Mining Word Boundaries from Speech for Cross-domain Chinese Word Segmentation

Anonymous ACL submission

Abstract

Inspired by early research on exploring naturally annotated data for Chinese Word Segmentation (CWS), and also by recent research on integration of speech and text processing, this work for the first time proposes to explicitly mine word boundaries from parallel speech-text data. We employ the Montreal Forced Aligner (MFA) toolkit to perform character-level alignment on speech-text data, giving pauses as candidate word boundaries. Based on detailed analysis of collected pauses, we propose an effective probability-based strategy for filtering unreliable word boundaries. To more effectively utilize word boundaries as extra training data, we also propose a robust complete-then-train (CTT) strategy. We conduct cross-domain CWS experiments on two target domains, i.e., ZX and AISHELL2. We have also annotated about 900 sentences as the evaluation data of AISHELL2. Experiments demonstrate the effectiveness of our proposed approach.

1 Introduction

As a fundamental task in Chinese language processing, CWS aims to segment an input character sequence into a word sequence, since words, instead of characters, are the basic meaning unit in Chinese. Figure 1 gives an example of the CWS task, along with the speech signals.

With the rapid progress of deep learning techniques, especially the proposal of pre-trained language models like BERT (Devlin et al., 2019), CWS models have achieve very high performance when there is abundant training data from the same domain with the test data (Tian et al., 2020; Huang et al., 2020b). Therefore, recent studies on CWS pay more attention to the cross-domain scenarios (Huang et al., 2020a; Ke et al., 2021).

Meanwhile, considering the high cost of manually annotating high-quality CWS data, it had been an attractive research direction to explore naturally annotated CWS data from different channels. For instance, anchor texts in HMLT-format web documents imply reliable word boundaries (Jiang et al., 2013; Yang and Vozila, 2014); domain-aware dictionaries can match words accurately in target domain texts (Liu et al., 2014). These studies illustrate that such information can be used as partial annotations for training CWS models.

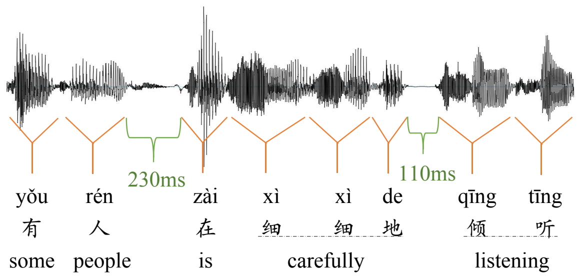

Figure 1: An example of speech-text alignment data. The correct segmentation result is "有/人/在/细细/地/倾听", translated as "some people is carefully listening".

Another interesting research line in recent years is the multi-modal integration of speech and texts, mainly due to the adoption of unified model architectures in both speech processing (Baevski et al., 2020; Hsu et al., 2021) and NLP fields (Devlin et al., 2019; Lewis et al., 2020) in the deep learning era. These approaches can be broadly divided into three categories, i.e., 1) using speech as extra features for NLP (Zhang et al., 2021), 2) multi-task learning (MTL) with cross-attention interaction (Sui et al., 2021), and 3) end-to-end language analysis from speech (Chen et al., 2022). Among these, a work (Zhang et al., 2021) is closely related with ours. They extract extra features from speech to enhance CWS on corresponding texts.

Following previous Inspired by progress of research directions discussed above, we propose for the first time to explicitly utilizes pauses in speech as word boundary annotations. The basic motivation is that when uttering a Chinese sentence, people often pause after finishing some complete meaning in the middle of the sentence, to breath or to make the speech easier to understand. Considering that words are the basic meaning unit, we hypothesize that the pause information can be utilized to help CWS.

Following previous works on cross-domain CWS, we employ Chinese Penn Treebank 5 (CTB5) (Xue et al., 2005) as the source domain, and use the widely used ZhuXian ("Jade Dynasty" in English, abbreviated as ZX) data as the target domain (Zhang et al., 2014). We collect and clean the parallel speech-text corpus of ZX for mining word boundaries. To more thoroughly evaluate the models, we use AISHELL2 as the second target domain, which is a publicly available data for automatic speech recognition (ASR) (Du et al., 2018). The contributions of our work are as follows.

- We have manually annotated about 900 sentences as the dev/test evaluation data for the AISHELL2 domain.

- We employ the MFA toolkit (McAuliffe et al., 2017) to perform character-level alignment on speech-text corpora, and conduct detailed analysis on the collect pauses.

- We propose an effective probability-based strategy for filtering unreliable word boundaries, and a robust CTT strategy to make use of the word boundaries as naturally annotated data.

- Experiments on both ZX and AISHELL demonstrate the effectiveness of our proposed approach.

Our code and newly annotated data will be released at `github`.

2 Mining Word Boundaries from Speech

This section describes how we collect speech pauses from parallel speech-text data, which consists of two steps. First,we prepare parallel speech-text data. Second, we utilize a GMM-HMM based model to obtain character-level speech-text alignments. Based on the alignments, we can obtain the pause duration between characters. Finally, we conduct detailed analysis on pauses and propose a simple filtering strategy to keep reliable pauses as word boundaries.

Corpus	Item	Train	Dev	Test
CTB5	# Sent	18,104	352	348
	# Word	493,932	6,821	8,008
ZX	# Sent		788	1,394
	# Word		20,393	34,355
AISHELL2 (Annotated)	# Sent		300	581
	# Word		2,091	3,821
Speech-text Data			**# Pause**	**# Sent**
ZX	all		—	25,038
	containing pause		203,842	25,016
	after filtering ($p^{\mathrm{B}} \geq 0.1$)		198,361	25,007
	after filtering ($p^{\mathrm{B}} \geq 0.9$)		197,540	24,964
AISHELL2	all		—	847,662
	containing pause		537,986	324,577
	after filtering ($p^{\mathrm{B}} \geq 0.1$)		457,007	294,694
	after filtering ($p^{\mathrm{B}} \geq 0.9$)		442,633	286,608

Table 1: Statistics of data used in our experiments. p^{B} means the probability threshold for filtering pauses.

2.1 Preparing Speech-Text Parallel Data

In this work, we use CTB as the source domain and employ two target-domain datasets. Table 1 shows the data statistics.

(1) ZX. The first dataset is the ZhuXian (ZX) dataset for the web fiction domain, which was constructed by Zhang et al. (2014) and has been widely used in previous works on cross-domain word segmentation (Liu and Zhang, 2012; Ding et al., 2020; Jiang et al., 2021)

The ZX dataset contains about 5K sentences in total.[1] The ZhuXian fiction consists of about 30K sentences in total. In this work, we manage to derive word boundaries from speech for the remaining sentences that are not included in ZX-dev/test.

There are several versions of audio books for the ZhuXian fiction on the Internet, in which some reader reads aloud the texts. We select one version that is of high quality and has little background music. All audios are processed to be at a sampling frequency of 16kHz.

Cleansing. We apply several data cleansing or filtering strategies to improve the data quality. (1) Numbers like "1200" are transformed into their Chinese character form like "一千两百" (one thousand and two hundred). (2) Silent and special symbols in the texts like punctuation marks are removed. (3) Irrelevant blanks or noises in the

[1] Among them, 2,373 sentences are reserved for training, but usually are not used in cross-domain experiments.

beginning or end of the audio are removed. (4) Audios with background music are discarded. Finally, we collect 246 audio files amounting to 144 hours, each corresponding to a chapter of the fiction.

(2) AISHELL2. For the second domain, we adopt the AISHELL2 (Du et al., 2018) mandarin Chinese speech corpus, which contains about 1,000 hours of high-quality audio, corresponding to about one million transcription sentences.[2] The corpus covers 12 different domains that are closely related with application of speech recognition in smart home, autonomous driving, industrial production, etc.

One major feature of the AISHELL2 data, whose major use is as training data for ASR, is that the transcription texts do not contain punctuation marks. In fact, outputs of ASR models usually do not contain soundless symbols in written texts, including punctuation marks.

Instead of injecting punctuation marks in AISHELL2 transcription texts, which would be highly time-consuming and prone to annotation errors, we decide to perform word segmentation on transcription texts directly. We believe this is an interesting and useful scenario for word segmentation research. Text normalization procedures such as filling punctuation marks may be applied over the output word sequence.

To alleviate the mismatch between the AISHELL2 data and the source-domain training data, i.e., CTB, regarding punctuation marks, we employ a simple strategy that can boost the performance of the baseline model by large margin. For each sentence in CTB-train, we remove the punctuation marks in the sentence. With this strategy, the trained model can handle transcription texts well.

To evaluate the model on AISHELL2, we have manually annotated about 900 sentences in the original AISHELL2-dev/test, and use them as the dev/test evaluation datasets. We present more details about data annotation in Section 4.1.

2.2 Character-level Speech-Text Alignment

In this paper, we try to derive word boundaries from speech based on pause information. The intuition is that if the speaker pauses for some time after uttering a char, then there may be a word boundary after the char. The key challenge for implementing this idea is how to obtain accurate character-level alignments between speech signals and the corresponding sentence.

In the past decade, end-to-end Transformer based models have become the dominate ASR approach due to its superior performance (Gulati et al., 2020; Zhang et al., 2023; Pratap et al., 2023). With an extra connectionist temporal classification (CTC) component, the model can explicitly produce alignment. However, our early experiments reveal that the Transformer-CTC based models suffer from a severe peak alignment issue, meaning that every character is usually aligned to a single speech frame, leaving most of the frames aligned to blanks. This finding is consistent with previous results (Senior et al., 2015; Zeyer et al., 2021).

Instead, we employ the MFA toolkit with its GMM-HMM implementation to obtain alignment between text and speech (McAuliffe et al., 2017). We employ both monophone and triphone GMMs.

Given a speech, we use the default frame window length of 25ms and the default frame offset of 10ms. For each frame, the acoustic features are the standard Mel-frequency cepstral coefficients (MFCCs). Formally, we represent speech as $\mathbf{x} = x_0...x_i...x_n$, where x_i is an MFCC feature vector, and the corresponding transcription as $\mathbf{y} = y_0...y_i...y_m$, where y_i denotes a token. The objective of GMM-HMM is two fold: 1) to determine which phonemes correspond to a token, and 2) to determine which frames (e.g., $x_k...x_l$) correspond to a phoneme. Combining the results, we can obtain the time range for each token. The model works under the unsupervised scenario, and apply the expectation-maximization (EM) algorithm (Moon, 1996) on the training speech-text pairs.

We continue training the pre-trained mandarin model in the MFA toolkit[3] using our parallel speech-text data at hand, either ZX or AISHELL2. In our context, a token y_i corresponds to a character.[4] Suppose y_i is aligned to $x_{b_i}...x_{e_i}$, also denoted as (b_i, e_i), where b_i and e_i are the beginning and end indices of frames. Then we can calculate the pause duration between two adjacent characters,

[2] We sincerely thank the Beijing AISHELL Technology Co., Ltd for sharing the data.

[3] https://mfa-models.readthedocs.io/en/latest/acoustic

[4] By default, the mandarin model in the MFA toolkit can only perform alignment at the word level, since the acoustic dictionary is word-based and polyphonic characters only have one entry, corresponding to the most frequent pronunciation. To handle this issue, we extend the acoustic dictionary by leveraging a Pinyin-based Chinese lexicon (both words and characters). We will release the related resource and the scripts.

for instance y_i and y_{i+1} as follows.

$$d(y_i, y_{i+1}) = (b_{i+1} - e_i) \times \mathit{10ms} \qquad (1)$$

Figure 1 gives an example. There are two pauses in the sentence, with duration of *230ms* and *110ms* respectively.

2.3 Filtering Pauses

At the beginning, our plan was to filter unreliable word boundaries based on a global pause duration threshold. For instance, if $d(y_i, y_{i+1}) < \mathit{50ms}$, then we discard the pause and not consider it as a boundaries. In other words, we only keep pauses with $d(y_i, y_{i+1}) \geq \mathit{50ms}$ as boundaries. However, our analysis shows that pauses with short duration are equally helpful.

Then we turn to another simple probability-based filtering strategy. The idea is to let the baseline model trained on the source-domain data (i.e., CTB) to judge. If the baseline model has a very small probability to support the boundary, then we discard it.

Following previous works, we adopt the BERT-based CRF model as our baseline model, and employ the label set of {B, M, E, S}, meaning "beginning", "middle", "end", and "single-char", respectively. Given an input char sequence $\mathbf{y} = y_0...y_m$, we denote a label sequence as $\mathbf{z} = z_0...z_m$. The marginal probability of a label bigram at given positions i and $i+1$, for instance E_S, is:

$$p(\mathsf{E\_S}|\mathbf{y}, i) = \sum_{\mathbf{z}: z_i = \mathsf{E}, z_{i+1} = \mathsf{S}} p(\mathbf{z}|\mathbf{y}). \qquad (2)$$

Then the probability that there is a boundary between y_i and y_{i+1} is:

$$p^{\mathsf{B}}(\mathbf{y}, i) = \sum_{l \in \{\mathsf{S\_S,\ S\_B,\ E\_S,\ E\_B}\}} p(l|\mathbf{y}, i). \qquad (3)$$

And the probability that there is no boundary is:

$$1 - p^{\mathsf{B}}(\mathbf{y}, i) = \sum_{l \in \{\mathsf{B\_M,\ B\_E,\ M\_M,\ M\_E}\}} p(l|\mathbf{y}, i). \qquad (4)$$

Please note that illegal label bigrams (a.k.a. illegal transitions) such as B_B are forbidden and always get zero probability.

According to our experiments and analysis, our final approach keeps all pauses having $p^{\mathsf{B}} \geq 0.1$, regardless of the pause duration.

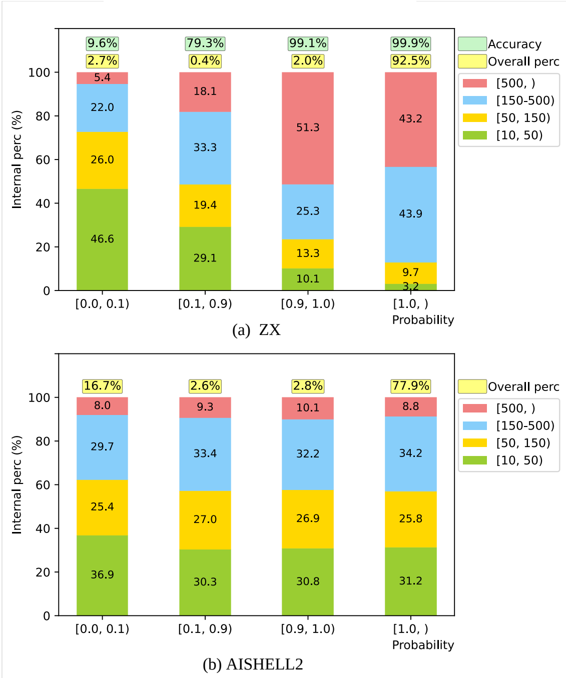

Figure 2: Statistics of pauses regarding probability/accuracy of being boundaries and duration distribution. Probabilities are grouped into four bins, i.e., $[0.0, 0.1)$, $[0.1, 0.9)$, $[0.9, 1.0)$, and $[1.0,)$. The overall percentage means the proportion of pauses belonging to a given probability bin against all pauses. Pause durations are divided into four bins, i.e., $[10, 50)$, $[50, 150)$, $[150, 500)$, and $[500,)$, in the unit of *ms*. Given a probability bin, the internal percentage means the proportion of pauses belonging to a given duration bin against all pauses in the probability bin. For the ZX data, accuracy means the proportion of pauses that are really word boundaries according to further verification.

2.4 Analysis of Pauses

The lower part of Table 1 presents the overall statistics of pauses in both ZX and AISHELL2, with and witout filtering. One notable difference between the two datasets is that pauses are much sparser in the latter. Almost all sentences in ZX contain pauses ($\geq 10ms$), and for sentences that do contain pauses, the average number of pauses is about 8. In contrast, less than 40% of sentences in AISHELL2 contain pauses, and the average number is only 1.7. We believe the major reason is that the sentences are much longer in ZX than in AISHELL2. Each sentence contains about 25 words in average in the former, while only about 7 in the latter.

Figure 2 provides more details about the pauses. We group probability of $[0.1, 0.9)$ into one bin for two reasons. First, the total percentage of pauses falling into the bin is still not high. Second, pauses in the bin scatter quite evenly in terms of probability. Our experiments show that despite the low overall percentage, pauses in this bin are quite valuable for improving model performance.

From the aspect of *overall percentage*, the most notable difference is that the percentages for the first two probability bins, i.e., $[0.0, 0.1)$ and $[0.1, 0.9)$, are much higher in AISHELL2 than in ZX ($2.7 \rightarrow 15.7$ and $0.4 \rightarrow 2.6$).

From the aspect of *internal percentage*, we can see that pauses of different duration bins have similar distribution in the four probability bins in AISHELL2. In contrast, in ZX the percentages of smaller pause duration, i.e., $[10, 50)$ and $[50, 150)$, decrease consistently as the probability increases.

For ZX, we also manage to report the accuracy for each probability bin, in order to gain more insights. Instead of performing manual annotation, we notice that the ZX data with WS annotations are a part of the transcription texts and thus evaluate the accuracy of pauses as word boundaries over the overlapping sentences, using annotated WS information as the gold-standard.[5]

It is clear that accuracy increases consistently as the probability becomes higher. Most of pauses falling into the $[0.0, 0.1)$ bin are incorrect boundaries, and thus should be excluded.

Pauses with high probability, i.e., $[0.9, 1.0)$ and $[1.0,)$, have almost perfect accuracy and should be included. However, from the perspective of model training, we suspect that this part may not be very useful, since the baseline model is already quite certain about these boundaries.

Most importantly, pauses in the $[0.1, 0.9)$ have 79.3% accuracy, which is much higher than that for the $[0.0, 0.1)$ bin. Our experiments show that these pauses are very useful for the model.

3 Utilizing Pauses as Word Boundaries

Word boundaries as naturally annotated CWS data. In fact, quite a few previous studies try to explore word boundaries from different channels and use them as naturally annotated CWS data (Jiang et al., 2013; Liu et al., 2014; Yang and Vozila, 2014). Under a sequence labeling framework, word boundaries can be naturally treated as partial annotations and used to construct a constrained label space.

[5] Due to several factors, including transcription mistakes, difference in the fiction versions, difference in sentence segmentation procedures, etc, we collect about 2K overlapping sentences that appear both in the transcription texts and the ZX evaluation data.

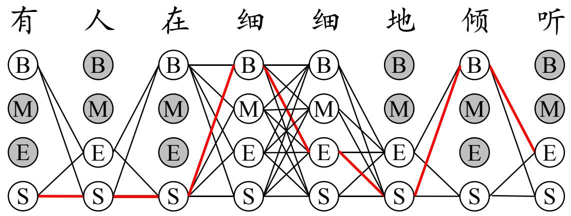

Figure 3: Constrained label space for the sentence in Figure 1, in which we obtain two boundaries "有人/在细细地/倾听". Illegal labels are marked as gray. The red thick lines present a legal path that may be selected by a model.

Figure 3 gives an example. Due to the boundary "人 (people)/在 (is)", the left-side char can only be either a single-char word or the end of a word, where as the right-side char can only be either a single-char word or starting a word. A similar explanation goes to the second boundary.

3.1 Problem with the Partial-CRF strategy

To make use of partially annotated training samples, shown in Figure 3, we first employ the partial-CRF strategy (Liu et al., 2014), which is theoretically elegant. The basic idea is that instead of maximizing the probability of a single gold-standard label sequence, the training objective is to maximize the sum of probabilities of all legal paths in the constrained space, which can be efficiently computed via a variant Forward algorithm.

However, our experiments show that this strategy performs terribly when the model is trained on both CTB-Train and the target-domain data with only boundaries. Further analysis show that the models predicts an extreme high percent of the S label to the target-domain sentences (i.e., most words being single-char). We suspect the major reason is that all characters in the constrained space can be labeled as "S" tags, as shown in Figure 3, and the model fails to transfer from CTB to the target domain the knowledge of when/how to compose multi-char words.

3.2 The Complete-Then-Train (CTT) Strategy

To address the above issue, we present a simple yet effective CTT strategy. The basic idea is converting partial annotations into full annotations by letting a basic model select an optimal sequence in the constrained space. Figure 4 illustrates the strategy, consisting of three steps. First, we train a CWS model (i.e., baseline) on the basic CWS training dataset without using naturally annotated data.

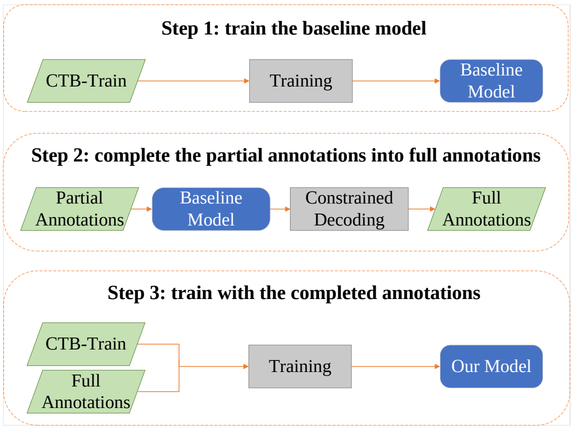

Figure 4: The CTT training strategy.

Second, we employ the basic model to complete partial annotations into full ones. More concretely, the basic model selects an optimal label sequence from the constrained space via constrained Viterbi decoding. For example, we suppose the model selects the path marked by red thick lines in Figure 3. Finally, we use both basic CWS data and completed data to train the full model.

4 Experiments

4.1 Annotation Details for AISHELL2

Upon release, the AISHELL2 data sets aside 2,500 sentences and 3,000 sentences, serving as the dev and test sets, respectively. We apply the baseline models and our full models to the 5,500 sentences. From the sentences that receive different results from a baseline model and a full model, we randomly select 900 sentences for manual annotation.

Two postgraduate students participate the data annotation. Our annotation process consists of two stages. At the first stage, each sentence is annotated by the two annotators, and the differences are resolved by further discussion. During this stage, the annotators becomes familiar with the segmentation guidelines of CTB (Xia, 2000). At the second stage, one annotator (the first author of this submission) annotates all left sentences. We plan to annotate more sentences to make the experiment conclusions more solid.

To speed up annotation, we provide the results of the two models with differences highlighted. Meanwhile, the models' results are randomly given, so that the annotator cannot tell which results are from which model, avoiding the risk of favoring our own approach. Table 2 illustrates the annotation process.

Item	Sentence
	邀请上朋友办个晚宴 Invite friends to host a dinner party
Results of Model 1	邀* / 请上* / 朋友 / 办 / 个 / 晚宴 / Invite / please up / friends / to host / a / dinner party
Results of Model 2	邀* / 请* / 上* / 朋友 / 办 / 个 / 晚宴 / Invite / please / up / friends / to host / a / dinner party
Annotation Results	邀 请 / 上 / 朋 友 / 办 / 个 / 晚 宴 / Invite / friends / to host / a / dinner party

Table 2: Illustration of the annotation process of the AISHELL2 dev/test data. Please notice that results of models are randomly given, so that the annotator cannot favor our own approach.

After removing sentences that cannot be labeled due to noise or transcription errors, we obtain 881 sentences in total. We split them into a dev set and a test set. Table 1 shows the data statistics.

4.2 Settings

For the evaluation, we employ the standard metrics of precision (P), recall (R), and the F1 score.

As discussed in Section 2.3, we regard CWS as a sequence labeling task and employ the BERT-CRF baseline model.[6] We use AdamW with an initial learning rate of 5e-5, and a mini-batch size of 1000 characters. The dropout ratio is 0.1 for all models. We train each model for 10 epochs.

Following previous works on cross-domain word segmentation on ZX, we use CTB5-train as the training data, and use the target-domain dev data to select the best epoch number.

To be more convincing, we train each model three times with three different random seeds and present the average and standard deviation.[7]

4.3 Results

Table 3 presents the main results. Compared with previous results on ZX, our baseline model already achieve very good performance.

Most importantly, we can see that our final models using filtered pauses as word boundary ($p^{\mathrm{B}} \geq 0.1$) achieves significant improvement boost by 0.45 and 1.44 in F1 score on ZX-test and AISHELL2-test, respectively, compared with the baseline models.

[6] https://huggingface.co/bert-base-chinese

[7] $\sigma = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_i - \bar{x})^2}$

	P	R	F1	P	R	F1
Models		ZX-dev			ZX-test	
Baseline	94.16	94.39	$94.27_{\pm 0.21}$	93.16	93.82	$93.49_{\pm 0.22}$
Using word boundaries						
w/o filtering	94.18	94.34	$94.26_{\pm 0.49}$	93.69	94.03	$93.86_{\pm 0.36}$
w/ filtering ($p^{\mathsf{B}} \geq 0.9$)	94.27	94.64	$94.45_{\pm 0.20}$	93.46	94.08	$93.77_{\pm 0.25}$
w/ filtering ($p^{\mathsf{B}} \geq 0.1$)	94.23	94.78	$\mathbf{94.50}_{\pm 0.27}$	93.56	94.32	$\mathbf{93.94}_{\pm 0.20}$
			Previous Results			
Ding et al. (2020)		—				90.90
Luo et al. (2022)		—				91.11
Higashiyama et al. (2020)		—				93.30
Models		AISHELL2-dev			AISHELL2-test	
Baseline	83.56	86.25	$84.86_{\pm 1.55}$	86.11	88.94	$87.48_{\pm 1.76}$
Using word boundaries						
w/o filtering	85.04	88.07	$86.52_{\pm 1.07}$	86.29	89.21	$87.71_{\pm 1.12}$
w/ filtering ($p^{\mathsf{B}} \geq 0.9$)	85.38	87.87	$86.60_{\pm 0.43}$	87.64	89.72	$88.66_{\pm 0.20}$
w/ filtering ($p^{\mathsf{B}} \geq 0.1$)	85.65	87.77	$\mathbf{86.69}_{\pm 0.45}$	87.94	89.94	$\mathbf{88.92}_{\pm 0.39}$

Table 3: Main results on both datasets.

Effect of filtering pauses. Compared with the results of models without filtering pauses, our final models ($p^{\mathsf{B}} \geq 0.1$) are consistently superior in F1 scores.

Usefulness of pauses with probability of $[0.1, 0.9)$. On the one hand, compared with using $p^{\mathsf{B}} \geq 0.9$, our final models ($p^{\mathsf{B}} \geq 0.1$) are consistently superior in F1 scores. On the other hand, compared with not filtering pauses, the models only using pauses of $p^{\mathsf{B}} \geq 0.9$ have even lower F1 score in ZX-test and AISHELL2-test. These two aspects reflect the usefulness of pauses with probability of $[0.1, 0.9)$.

5 Related Works

5.1 Integrated Speech and Text Processing

In the deep learning, the Transformer-based model architecture becomes popular in both speech processing and NLP fields. The same architecture makes it convenient to process speech and textual data in an integrated manner. Intuitively, speech and text can provide complementary useful features. We summarize recent works into three categories.

(1) Speech as extra features for NLP. The most straightforward way is to extract features from speech and use them as extra inputs for an NLP model. Zhang et al. (2021) present an interesting pioneer effort and use speech features to help CWS, which is closely with our work. Their approach requires parallel speech-text data in both training and test phases, with WS annotations and the character-frame alignments. They manually annotate 250 sentences and split them into training-test data. Experiments show that extra speech features are beneficial.

Different from their work, ours emphasis on the use of pause information in speech. We do not need WS annotations for the text data and automatically derive character/frame alignments. In the test phase, our CWS model performs on on text data, rather than parallel speech-text data.

(2) MTL with cross-attention interaction. Given parallel speech-text data, Sui et al. (2021) present a multi-task learning approach that performs NER and ASR at the same time. They first use separate encoders for the two types of inputs, and then employ the cross-attention mechanism to achieve multi-model interaction.

(3) End-to-End language analysis from speech. Several works propose to directly derive language analysis results from speech inputs in an end-to-end manner. Ghannay et al. (2018) embed NE labels into texts and train a model that transcribes speech into texts and treats NE labels as normal tokens. They conduct experiments on French NER. Yadav et al. (2020) applie the approach to English NER and propose a new label embedding scheme. Chen et al. (2022) present a Chinese datasets of parallel speech-text data with NE annotations, and systematically compare the pipeline and end-to-end approaches.

Wu et al. (2022) propose an end-to-end relation extraction model that transcribes speech into (entity, entity, relation) triples, and totally ignores the full text (not performing ASR). However, their experiments show that the end-to-end approach is inferior to the pipeline model, i.e., first ASR and then relation extraction on texts.

Utilizing speech pauses. Fleck (2008) make use of speech pauses to help English ASR, and more specifically to help transforming phonemes into words. The pauses are output by a previous ASR component and are embedded in the phoneme sequence. They propose to use the pauses to segment the phoneme sequence into several fragments and transform them into words separately.

5.2 Naturally annotated CWS data

Mining naturally annotated data. Previous studies try to mine naturally annotated CWS data from different channels. Jiang et al. (2013) hypothesize that anchor texts (i.e., for hyperlinks) in HTML-format web documents are very likely to correspond to complete meaning units, and thus can be explored to obtain at least two word boundaries. In the cross-domain scenario, Liu et al. (2014) use a domain-related dictionary and perform maximum matching on unlabeled target-domain text, treating matched texts as annotated words.

Utilizing naturally annotated data. Above naturally annotated data are in two forms. In the first form, some word boundaries in the sentence are given, whereas in the second, some words are given. Both forms can be treated as partial annotations, in contrast to full annotations, and be encoded as constrained label space as shown in Figure 3.

Jiang et al. (2013) proposes a constrained decoding approach to learn from partially annotated data with word boundaries. They use a max-margin training loss. For each training sentence, they first obtain an optimal label sequence from the constrained space and use it as gold-standard reference in an online fashion.

Some researchers employ the CRF (Liu et al., 2014; Yang and Vozila, 2014) to extend the loss for learning from partial/incomplete annotations. In this work, we also use this approach, but obtain inferior performance probably due to the issue of pervasive "S" labels. We propose a simple yet effective CTT strategy.

6 Conclusion

This paper for the first time proposes to explicitly mine word boundaries from speech-text data as extra naturally annotated training data for cross-domain CWS. Initially, we collect speech-text data from the web fiction domain (ZX) and annotate a part of original AISHELL2-dev/test datasets for CWS evaluation. Secondly, we perform character-level alignment on the speech-text data to mine word boundaries. Thirdly, we employ the baseline to calculate the marginal probability of word boundaries. By analyzing the accuracy across four probability range, we filter out word boundaries with probabilities lower than 0.1. Finally, we employ a CTT method to leverage mined word boundaries as extra training data to improve CWS model performance in cross-domain scenarios. Our experiments demonstrate that mined word boundaries significantly enhance CWS via the CTT method. Upon analysis, we find that filtering boundaries is crucial to the efficacy of the CTT method.

Limitations

We believe our work has built a solid foundation for future research on this direction. Meanwhile we are aware that our work is limited in and can be improved from several aspects.

First, our approach relies on accurate character-level alignment between speech and texts. So far, we use MFA as a black-box and our early trails showed that the end-to-end Transformer-CTC model is inferior. Therefore, our proposed approach may be more effective with improved alignment quality.

Second, this work only utilizes pauses detected by character-level aligner to derive word boundaries, but ignore other rich features in speech. For example, intonation or pitch change may also be helpful.

Finally, as discussed in 4.1, we plan to annotate more evaluation data for AISHELL2 to make the experiments more solid.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang. 2022. Aishell-NER: named entity recognition from Chinese speech. In *Proceedings of ICASSP*, pages 8352–8356. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota.

Ning Ding, Dingkun Long, Guangwei Xu, Muhua Zhu, Pengjun Xie, Xiaobin Wang, and Haitao Zheng. 2020. Coupling distant annotation and adversarial training for cross-domain Chinese word segmentation. In *Proceedings of ACL*, pages 6662–6671, Online.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. AISHELL-2: transforming mandarin ASR research into industrial scale. *ArXiv preprint*.

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138, Columbus, Ohio.

Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *Workshop of SLT*, pages 692–699. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of INTERSPEECH*, pages 5036–5040.

Shohei Higashiyama, Masao Utiyama, Yuji Matsumoto, Taro Watanabe, and Eiichiro Sumita. 2020. Auxiliary lexicon word prediction for cross-domain word segmentation. *Journal of Natural Language Processing*, (3):573–598.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 3451–3460.

Kaiyu Huang, Degen Huang, Zhuang Liu, and Fengran Mo. 2020a. A joint multiple criteria model in transfer learning for cross-domain Chinese word segmentation. In *Proceedings of EMNLP*, pages 3873–3882, Online.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2020b. Towards fast and accurate neural Chinese word segmentation with multi-criteria learning. In *Proceedings of COLING*, pages 2062–2072, Barcelona, Spain (Online).

Peijie Jiang, Dingkun Long, Yueheng Sun, Meishan Zhang, Guangwei Xu, and Pengjun Xie. 2021. A fine-grained domain adaption model for joint word segmentation and POS tagging. In *Proceedings of EMNLP*, pages 3587–3598, Online and Punta Cana, Dominican Republic.

Wenbin Jiang, Meng Sun, Yajuan Lü, Yating Yang, and Qun Liu. 2013. Discriminative learning with natural annotations: Word segmentation as a case study. In *Proceedings of ACL*, pages 761–769, Sofia, Bulgaria.

Zhen Ke, Liang Shi, Songtao Sun, Erli Meng, Bin Wang, and Xipeng Qiu. 2021. Pre-training with meta learning for Chinese word segmentation. In *Proceedings of NAACL*, pages 5514–5523, Online.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880, Online.

Yang Liu and Yue Zhang. 2012. Unsupervised domain adaptation for joint segmentation and POS-tagging. In *Proceedings of COLING 2012: Posters*, pages 745–754, Mumbai, India.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for CRF-based Chinese word segmentation using free annotations. In *Proceedings of EMNLP*, pages 864–874, Doha, Qatar.

Zhiyong Luo, Mingming Zhang, Yujiao Han, and Zhilin Zhao. 2022. Semi-supervised CRF Chinese word segmentation based on neural network (in Chinese). In *Proceedings of CCL*, pages 644–655, Nanchang, China.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Proceedings of INTERSPEECH*, pages 498–502.

Todd K Moon. 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine*, (6):47–60.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *ArXiv preprint*.

Andrew Senior, Haşim Sak, Félix de Chaumont Quitry, Tara Sainath, and Kanishka Rao. 2015. Acoustic modelling with CD-CTC-SMBR LSTM RNNs. In *Workshop of ASRU*, pages 604–609.

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, and Jun Zhao. 2021. A large-scale Chinese multimodal NER dataset with speech clues. In *Proceedings of ACL*, pages 2807–2818, Online.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020. Improving Chinese word segmentation with wordhood memory networks. In *Proceedings of ACL*, pages 8274–8285, Online.

Tongtong Wu, Guitao Wang, Jinming Zhao, Zhaoran Liu, Guilin Qi, Yuan-Fang Li, and Gholamreza Haffari. 2022. Towards relation extraction from speech. In *Proceedings of EMNLP*, pages 10751–10762, Abu Dhabi, United Arab Emirates.

Fei Xia. 2000. The segmentation guidelines for the penn Chinese treebank (3.0). *University of Pennsylvania Technical Report, IRCS00-06*.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese treebank: phrase structure annotation of a large corpus. *Natural language engineering*, (2):207–238.

Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from English speech. In *Proceedings of INTERSPEECH*, pages 4268–4272.

Fan Yang and Paul Vozila. 2014. Semi-supervised Chinese word segmentation using partial-label learning with conditional random fields. In *Proceedings of EMNLP*, pages 90–98, Doha, Qatar.

Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2021. Why does CTC result in peaky behavior? *ArXiv preprint*.

Dong Zhang, Zheng Hu, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. More than text: multi-modal Chinese word segmentation. In *Proceedings of ACL*, pages 550–557, Online.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Type-supervised domain adaptation for joint segmentation and POS-tagging. In *Proceedings of EACL*, pages 588–597, Gothenburg, Sweden.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara N. Sainath, Pedro J. Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Franccoise Beaufays, and Yonghui Wu. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *ArXiv preprint*.