A Large-Scale Chinese Multimodal NER Dataset with Speech Clues

Dianbo Sui, Zhengkun Tian, Yubo Chen, Kang Liu, Jun Zhao
National Laboratory of Pattern Recognition, Institute of Automation, CAS
School of Artificial Intelligence, University of Chinese Academy of Sciences
{dianbo.sui, zhengkun.tian, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

In this paper, we aim to explore an uncharted territory, which is Chinese multimodal named entity recognition (NER) with both textual and acoustic contents. To achieve this, we construct a large-scale human-annotated Chinese multimodal NER dataset, named CNERTA. Our corpus totally contains 42,987 annotated sentences accompanying by 71 hours of speech data. Based on this dataset, we propose a family of strong and representative baseline models, which can leverage textual features or multimodal features. Upon these baselines, to capture the natural monotonic alignment between the textual modality and the acoustic modality, we further propose a simple multimodal multitask model by introducing a speech-to-text alignment auxiliary task. Through extensive experiments, we observe that: (1) Progressive performance boosts as we move from unimodal to multimodal, verifying the necessity of integrating speech clues into Chinese NER. (2) Our proposed model yields state-of-the-art (SoTA) results on CNERTA, demonstrating its effectiveness. For further research, the annotated dataset is publicly available at http://github.com/DianboWork/CNERTA.

1 Introduction

"Speech is a part of thought."
— Oliver Sacks, *Seeing Voices*

As a fundamental subtask of information extraction, named entity recognition (NER) aims to locate and classify named entities mentioned in unstructured texts into predefined semantic categories, such as person names, locations and organizations. NER plays a crucial role in many natural language processing (NLP) tasks, including relation extraction (Zelenko et al., 2003), question answering (Mollá et al., 2006) and summarization (Aramaki et al., 2009).

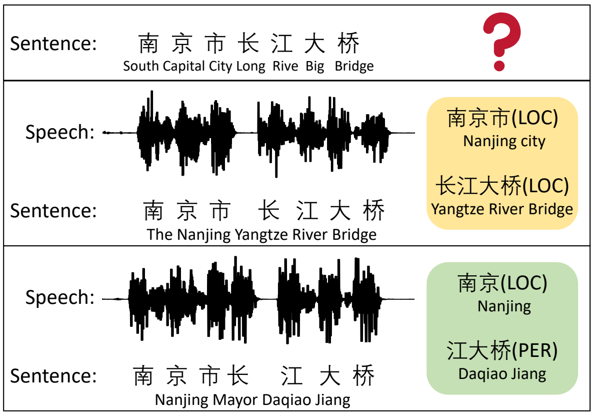

Figure 1: The given sentence "南京市长江大桥" can be segmented into "[南京市] [长江大桥]" or "[南京] [市长] [江大桥]". Only based on textual contents, it is difficult to infer NER tags. But the speech waveforms of these two segmentations are radically different.

Most of the research on NER, such as Lample et al. (2016); Ma and Hovy (2016); Chiu and Nichols (2016), only relies on the textual modality to infer tags. However, when texts are noisy or short, and it is not sufficient to locate and classify named entities accurately only based on textual information (Baldwin et al., 2015; Lu et al., 2018). One promising solution is to introduce other modalities as the supplement of the textual modality. So far, some studies on multimodal NER, such as Moon et al. (2018); Zhang et al. (2018); Lu et al. (2018); Arshad et al. (2019); Asgari-Chenaghlu et al. (2020); Yu et al. (2020); Chen et al. (2020); Sun et al. (2020), have attempted to couple the textual modality with the visual modality and witnessed a stable improvement.

In this work, we also focus on multimodal NER. But differently from previous studies, we pay special attention to Chinese multimodal NER with both textual and acoustic contents. The motivation comes from two aspects:

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing, pages 2807–2818
August 1–6, 2021. ©2021 Association for Computational Linguistics

First, despite much recent success in multimodal NER, current studies on this topic are limited in English, and totally skirt other languages. Meanwhile, previous work on Chinese NER, such as Xu et al. (2013); Peng and Dredze (2016a); Zhang and Yang (2018); Cao et al. (2018); Sui et al. (2019); Gui et al. (2019); Ma et al. (2020); Li et al. (2020), totally ignores valuable multimodal information. With around 1.3 billion native speakers and the wide spread of short-form video apps in China, it is necessary and urgent to carry out research on Chinese multimodal NER.

Second, unlike the static visual modality, the time-varying acoustic modality plays a unique role in Chinese NER, especially in providing precise word segmentation information. In detail, different from English, Chinese is an ideographic language featured by no word delimiter between words in written. This language characteristic is one of the major roadblocks in Chinese NER, since named entity boundaries are usually word boundaries (Zhang and Yang, 2018). Fortunately, cues contained in the fluent acoustic modality, especially pauses between adjacent words, are able to aid the NER model in discovering word boundaries. A classic example shown in Figure 1 can perfectly illustrate this point. In this example, the sentence with ambiguous word segmentation would be disambiguated with the aid of the acoustic modality, which would absolutely assist the model to infer correct NER tags.

In this work, we make the following efforts to advance multimodal NER:

First, we construct a large-scale human-annotated **C**hinese **NER** dataset with **T**extual and **A**coustic contents, named CNERTA. Specifically, we annotate all occurrences of 3 entity types (person name, location and organization) in 42,987 sentences originating from the transcripts of Aishell-1 (Bu et al., 2017), a corpus that has been widely employed in Mandarin speech recognition research in recent years (Shan et al., 2019; Li et al., 2019; Tian et al., 2020). In particular, unlike previous multimodal NER datasets (Moon et al., 2018; Zhang et al., 2018; Lu et al., 2018) are all flatly annotated, not only the topmost entities but also nested entities are annotated in CNERTA.

Second, based on CNERTA, we establish a family of strong and representative baselines. In detail, we first investigate the performance of several classic text-only models on our dataset, including BiLSTM-CRF (Lample et al., 2016) and BERT-CRF (Devlin et al., 2019). Then, since introducing a lexicon has been proven as an effective way to incorporate word information in Chinese NER (Zhang and Yang, 2018), we implement several lexicon-enhanced models, such as Lattice-LSTM (Zhang and Yang, 2018) and ZEN (Diao et al., 2020), to explore whether the acoustic modality can provide word information beyond the lexicon. Finally, to verify the effectiveness of introducing the acoustic modality, we test some widely used multimodal models, such as CMA (Tsai et al., 2019) and MMI (Yu et al., 2020), on our dataset.

Third, upon these strong baselines, we further propose a simple **M**ulti-**M**odal **M**ulti-**T**ask model (short for **M3T**) to make better use of the pause information in the acoustic modality. Specifically, different from coupling the visual modality with the textual modality, there is a monotonic alignment between the acoustic modality and the textual modality. Armed with such an alignment, the position of each Chinese character in the continuous speech would be determined, which would make it easy to discover pauses between adjacent words. Therefore, to automatically estimate this desired alignment, we introduce a speech-to-text alignment auxiliary task and propose a hybrid CTC/Tagging loss. In the hybrid loss, a masked CTC loss (Graves et al., 2006) is designed for enforcing a monotonic alignment between speech and text sequences.

The primary contributions of this work can be summarized as follows:

- We construct CNERTA, the first human-annotated Chinese multimodal NER dataset, where each annotated sentence is paired with its corresponding speech data. To our best knowledge, this dataset is not only the largest multimodal NER dataset, but also the largest Chinese nested NER dataset.

- We establish a family of baselines to leverage textual features or multimodal features. Through various experiments, we observe consistent performance boosts originating from acoustic features, which verifies the significant merits of integrating acoustic features for Chinese NER.

- We further propose a multimodal multitask method by introducing a speech-to-text alignment auxiliary task. By jointly solving the tagging task and the alignment task, the proposed method can yield SoTA results on CNERTA.

2 Related Work

Mutlimodal NER: As multimedia technology evolves, processing multimodal data is becoming a burning issue. As a basic NLP tool, multimodal NER attracts increasing attention in recent years. Most of studies on multimodal NER focus on leveraging the associate images to better identify the named entities contained in the text. Specifically, Moon et al. (2018) propose a multimodal NER network with modality attention to fuse textual and visual information. To model inter-modal interactions and filter out the noise in the visual context, Zhang et al. (2018) propose an adaptive co-attention network and a gated visual attention mechanism for multimodal NER. As transformer-based models (Vaswani et al., 2017; Devlin et al., 2019) become the mainstream method in NLP, researchers turn to study how to fuse visual clues in transformers structure. Chen et al. (2020) use captions to represent images as text and adopt transformer-based sequence labeling models to connect multimodal information. Yu et al. (2020) propose a Multimodal Transformer model, which empowers transformer with a multimodal interaction module to capture the inter-modality dynamics between words and images. But different from them, we aim to explore an unexplored territory in this work, which is Chinese multimodal NER with both speech and textual contents.

Chinese NER: Compared with English NER, Chinese NER is more complicated since the written text in Chinese is not naturally segmented. Therefore, how to incorporate word information is the key challenge in Chinese NER. There are three main ways to fuse word information in Chinese NER. The first one is the pipeline method. In the pipeline method, Chinese word segmentation (CWS) is first applied and then a word-based NER model is used. The second one is to learn CWS and NER tasks jointly (Xu et al., 2013; Peng and Dredze, 2016b; Cao et al., 2018; Wu et al., 2019). In such a way, the word boundary information in the CWS task can be transferred to the NER model. The third one is to resort to an automatically constructed lexicon (Zhang and Yang, 2018; Ding et al., 2019; Liu et al., 2019a; Sui et al., 2019; Gui et al., 2019; Li et al., 2020; Ma et al., 2020; Xue et al., 2020). Different from all previous studies, we focus on use speech clues to incorporate word information in Chinese NER.

	Train	Dev	Test
Audio Duration	56.68h	7.50h	7.59h
Avg Sent Len	19.69	19.77	19.75
Max Sent Len	39	44	39
Prop Nested Ent	31.25%	29.50%	28.35%
# Instance	34,102	4,440	4,445
# Entity	23,805	5,889	7,263
# ORG	7,066	2,187	2,794
# PER	5,846	1,116	1,072
# LOC	10,893	2,586	3,397

Table 1: The statistics of training, development and test folds of the annotated corpus. Here, "Avg" denotes average, "Sent" denotes sentence, "Len" denotes length, "Prop" denotes proportion, "Ent" denotes entity and "#" denotes number.

3 Dataset Acquisition and Comparison

In this work, we aim to explore Chinese NER with both speech and textual clues. But we are not aware of any such existing corpus, hence we are motivated to collect one. In this section, we will discuss the data acquisition process, subsequently present statistics of the dataset and compare the annotated dataset with other widely-used NER datasets.

3.1 Dataset Acquisition

The main challenge in data acquisition is to find a large-scale dataset, which includes texts and the corresponding speech data. One possible way is to attach speech data to current existing Chinese NER datasets. However, it is costly to gather hundreds of participants in the recording. Therefore, we take a different way, manually annotating NER tags on a speech recognition dataset from scratch. In detail, our annotated dataset is based on Aishell-1 (Bu et al., 2017) dataset, which is a large-scale Mandarin automatic speech recognition dataset. In this dataset, text transcriptions are chosen from five domains: "Finance", "Science and Technology", "Sport", "Entertainments" and "News". There are 400 participants in the recording, and the gender of participants is balanced with 47% male and 53% female. Speech utterances are recorded via three categories of devices in parallel, which are a high fidelity microphone working at 44.1 kHz, 16-bit, Android phones working at 16 kHz, 16-bit, and Apple iPhones working at 16 kHz, 16-bit.

To ensure the quality of annotation, we design two rounds in the annotation procedure. In the first

Dataset	# Train	# Dev	# Test	# Total	Language	Structure	Modality
MSRA	46,364	-	4,365	50,729	Chinese	Flat	Text
OntoNotes	15,724	4301	4,346	24,371	Chinese	Flat	Text
Weibo NER	1,350	271	270	1,891	Chinese	Flat	Text
Resume	3,821	463	477	4,761	Chinese	Flat	Text
GENIA	15,022	1,669	1,854	18,545	English	Nested	Text
JNLPBA	20,546	-	4,260	24,806	English	Nested	Text
ACE-2004	6,198	742	809	7,749	English	Nested	Text
ACE-2005	7,285	968	1,058	9,311	English	Nested	Text
Twitter-2015	4,000	1,000	3,257	8,257	English	Flat	Text + Image
Twitter-2017	3,373	723	723	4,819	English	Flat	Text + Image
CNERTA	34,102	4,440	4,445	42,987	Chinese	Nested	Text + Speech

Table 2: A comparison between CNERTA and other existing widely-used NER datasets.

round, we use Brat (Stenetorp et al., 2012) as the annotation tool and ask 3 internal annotators (including the first author of this paper) to perform annotation, who are very familiar with this task. They independently identify and classify named entities in the transcriptions with more than 17 characters. Cohen's kappa coefficient (Cohen, 1960) is used to measure the inter-annotator agreements. After the first round, κ = 0.965, which shows the quality of CNERTA is satisfactory. But there are still some sentences for which annotators give out different annotations. For those sentences, the annotators check the disagreed annotations carefully and discuss to reach the agreements for all cases.

After we finish the annotation process, we split the dataset into three parts: training, development, and test set. Table 1 shows the high level statistics of data splits for CNERTA.

3.2 Dataset Comparison

We compare CNERTA with several widely used NER datasets in Table 2. Specifically, we first compare our corpus with some Chinese NER datasets, such as MSRA (Levow, 2006), OntoNotes (Weischedel et al., 2011), Weibo NER (Peng and Dredze, 2016a) and Resume (Zhang and Yang, 2018). Then, we compare our corpus with several widely used nested NER datasets, like GENIA (Kim et al., 2003), JNLPBA (Collier and Kim, 2004), ACE-2004 (Doddington et al., 2004) and ACE-2005 (Walker et al., 2004). Finally, multimodal NER datasets, including Twitter-2015 (Zhang et al., 2018) and Twitter-2017 (Lu et al., 2018), are compared with our corpus.

From Table 2, we observe that our corpus has unique value compared with the existing datasets. The value is reflected in the following aspects: (1) CNERTA is a large-scale dataset; (2) CNERTA is the first Chinese multimodal dataset; (3) Not only the topmost entities but also nested entities are annotated; (4) Among these datasets, the acoustic modality is only introduced in CNERTA.

4 Preliminaries

4.1 Task Description

Given a text $X = x_1, x_2, ..., x_n$ and its corresponding speech $S = s_1, s_2, ..., s_t$, where x_i denotes the i-th Chinese character and s_j denotes the j-th waveform frame, the goal of the task is to leverage textual and speech clues to identify and classify all named entities contained in the text.

4.2 Nested Structure Linearization

Unlike flat NER, named entities may overlap and also be labeled with more than one label in nested NER. To solve nested NER, we follow Straková et al. (2019) to encode the nested entity structure into a CoNLL-like, per-character BIO encoding (Ramshaw and Marcus, 1995). There are two rules to guide the linearization: (1) entity mentions starting earlier have priority over entities starting later, and (2) for mentions with the same beginning, longer entity mentions have priority over shorter ones. A multilabel for a given Chinese character is a concatenation of all intersecting entity mentions, from the highest priority to the lowest. For more details, we refer readers to Straková et al. (2019).

4.3 Acoustic Encoder

The acoustic encoder is used to map raw speech signals into continuous space. There are three parts in the proposed acoustic encoder: a speech processing layer, a convolution front end and a transformer-based encoder.

Specifically, in the speech processing layer, a speech signal first goes through a pre-emphasis filter; then gets sliced into frames and a window function is applied to each frame; afterwards, a Short-Time Fourier transform (Kwok and Jones, 2000) is employed on each frame and the power spectrum is calculated; and subsequently, the filter banks (Ravindran et al., 2003) are computed. Then, we use a convolution front end to down-sample the long acoustic features. In the convolution front end, following Dong et al. (2018); Tian et al. (2020), two 3×3 CNN layers with stride 2 are stacked for both time and frequency dimensions. Afterwards, in order to enable the acoustic encoder to attend by relative positions, the positional encoding is added to the output of the convolution front end. Finally, to effectively capture long-term dependencies, down-sampled acoustic features flow through the transformer-based encoder (Vaswani et al., 2017). The transformer-based encoder is a stack of 6 identical layers, each of which is composed of a self-attention sub-layer and a feed-forward network.

5 Baselines

Based on the annotated dataset, a family of strong and representative baselines is established, including (1) text-only models presented in Section 5.1, (2) lexicon-enhanced models shown in Section 5.2 and (3) multimodal models introduced in Section 5.3.

5.1 Text-Only Model

Open-Source NLP Toolkit: Many open-source NLP toolkits, such as spaCy (Honnibal et al., 2020) and Stanza (Qi et al., 2020), support Chinese NER. In spaCy, a multitask CNN is employed. In Stanza, a contextualized string representation based tagger from Akbik et al. (2018) is adopted. In both spaCy and Stanza, the tagger is trained on OntoNote (Weischedel et al., 2011). To map the output of taggers to CNERTA's label space, expert-designed rules are used, such as PERSON → PER. Since these toolkits are only designed for flat structure, we do not evaluate these toolkits in nested settings.

BiLSTM-CRF: Featured by a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) as the textual encoder and conditional random fields (CRF) (Lafferty et al., 2001) as the decoder, the widely used BiLSTM-CRF (Lample et al., 2016) is adopted as an important baseline.

PLM-CRF: Instead of training a model from scratch, we also adopt the framework of fine-tuning a pretrained language model (PLM) on a downstream task (Radford et al., 2018). In this framework, we adopt **BERT** (Devlin et al., 2019) as the textual encoder and use CRF as the decoder. In addition to initializing the textual encoder with the original pretrained BERT model, a SoTA Chinese pretrained language model, called **MacBERT** (Cui et al., 2020), is used. Compared with BERT, MacBERT is built upon RoBERTa (Liu et al., 2019b) and the original MLM task in BERT is replaced with the MLM as correction task. For more details, we refer readers to Cui et al. (2020).

5.2 Lexicon-Enhanced Model:

A drawback of the text-only methods mentioned above is that explicit word and word sequence information is not fully exploited, which can be potentially useful. With this consideration, we also adopt lexicon-enhance models to incorporate word lexicons. (1) **Lattice-LSTM** (Zhang and Yang, 2018) is a classic method that can encode a sequence of input characters as well as all potential words that match a lexicon. (2) **ZEN** (Diao et al., 2020) is a pretrained Chinese text encoder enhanced by an n-gram lexicon. In ZEN, n-gram contexts are extracted, encoded and integrated with the character encoder. For more details about Lattice-LSTM and ZEN, we refer readers to Zhang and Yang (2018) and Diao et al. (2020).

5.3 Multimodal Model

To leverage the acoustic modality, several multimodal models are introduced. In these models, fusion modules are built on the top of the acoustic encoder and the textual encoder, which are designed for capturing the interaction between the textual hidden representations $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$; $\mathbf{x}_i \in \mathbb{R}^d$ and the acoustic representations $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{t'}]$; $\mathbf{s}_j \in \mathbb{R}^d$. We present two representative fusion modules, which are Cross-Modal Attention (CMA) module (Tsai et al., 2019) and Multimodal Interaction (MMI) module (Yu et al., 2020).

Cross-Modal Attention Module (CMA): Given the textual hidden representations $\mathbf{X} \in \mathbb{R}^{d \times n}$ and the acoustic representations $\mathbf{S} \in \mathbb{R}^{d \times t'}$, we first employ a m-head cross-modal attention mechanism (Tsai et al., 2019), by treating $\mathbf{X}$ as queries, and $\mathbf{S}$ as keys and values:

$$\mathrm{CA_i}(\mathbf{X}, \mathbf{S}) = \mathrm{softmax}(\frac{[\mathbf{W_{q_i}X}]^\mathsf{T}[\mathbf{W_{k_i}S}]}{\sqrt{\mathrm{d/m}}})[\mathbf{W_{v_i}S}]$$

$$\text{MH-CA}(\mathbf{X}, \mathbf{S}) = \mathbf{W}'[\mathrm{CA_1}(\mathbf{X}, \mathbf{S}), ..., \mathrm{CA_m}(\mathbf{X}, \mathbf{S})]$$

where $\mathrm{CA_i}$ refers to the i-th head of cross-modal attention, and $\{\mathbf{W}_{q_i}, \mathbf{W}_{k_i}, \mathbf{W}_{v_i}\} \in \mathbb{R}^{d/m \times d}$, $\mathbf{W}' \in \mathbb{R}^{d \times d}$ denote the weight matrices for the query, key, value and multi-head attention, respectively. Then, we stack the following sub-layers on top:

$$\begin{aligned} \hat{\mathbf{F}} &= \mathrm{LN}(\mathbf{X} + \text{MH-CA}(\mathbf{X}, \mathbf{S})) \\ \mathbf{F} &= \mathrm{LN}(\hat{\mathbf{F}} + \mathrm{FFN}(\hat{\mathbf{F}})) \end{aligned} \tag{1}$$

where LN means layer normalization (Ba et al., 2016) and FFN means a fully connected feed-forward network, which consists of two linear transformation with a ReLU activation (Nair and Hinton, 2010). Finally, the new textual representations $\mathbf{F} \in \mathbb{R}^{d \times n}$, which are enhanced by acoustic features, are fed into the CRF decoder to infer NER tags.

Multimodal Interaction Module (MMI): A stack of cross-modal attention layer mentioned above makes up the multimodal interaction module. Since the architecture of MMI is too complex and is not the core of this paper, we will not introduce it in the main text. For more details about MMI, we refer readers to Yu et al. (2020).

6 Proposed Method

Previous multimodal methods ignore a natural monotonic alignment between the acoustic modality and the textual modality. To capture this alignment, we propose a multimodal multitask model, called **M3T**. The framework of the proposed method is shown in Figure 2.

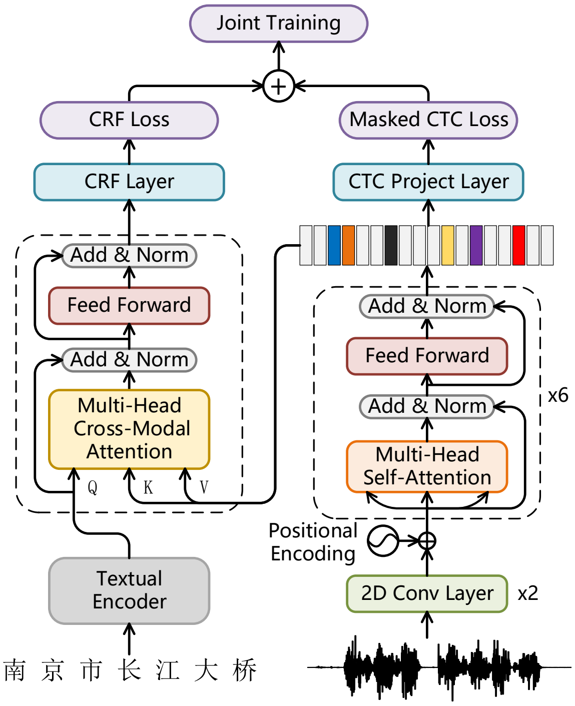

Figure 2: Overall architecture of the proposed multi-modal multitask model.

In the M3T model, we adopt the CMA module to fuse acoustic information into the textual representations. Besides, a CTC project layer is built upon the acoustic encoder, and the loss function is a combination of masked CTC loss and CRF loss. Specifically, through the CTC project layer, each acoustic representation $\mathbf{s}_i \in \mathbb{R}^d$ is first mapped to the total size of model units (in this paper, the model unit is the Chinese character) and then is passed through a logit function:

$$\mathbf{G} = \mathrm{logit}(\mathbf{W}_\mathrm{v}^\mathsf{T}\mathbf{S}) \tag{2}$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times |V|}$ and $|V|$ is the total size of Chinese characters. Unlike automatic speech recognition, only the characters in the given text need to be aligned rather than the entire model units. Therefore, we only keep these rows unchanged, whose corresponding characters are contained in the given text, and fill the other rows in $\mathbf{G} \in \mathbb{R}^{|V| \times t'}$ with the value $-\infty$. The masked tensor $\mathbf{G}$ is then fed into CTC loss. Finally, to jointly solve the tagging task and the alignment task, a hybrid loss of combining the masked CTC loss with the CRF loss is used:

$$\mathcal{L} = \mathcal{L}_{crf} + \lambda \mathcal{L}_{ctc} \tag{3}$$

where λ is a hyperparameter.

7 Experiments

In this section, we carry out various experiments to investigate the effectiveness of introducing the acoustic modality. In addition, we empirically compare the proposed model and these baselines under different settings. Following previous studies in NER (Zhang and Yang, 2018), standard precision (P), recall (R) and F1-score (F1) are used as evaluation metrics.

Model	Resource	Flat NER				Nested NER			
		P	R	F1	Δ	P	R	F1	Δ
spaCy	Text	64.74	23.01	33.94	-	-	-	-	-
Stanza	Text	49.65	27.34	35.27	-	-	-	-	-
BiLSTM-CRF	Text	64.43	62.58	63.49	-	70.72	59.16	64.43	-
Lattice LSTM	Text+Lexicon	67.13	68.34	67.73	-	77.92	60.89	68.36	-
BERT-CRF	Text	74.47	76.34	75.39	-	80.03	72.34	75.99	-
MacBERT-CRF	Text	75.10	78.70	76.86	-	81.22	73.67	77.26	-
ZEN-CRF	Text+Lexicon	74.85	77.78	76.28	-	81.18	72.12	76.38	-
BiLSTM-CMA-CRF	Text+Speech	66.56	65.28	65.92	↑ 2.43	75.28	61.25	67.54	↑ 3.11
BERT-CMA-CRF	Text+Speech	75.67	78.49	77.05	↑ 1.66	81.50	74.48	77.83	↑ 1.84
MacBERT-CMA-CRF	Text+Speech	75.94	81.37	78.56	↑ 1.70	81.20	76.80	78.94	↑ 1.68
ZEN-CMA-CRF	Text+Lexicon+Speech	77.26	78.07	77.66	↑ 1.38	81.56	74.94	78.11	↑ 1.73
BiLSTM-MMI-CRF	Text+Speech	66.63	64.78	65.82	↑ 2.33	77.78	59.11	67.17	↑ 2.74
BERT-MMI-CRF	Text+Speech	75.37	79.62	77.44	↑ 2.05	80.95	74.98	77.85	↑ 1.86
MacBERT-MMI-CRF	Text+Speech	76.75	80.91	78.77	↑ 1.91	81.18	77.21	79.14	↑ 1.88
ZEN-MMI-CRF	Text+Lexicon+Speech	76.30	79.45	77.84	↑ 1.56	81.11	75.36	78.13	↑ 1.75
BiLSTM-M3T	Text+Speech	**69.85**	**66.24**	**68.00**	↑ 4.51	**79.17**	**60.39**	**68.52**	↑ 4.09
BERT-M3T	Text+Speech	**77.71**	**80.60**	**79.13**	↑ 3.74	**83.46**	**75.81**	**79.45**	↑ 3.46
MacBERT-M3T	Text+Speech	**78.74**	**82.02**	**80.35**	↑ 3.49	**83.99**	**77.46**	**80.59**	↑ 3.33
ZEN-M3T	Text+Lexicon+Speech	**78.66**	**79.78**	**79.21**	↑ 2.93	**82.99**	**76.41**	**79.57**	↑ 3.19

Table 3: Precision (%) , Recall (%) and F1 score (%) of baselines and our proposed method on CNERTA. Δ means the points higher than the corresponding baselines without using the acoustic modality.

7.1 Implementation Details

LSTM-Based Baselines: We use the 50-dimensional character embeddings, which are pre-trained on Chinese Giga-Word [*] using word2vec (Mikolov et al., 2013). The dimensionality of LSTM hidden states is set to 300 and the initial learning rate is set to 0.001. We train the models using 100 epochs with a batch size of 16.

Lexicon: The lexicon used in Lattice-LSTM is the same as Zhang and Yang (2018) and the lexicon used in ZEN is the same as Diao et al. (2020). Due to low speed in training and inference, we only employ Lattice-LSTM in unimodal settings.

Pretrained Language Model Fine-Tuning: We use the base models of BERT (Devlin et al., 2019), MacBERT (Cui et al., 2020) and ZEN (Diao et al., 2020). The initial learning rate of pretrained language model is set to 1×10^{-5}. We fine-tune models using 10 epochs with a batch size of 16.

Computing Infrastructure: All experiments are conducted on an NVIDIA GeForce RTX 2080 Ti (11 GB of memory).

7.2 Main Results

Table 3 shows the results of baselines and our proposed model on CNERTA. From the table, we find:

(1) Introducing the acoustic modality can significantly boost the performance of the character-based models, such as BiLSTM-CRF, BERT-CRF and MacBERT-CRF. With the simple CMA module to introduce the acoustic modality, there is a more than 1.6% improvement in both flat NER and nested NER. Furthermore, by using the M3T model to leverage the acoustic modality, a more than 3% improvement can be brought in all cases. These experimental results demonstrate the effectiveness of introducing the acoustic modality in character-based NER models.

(2) Introducing the acoustic modality can improve the performance of lexicon-based models, such as ZEN-CRF. By introducing the acoustic modes in ZEN-CRF with the CMA module, the performance in flat NER and nested NER can be improved by 1.38% and 1.73%, respectively. Armed with the M3T model, the performance in flat NER and nested NER can be further improved by 2.93% and 3.19%. Although not as significant as the improvement of the character-based models, these

[*] https://catalog.ldc.upenn.edu/LDC2011T13

Sentence	Gold	BERT-M3T	BERT-CRF

沙特阿拉伯选手马斯拉赫以四十三秒九三预获得预赛第一
(Maslakh, from Saudi Arabia, won the first place in the preliminary contest with 43.93 seconds)
	沙特阿拉伯(LOC)
	马斯拉赫(PER)
		沙特阿拉伯(LOC)
		马斯拉赫(PER)
			沙特(LOC)
			阿拉伯(LOC)
			马斯拉赫(PER)

与她在首都机场吃了一碗牛肉面有很大关系 (It has a lot to do with a bowl of beef noodles eaten at the Capital Airport)	首都机场(LOC)	首都机场(LOC)	首都(LOC)
国际米兰日文官方推特公布了选手抵达时的照片 (Inter Milan's official Japanese Twitter released photos of the players when they arrived)	国际米兰(ORG)	国际米兰(ORG)	国际米兰日文(ORG)

卡巴里罗在毕尔巴鄂掷出了七十米六五的好成绩
(Kabariro threw a good result of 70.65m in Bilbao)
	卡巴里罗(PER)
	毕尔巴鄂(LOC)
		卡巴里罗(PER)
		毕尔(LOC)
		巴鄂(PER)
			卡巴里罗(PER)
			毕尔巴鄂(LOC)

Table 4: Case studies to illustrate the effectiveness of introducing the acoustic modality. Note that both BERT-M3T and BERT-CRF are trained in flat NER settings.

Structure	Model	# Type Error	# Boundary Error	Total
Flat NER	BERT-CRF	97 (10.68%)	811 (89.32%)	908
	BERT-M3T	94 (11.93%)	694 (88.07%)	788
Nested NER	BERT-CRF	125 (13.79%)	781 (86.20%)	906
	BERT-M3T	129 (15.21%)	719 (84.78%)	848

Table 5: The statistics of different errors that occur in the output of NER models on the development set.

results still prove that the acoustic modality can provide lexicon-based models with some information that does not contain in the large-scale lexicon.

(3) Our proposed method (M3T) can achieve the SoTA results on CNERTA. Compared with CMA (Tsai et al., 2019) and MMI (Yu et al., 2020), there is a significant improvement. We conjecture that is due to that the monotonic alignment between the acoustic modality and the textual modality is captured by the masked CTC loss and armed with this alignment, precise word boundary information contained in speech is leveraged by the model.

7.3 Error Analysis

As NER models established here are not yet as accurate as one would hope, some analyses of the errors that occur in the output of NER models are performed. We divide the error into type error and boundary error. The type error is defined as that the boundary of the predicted entity is correct but the predicted type is wrong, and the other errors are classified as boundary errors. The statistics of boundary errors and type errors are shown in Table 5. From the table, we find that: (1) Errors are mainly caused by mistakenly locating boundaries of entities. Therefore, discovering entity boundaries is the main challenge in Chinese NER. (2) Leveraging the acoustic modality can effectively reduce boundary errors. In nested NER, the number of errors decreases from 906 to 848, totally owning to the reduction of boundary errors, but the number of type errors increases, which may be due to overfitting or some random factors.

7.4 Case Studies

To visually show the effectiveness of introducing the acoustic modality, case studies on comparing the output of BERT-CRF and BERT-M3T are present in Table 4. From the table, we can observe that: without the acoustic modality, BERT-CRF is prone to locate some ambiguous entities mistakenly, such as "沙特阿拉伯" (Saudi Arabia), "首都机场"(Capital Airport), "国际米兰" (Inter Milan). But armed with the acoustic modality, these entities are located with complete accuracy. In the last case, BERT-M3T makes some mistakes. We listen to the corresponding audio clip and find that there is a long pause between "毕尔" and "巴鄂".

8 Conclusion and Future Work

In this paper, we explore Chinese multimodal NER with both textual and acoustic contents. To achieve this, we construct a large-scale manually annotated multimodal NER dataset，named CNERTA. Based on this dataset, we establish a family of baseline models. Furthermore, we propose a simple multimodal multitask method by introducing a speech-to-text alignment auxiliary task. Through extensive experiments, we prove that Chinese NER models can benefit from introducing the acoustic modality and our proposed model is effective.

In the future, we are interested in mining other information contained in speech, such as rhythm, emotion, pitch, accent and stress, to boost NER. Meanwhile, we will also work on designing some speech-text pretraining tasks for building a large-scale pretrained model with multimodal capabilities.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. We also thank Zhixing Tian, Tao Wang and Ye Bai for helpful suggestions.

This work is supported by the National Key Research and Development Program of China (Grant No. 2020AAA0106400), the National Natural Science Foundation of China (Grant No. 61922085 and Grant No. 61976211) and Beijing Academy of Artificial Intelligence (Grant No. BAAI2019QN0301).

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, and Kazuhiko Ohe. 2009. TEXT2TABLE: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the BioNLP 2009 Workshop*. Association for Computational Linguistics.

Omer Arshad, Ignazio Gallo, Shah Nawaz, and Alessandro Calefati. 2019. Aiding intra-text representations with visual context for multimodal named entity recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*.

Meysam Asgari-Chenaghlu, M Reza Feizi-Derakhshi, Leili Farzinvash, and Cina Motamed. 2020. A multimodal deep learning approach for named entity recognition from social media. *arXiv preprint arXiv:2001.06888*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2018. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2020. A caption is worth a thousand images: Investigating image captions for multimodal named entity recognition. *arXiv preprint arXiv:2010.12712*.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: Pre-training Chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Ruixue Ding, Pengjun Xie, Xiaoyan Zhang, Wei Lu, Linlin Li, and Luo Si. 2019. A neural multi-digraph model for chinese ner with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuanjing Huang. 2019. A lexicon-based graph neural network for Chinese NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*.

Henry K Kwok and Douglas L Jones. 2000. Improved instantaneous frequency estimation using an adaptive short-time fourier transform. *IEEE transactions on signal processing*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*.

Mohan Li, Min Liu, and Hattori Masanori. 2019. End-to-end speech recognition with adaptive computation steps. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019a. An encoding strategy based word-character LSTM for Chinese NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2020. Simplify the usage of lexicon in Chinese NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Diego Mollá, Menno van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian Language Technology Workshop 2006*.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the*

2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*.

Nanyun Peng and Mark Dredze. 2016a. Improving named entity recognition for Chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Nanyun Peng and Mark Dredze. 2016b. Improving named entity recognition for Chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Sourabh Ravindran, C Demirogulu, and David V Anderson. 2003. Speech recognition using filter-bank features. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*.

Changhao Shan, Chao Weng, Guangsen Wang, Dan Su, Min Luo, Dong Yu, and Lei Xie. 2019. Component fusion: Learning replaceable language model component for end-to-end speech recognition system. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Lin Sun, Jiquan Wang, Yindu Su, Fangsheng Weng, Yuxuan Sun, Zengwei Zheng, and Yuanyi Chen. 2020. RIVA: A pre-trained tweet multimodal model based on text-image relation for multimodal NER. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Zhengkun Tian, Jiangyan Yi, Ye Bai, Jianhua Tao, Shuai Zhang, and Zhengqi Wen. 2020. Synchronous transformers for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2004. ACE 2005 multilingual training corpus. In *LDC*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Fangzhao Wu, Junxin Liu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2019. Neural chinese named entity recognition via cnn-lstm-crf and joint training with word segmentation. In *The World Wide Web Conference*.

Yan Xu, Yining Wang, Tianren Liu, Jiahua Liu, Yubo Fan, Yi Qian, Junichi Tsujii, and Eric I Chang. 2013. Joint segmentation and named entity recognition using dual decomposition in chinese discharge summaries. *Journal of the American Medical Informatics Association*.

Mengge Xue, Bowen Yu, Tingwen Liu, Yue Zhang, Erli Meng, and Bin Wang. 2020. Porous lattice transformer encoder for Chinese NER. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.