punctuation, contributing to different meanings, respectively. For example, "I don't want anymore kids" means far away from "I don't want anymore, kids", suggesting the importance of a comma. Additionally, unaware of the speaker's tone, the model might find it difficult to determine whether a sentence should end in a full stop or a question mark.

Acoustic signals can help mitigate the ambiguity of punctuation models, considering rich information in speech audio such as pauses and intonation. Therefore, [9, 10, 11] proposed several multimodal punctuation restoration models. Generally, these multimodal models extracted acoustic features from speech audio and fused acoustic and lexical features by addition/concatenating. Their experiments validated that the speech audio of the text benefits punctuation.

Despite their effectiveness, previous multimodal systems faced with *modality missing* in real applications. Firstly, storage constraints or privacy policies may deny access to the corresponding audio, where previous multimodal systems failed to punctuate such audio-free sentences; Secondly, human-labeled punctuation audio is expensive and hard to obtain, which leads to sparser training sets for those multimodal models. Meanwhile, the audio-free unpunctuated text is readily available. Hence, an ideal punctuation model should utilize and punctuate both audio text and audio-free text.

To handle the modality missing problem, this paper proposes the UniPunc, a unified multimodal punctuation framework enabling punctuation on both audio and audio-free text. Specifically, for audio text, UniPunc first converts the two modalities into embedding by pretrained lexical model and acoustic model, while for the audio-free sentences, UniPunc introduces virtual embedding to simulate its audio. UniPunc then applies an attention-based module, named coordinated bootstrapper, to construct a cross-modal hybrid representation. Based on the hybrid representation, the model learns and predicts multimodal punctuation. The UniPunc combines the strengths of the previous unimodal and multimodal models: it can utilize a large amount of audio-free corpora as the training set and take advantage of acoustic input to reduce ambiguity in punctuation. Thus, the model is applicable to the punctuation of both audio texts and non-audio texts.

We highlight our main contributions as follows: *a)* We propose UniPunc, a new framework that serves punctuation

Yaoming and Liwei have equal contribution

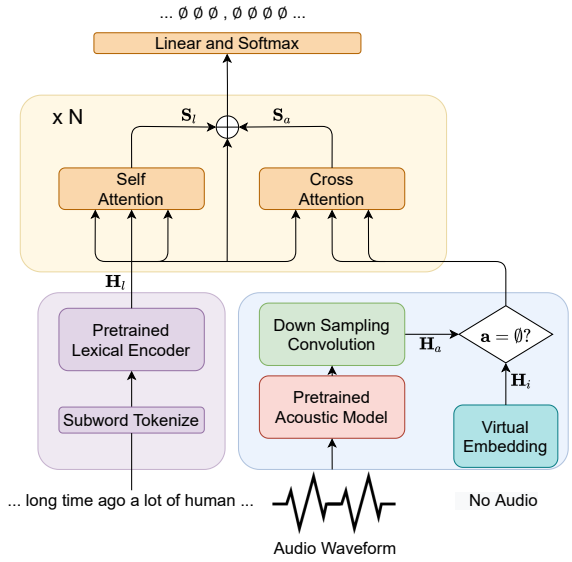

Fig. 1: Overall layout of UniPunc. Best viewed in color. Lexical encoder, acoustic assistant, and coordinate bootstrapper are blocked in purple, blue, and yellow respectively.

restoration task and utilizes both audio-free text corpus and audio speech transcripts. *b)* UniPunc achieves state-of-the-art performance on multimodal punctuation and outperforms various strong baselines by at least 0.8% overall F1 scores on real-world datasets. *c)* We discuss pervasiveness UniPunc: the introduced framework can be grafted into related work [5, 12], empowering them to process both audio and non-audio text while improving performance.

2. PROBLEM FORMULATION

Punctuation restoration is often modeled as a sequence labeling task [13]. Generally, the corpus of multimodal punctuation is a set of sentence-audio-punctuation triples and denoted as $S = \{\mathbf{x}, \mathbf{a}, \mathbf{y}\}$. Here $\mathbf{x}$ is an unpunctuated sentence of length T, and $\mathbf{a}$ is the corresponding speech audio. The model should predict the sequence of punctuation $\mathbf{y}$ given $\mathbf{x}$ and $\mathbf{a}$. The length of punctuation $\mathbf{y}$ is identical to the unpunctuated sentence $\mathbf{x}$ due to the nature of the sequence labeling task, *i.e.*, $|\mathbf{x}| = |\mathbf{y}|$. This study focuses on four types of punctuation, namely comma(,), full stop(.), question mark(?), and no punctuation.

As we have mentioned, our training set is a mix of audio text and audio-free text: on the one hand, we want to leverage large amounts of plain text in the absence of audio; on the other hand, we wish to make full use of acoustic features in the presence of audio. So is the case for the test set since we do not always get audio of the ASR transcription to be punctuated. Hence, the main challenge lies in modality missing for both training samples and evaluation data: *i.e.*, $\exists\ \mathbf{a} = \emptyset$.

3. METHOD

This section will introduce the proposed UniPunc in detail. UniPunc consists of three main components: lexicon encoder, acoustic assistant, and coordinate bootstrapper. The three components learn lexical features, possible acoustic properties, and cross-modal hybrid features, respectively. Fig. 1 shows the overall layout.

3.1. Lexical Encoder

The lexical encoder tokenizes unpunctuated sentence into subword sequence and converts it into the contextual embedding sequence $\mathbf{H}_l = (h_1^l, ..., h_n^l)$. We leverage pretrained NLP model, e.g, BERT [14] as our backbone model. We treat the pretrained model as a sub-module of UniPunc and finetune the model on our task-specific data.

3.2. Acoustic Assistant

The acoustic assistant consists of an acoustic feature extractor, a down-sampling network, and a virtual embedding. It converts possible speech audio $\mathbf{a}$ into acoustic embeddings $\mathbf{H}_a$ or virtual embedding $\mathbf{H}_i$.

For the audio-annotated transcriptions, UniPunc converts the audio signal into acoustic features via a pretrained acoustic feature extractor. Generally, the acoustic feature extractor is first pretrained on unlabelled audio datasets through self-supervised training and is then used as a submodule of the whole model and finetuned on the downstream punctuation task. Then UniPunc applies a down-sampling network further to shorten the extracted acoustic features' length and get acoustic embedding $\mathbf{H}_a = (h_1^a, ..., h_m^a)$. The purpose is to make the length of acoustic embeddings close to the sentence embedding so that the model can better align cross-modal information. We choose a multilayer convolutional network as the core component of the down-sampling network.

For the audio-free sentences, we propose to use virtual embedding $\mathbf{H}_i$ simulating the possible missing acoustic feature, *i.e.*, $\mathbf{H}_a = \mathbf{H}_i$ **if** $\mathbf{a} = \emptyset$. We set virtual embedding as a fixed-length array of learnable parameters and expect it to learn the representations of absent audio. In the training process, the model learns the shared latent space of acoustic and virtual embedding, which supplies acoustic information for the subsequent coordinate bootstrapper.

3.3. Coordinate Bootstrapper

We propose the coordinate bootstrapper for jointly training audio-free and audio text to overcome the missing modality problem. The coordinate bootstrapper jointly exploits the acoustic and lexical features and applies an attention-based operation to learn a hybrid representation across two modalities.

Specifically, UniPunc first conducts self-attention on the contextual embedding $\mathbf{H}_l$ to capture the long-range dependencies $\mathbf{S}_l$ within the unpunctuated sentence and apply cross-attention between the contextual embedding $\mathbf{H}_l$ and acoustic embedding $\mathbf{H}_a$ to formulate a cross-modal representation $\mathbf{S}_a$:

$$\mathbf{S}_l = \mathrm{Att}(\mathbf{H}_l, \mathbf{H}_l, \mathbf{H}_l) \quad (1)$$

$$\mathbf{S}_a = \mathrm{Att}(\mathbf{H}_a, \mathbf{H}_a, \mathbf{H}_l) \quad (2)$$

where $\mathrm{Att}(q, k, v) = \mathrm{softmax}(\frac{qk^\top}{\sqrt{d_k}})v$ is the attention operation proposed by [15], and d_k is the dimension size of the model. Note that for modality missing samples, we substitute the acoustic embedding $\mathbf{H}_a$ with virtual embedding $\mathbf{H}_i$, in which case $\mathbf{S}_a = \mathrm{Att}(\mathbf{H}_i, \mathbf{H}_i, \mathbf{H}_l)$ **if** $\mathbf{a} = \emptyset$.

Then the UniPunc acquires hybrid representation $\mathbf{H}_h$ by adding the attended representation along with a residual connection [16].

$$\mathbf{H}_h = \mathbf{S}_l + \mathbf{S}_a + \mathbf{H}_l \quad (3)$$

Like other attention blocks, the coordinate bootstrapper can be stacked to multiple layers to increase the model capacity further.

Finally, the output classifier layer consists of a linear projection and a softmax activation function. UniPunc inputs $\mathbf{H}$ to the classifier layer to predict the punctuation sequence $\hat{\mathbf{y}}$.

In this way, we enable the representations of audio samples and no audio samples to share the same embedding space. Thus, the model can receive mixed data in the same training batch, and the trained model is able to punctuate both audio text and audio-free text.

It should be emphasized that UniPunc makes a pervasive framework in solving modality missing in multimodal punctuation restoration tasks. For the previously proposed punctuation model, a proper adaptation of the proposed coordinate bootstrapper and acoustic assistant can empower them to address modality missing samples. Section 5 discuss the the pervasiveness of coordinate bootstrapper upon previous models [5, 12] via experiment.

4. EXPERIMENT

4.1. Datasets

We conduct our experiment mainly on two real-world corpus: MuST-C [17][1] and Multilingual TEDx(mTEDx) [18][2], whose audio are both originated from TED talks. We constructed three sets of data based on two corpora: *1)* English-Audio: This set contains the English audio and sentences in MuST-C. Each sample is with audio. *2)* English-Mixed: This set contains all English audio sentences and audio-free sentences from two corpora. Note that English-Audio is a subset of English-Mixed.

[1] https://ict.fbk.eu/must-c/

[2] http://www.openslr.org/100/

Table 1: Statistical information of the training/test data on the two sets, where the first two rows are of the training set. The *sent. len* and *audio len.* denotes the *average* length of sentence and audio counted by words and seconds, respectively.

	# of sent.	# of audio	sent. len.	audio len.
English-Audio (Train)	99381	99381	42.5	14.7
English-Mixed (Train)	142441	99381	44.1	14.7
English-Audio (Test)	490	120940	54.7	15.7
English-Mixed (Test)	2298	120940	52.0	15.7

We re-partition the data for each sample corresponding to an audio duration roughly range from 10 to 30 seconds. Tab. 1 shows statistical information of the text and audio about the training and test sets in detail.

4.2. Configurations and Baselines

We set most modules at a dimension of 768. We choose BERT [14] and its subword tokenizer as pretrained lexical encoder and Wav2Vec 2.0 [21][3] as pretrained acoustic model. For two English datasets, we use BERT-base uncased and Wav2Vec 2.0-base no tuned version. We use a two-layer Convolution network of stride 5, kernel size of 15 as the downsampling network. The layer number of coordinate bootstrapper is 2. The sequence length of virtual embedding is 5. We use learning rate of 0.00001 with Adam [22], dropout rate of 0.1, and Noam learning rate scheduler of warm-up step 8000.

We compare the performance of UniPunc with various baselines and SOTA systems, including: LSTM-T [13], Att-GRU [20], BERT [7], SAPR [8], Self-Att-Word-Speech[10] and MuSe [11]. We also compare to TTS punctuation data augmentation [19].

All unimodal baselines are trained on English-Mixed corpus, while multimodal ones are trained on English-Audio since they cannot handle audio-free samples. For UniPunc, we first trained it jointly on English-Mix, denoted as UniPunc-Mixed. To facilitate comparison with other multimodal baselines, we trained a UniPunc variant solely on the English-Audio, denoted as UniPunc-Audio. All models and baselines are evaluated in terms of precision(P), recall(R), and F1 score(F1) on three punctuations.

We implement UniPunc by Fairseq [23]. All experimental data are publicly available online. We will release our code and data split after paper acceptance for reproducibility.

5. RESULTS

5.1. Main Results

We report the performance of our model and other baselines on English datasets in Tab. 2.

[3] The pretrained acoustic model used in the MuSe was Wav2Vec, which we also replace it with a newer and better version of Wav2Vec 2.0 in our implement for a fair comparison with UniPunc.

Table 2: Results for three punctuations on two test set.

Test Set	Model	Comma			Full Stop			Question Mark			Overall		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
English-Audio	LSTM-T [13]	68.8	50.7	56.9	75.8	74.9	74.3	29.8	26.5	27.2	72.0	60.0	65.1
	Self-Att-Word-Speech [10]	71.1	58.1	62.7	81.1	76.4	77.5	31.9	30.1	31.0	75.5	65.3	69.7
	BERT [7]	79.1	66.9	72.1	84.5	81.7	82.8	78.4	75.7	75.5	81.5	73.5	77.1
	MuSe [11]	78.5	68.7	73.2	83.0	84.5	83.6	81.4	79.7	79.4	80.6	75.4	77.9
	MuSe [11]+TTS [19]	75.3	78.9	77.1	80.0	77.6	78.8	78.0	80.9	79.5	75.6	79.9	77.8
	UniPunc-Audio (Ours)	78.9	70.2	74.2	81.7	86.0	83.7	84.3	80.1	80.8	80.3	76.8	78.5
	UniPunc-Mix (Ours)	69.9	79.7	74.5	85.1	82.6	83.8	79.0	81.9	80.4	76.3	81.0	78.6
English-Mixed	BiLSTM [5]	61.0	45.0	51.6	59.5	54.7	56.8	58.8	47.7	50.1	60.5	48.3	53.7
	Att-GRU [20]	61.1	47.5	52.0	75.7	66.9	69.6	27.3	26.5	26.0	67.1	55.5	60.3
	SAPR [8]	72.1	58.6	64.4	78.3	75.9	76.9	76.5	56.9	63.7	74.8	65.3	69.7
	BERT [7]	74.2	68.0	70.8	82.6	81.2	81.8	78.7	79.5	77.4	78.0	73.6	75.7
	UniPunc-Mix (Ours)	73.2	71.2	72.1	82.5	82.2	82.2	79.0	76.0	76.1	77.3	75.8	76.5

Table 3: F1 score for pervasive experiment on English-Mixed test set. CD denotes content dropout.

	Comma	Full Stop	Question Mark	Overall
BiLSTM [5]	51.6	56.8	50.1	53.7
BiLSTM-UniPunc	55.7	58.4	46.5	56.6
CD [12]-BERT	71.4	81.9	69.7	76.1
CD-UniPunc	71.5	82.1	76.7	76.9

Compare with other multimodal models: We compare UniPunc with other baselines on the English-Audio test set and get the following main conclusions: 1. UniPunc-mixed's overall F1 score exceeds all baselines by at least 0.7 points, indicating the effectiveness of our framework. 2. UniPunc-Audio also outperforms all multimodal baselines, suggesting that even without introducing extra audio-free corpus, UniPunc makes a strong model on multimodal sequence labeling; 3. The overall F1 score of UniPunc-Mix is slightly higher than that of UniPunc-Audio, indicating that training with audio-free sentences indeed improves the performance, especially in punctuating commas and full stops. 4. Using audio-free text mainly improves the recall of punctuation (compare Muse *vs*. Muse+TTS, UniPunc-Audio *vs*. UniPunc-Mixed)

Compare with other unimodal models: We also compare UniPunc-mixed and other unimodal punctuation models on the English-Mixed test set; we also list BERT's performance on English-Audio as it is the strongest unimodal baseline. Our conclusions are twofold: 1. UniPunc outperforms all baselines with at least 0.8 overall F1 scores on English-Mixed, which proves that its learned hybrid representation is very effective; 2. UniPunc far outperforms BERT on English-Audio 1.5 overall F1 score, which shows that UniPunc effectively represents the acoustic features in speech, which is especially obvious for punctuating question marks.

We also conduct a case study and evaluate UniPunc's performance on multilingual data from mTEDx. We find UniPunc punctuates closer to humans, and can better tell the pauses of commas and fulls stops as well as the tone of questions. In addition, we also find that UniPunc has better punctuation performance than other baselines on multilingual punctuation, which suggests that UniPunc has better robustness and generalization. The examples of case study is available at our code base.

5.2. Pervasiveness of UniPunc

This subsection explores the pervasiveness of UniPunc framework for punctuation restoration on the mixed datasets by grafting onto two previous approaches, namely BiLSTM [5] and content dropout [12]. Specifically, for the BiLSTM model, we introduce an acoustic assistant to extract potential acoustic features and coordinate bootstrapper to learn hybrid representation. We also graft UniPunc with content dropout and compare it with the lexical content dropout BERT. Tab. 3 shows the overall results.

The experiments show that our framework is pervasive for solving the modality missing in punctuation, and enables the previous unimodal model to handle multimodal corpus. By grafting UniPunc onto the BiLSTM, our module greatly improves the performance of the original model by 2.9 F1 scores. In particular, when UniPunc and content dropout are used jointly, the model achieves an F1 score of 76.9, further improving the overall performance.

6. CONCLUSION

This paper focuses on the modality missing problem in multimodal punctuation tasks, as unpunctuated sentences are a mixture of audio text and audio-free text in real applications. We devise a new unified multimodal punctuation framework, named UniPunc. UniPunc can learn a hybrid representation for both audio and audio-free corpus, based on which UniPunc allows punctuating both kinds of sentences. We experiment on two real-world datasets, and find that UniPunc surpasses all previous strong multimodal and unimodal baselines by at least 0.8 overall F1 scores. Extensive experiments show that our framework is pervasive and can empower other models to process modality missing sentences.

7. REFERENCES

[1] Máté Ákos Tündik, György Szaszák, Gábor Gosztolya, and András Beke, "User-centric evaluation of automatic punctuation in ASR closed captioning," in *Interspeech 2018*, B. Yegnanarayana, Ed. 2018, pp. 2628–2632, ISCA.

[2] Vincent Vandeghinste, Lyan Verwimp, Joris Pelemans, and Patrick Wambacq, "A comparison of different punctuation prediction approaches in a translation context," in *EAMT 2018*, pp. 269–278.

[3] John Makhoul, Alex Baron, Ivan Bulyko, Long Nguyen, Lance A. Ramshaw, David Stallard, Richard M. Schwartz, and Bing Xiang, "The effects of speech recognition and punctuation on information extraction performance," in *INTERSPEECH 2005*, pp. 57–60.

[4] Agustin Gravano, Martin Jansche, and Michiel Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *ICASSP*. IEEE, 2009, pp. 4741–4744.

[5] Ottokar Tilk and Tanel Alumäe, "Lstm for punctuation restoration in speech transcripts," in *INTERSPEECH 2015*, 2015.

[6] Maury Courtland, Adam Faulkner, and Gayle McElvain, "Efficient automatic punctuation restoration using bidirectional transformers with robust inference," in *IWSLT 2020*, pp. 272–279.

[7] Karan Makhija, Thi-Nga Ho, and Eng-Siong Chng, "Transfer learning for punctuation prediction," in *APSIPA ASC, 2019*. IEEE, pp. 268–273.

[8] Feng Wang, Wei Chen, Zhen Yang, and Bo Xu, "Self-attention based network for punctuation restoration," in *ICPR 2018*, pp. 2803–2808.

[9] Ondřej Klejch, Peter Bell, and Steve Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *ICASSP 2017*. IEEE, 2017, pp. 5700–5704.

[10] Jiangyan Yi and Jianhua Tao, "Self-attention based model for punctuation prediction using word and speech embeddings," in *ICASSP 2019*. 2019, pp. 7270–7274, IEEE.

[11] Monica Sunkara, Srikanth Ronanki, Dhanush Bekal, Sravan Bodapati, and Katrin Kirchhoff, "Multimodal semi-supervised learning framework for punctuation prediction in conversational speech," in *Interspeech 2020*, 2020, pp. 4911–4915.

[12] Andrew Silva, Barry-John Theobald, and Nicholas Apostoloff, "Multimodal punctuation prediction with contextual dropout," in *ICASSP 2021-2021*. IEEE, 2021, pp. 3980–3984.

[13] Piotr Zelasko, Piotr Szymanski, Jan Mizgajski, Adrian Szymczak, Yishay Carmiel, and Najim Dehak, "Punctuation prediction model for conversational speech," in *Interspeech 2018*. 2018, pp. 2633–2637, ISCA.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*, 2019, pp. 4171–4186.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS 2017*, 2017, pp. 5998–6008.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR 2016*, 2016, pp. 770–778.

[17] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," *Comput. Speech Lang.*, vol. 66, pp. 101155, 2021.

[18] Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post, "The multilingual tedx corpus for speech recognition and translation," *CoRR*, vol. abs/2102.01757, 2021.

[19] Daria Soboleva, Ondrej Skopek, Márius Šajgalík, Victor Cărbune, Felix Weissenberger, Julia Proskurnia, Bogdan Prisacari, Daniel Valcarce, Justin Lu, Rohit Prabhavalkar, et al., "Replacing human audio with synthetic audio for on-device unspoken punctuation prediction," in *ICASSP 2021*, pp. 7653–7657.

[20] Seokhwan Kim, "Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration," in *ICASSP 2019*. IEEE, 2019, pp. 7280–7284.

[21] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS 2020*.

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[23] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "FAIRSEQ: A fast, extensible toolkit for sequence modeling," in *NAACL demo*, 2019, pp. 48–53.