Correcting Pronoun Homophones with Subtle Semantics in Chinese Speech Recognition

Zhaobo Zhang, Rui Gan, Pingpeng Yuan, Hai Jin ✉
National Engineering Research Center for Big Data Technology and System
Service Computing Technology and System Laboratory
Cluster and Grid Computing Laboratory
Huazhong University of Science and Technology, Wuhan, China
zhang_zb, ruigan, ppyuan, hjin@hust.edu.cn

Abstract

Speech recognition is becoming prevalent in daily life. However, due to the similar semantic context of the entities and the overlap of Chinese pronunciation, the pronoun homophone, especially "他/她/它 (he/she/it)", (their pronunciation is "Tā") is usually recognized incorrectly. It poses a challenge to automatically correct them during the post-processing of Chinese speech recognition. In this paper, we propose three models to address the common confusion issues in this domain, tailored to various application scenarios. We implement the language model, the LSTM model with semantic features, and the rule-based assisted Ngram model, enabling our models to adapt to a wide range of requirements, from high-precision to low-resource offline devices. The extensive experiments show that our models achieve the highest recognition rate for "Tā" correction with improvements from 70% in the popular voice input methods up to 90%. Further ablation analysis underscores the effectiveness of our models in enhancing recognition accuracy. Therefore, our models improve the overall experience of Chinese speech recognition of "Tā" and reduce the burden of manual transcription corrections.

Keywords: Chinese speech recognition, natural language understanding, Chinese spelling correction

1. Introduction

Numerous studies on Chinese speech-to-text technology have been conducted. Traditional speech recognition systems typically rely on rule-based methods, converting speech signals through phoneme models and linguistic models (Fukuda et al., 2023). However, these methods often face challenges when dealing with Chinese because it is a syllabic language with a large number of homophones (Zhao et al., 2019; Gao et al., 2019).

Chinese spelling correction (CSC) is a key focus in automatic language recognition. Most errors arise from homophones. To tackle this, many studies and commercial speech software employ deep neural networks like RNN (Saon et al., 2021) and LSTM (Dokuz and Tufekci, 2021) for audio-to-text processing. The prevalent CSC model uses an end-to-end structure with pre-trained BERT (Zhang et al., 2021) to effectively differentiate between phonetics and characters. These models adeptly capture intricate speech features (Nassif et al., 2019).

However, the specific reference of pronouns may be related to multiple different noun entities in the context. Even with the use of advanced deep learning models, understanding the referential meaning of pronouns (Yu and Zhao, 2022) remains a challenge. Chinese faces even more formidable challenges in this regard, especially the third-person pronouns "他" (he), "她" (she), and "它" (it) which all share the same pronunciation and similar semantics. By checking the recognition result of existing voice input methods or models, we can determine that the accuracy of various current speech-to-text methods in recognizing "他/她/它 (he/she/it)" still needs to be improved. We are the first to notice this issue.

A common method for post-processing converted text is to use a language model to correct the recognition results (Karita et al., 2019). By analyzing contextual information, language models can determine the specific nouns or phrases being referred to. Additionally, some studies combine external knowledge bases and semantic analysis to improve recognition accuracy (Han et al., 2022). For example, enhancing end-to-end contextual speech recognition can be achieved by using fine-grained knowledge to alleviate confusion in context-specific phrases.

We are inspired by Chinese spelling correction and propose the "Tā Correct" scheme, as shown in Fig. 1, to improve the recognition effect. The scheme takes the sentence converted from speech, its context, and the "Tā" to be inspected as input into "Tā" classification models (Sec. 2) and outputs the correctly identified form of "Tā".

We offer three "Tā" classification models which incorporates and draws on the strengths of existing work, to confusion around the pronouns "他" (he), "她" (she), and "它" (it). Our models specifically address the widespread issue of third-person pronoun confusion in automatic speech recognition, a focus that separates them from previous solutions.

Firstly, we propose a strategy for fine-tuning language model through a "Tā" masked language modeling task to generate a focused attention on

LREC-COLING 2024, pages 4047–4058
20-25 May, 2024.

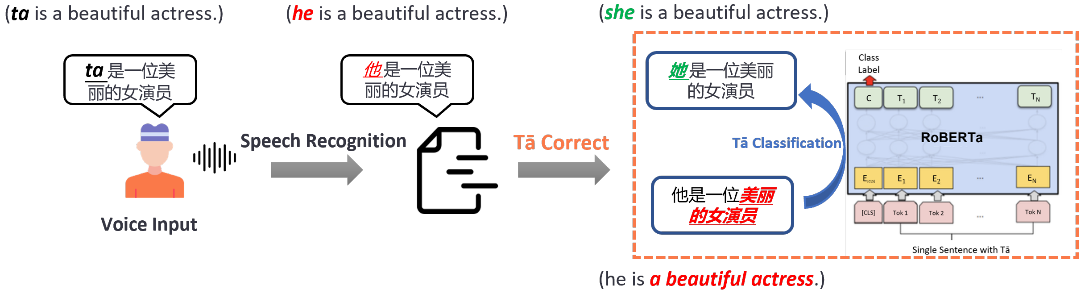


Figure 1: The workflow and application scenarios of our approach

"Tā". In this approach, by masking and predicting "Tā" in sentences, we improve the understanding of the RoBERTa (A Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019) model about the contextual environment of "Tā" in Chinese. This is achieved by optimizing the model's performance through extensive pre-training data.

Considering the phonetic and semantic features of the Chinese characters (Duan et al., 2019), we design another deep learning model that integrates various features with *long-short-term memory* (LSTM) to capture potential characteristics of Chinese phonetics and semantics, achieving high accuracy "Tā" recognition with relatively low computational costs.

Additionally, we present a low-consumption solution suitable for offline scenarios. We present an Ngram-based model that integrates a rule-based matching strategy, allowing for a more detailed capture of local information within the text. This strategy achieves the objectives of low resource utilization and robustness at the cost of slight reductions in accuracy.

In addition to the experiments comparing baselines, we evaluate the contributions of various factors through ablation experiments and measure their contribution to the improvement. Furthermore, we also conduct quantitative analysis to explore the impact of the number of "Tā", proportion of "Tā", and so on. Our models demonstrate their effectiveness and applicability with empirical validation when they address the confusion surrounding "他/她/它 (he/she/it)" in Chinese speech-to-text conversion. This has profound implications for advancing the maturity and application of Chinese speech recognition technology. Code and data: https://github.com/CGCL-codes/TaCorrect.

2. "Tā" Classification Models

2.1. Language Model

To construct a "Tā" correction model capable of effectively distinguishing between "他/她/它 (he/she/it)", we employ *masked language model* (MLM) task and design a "Tā" classification model grounded in the RoBERTa language model (Liu et al., 2019). It integrates a cloze MLM task, namely, **TaR**(oBERTa) in Fig. 2.

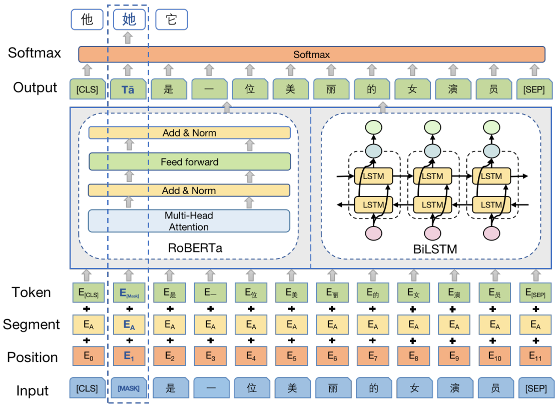


Figure 2: Architectural illustration for **TaR** & **TaL**

The primary objective of TaR is to address the challenge of recognizing the Chinese pronouns "Tā". The cloze task serves as an intermediary step to amplify the model's sensitivity to context, allowing it to learn distinct contextual features of the different "Tā" forms and thereby enhance "Tā" classification accuracy.

The RoBERTa model operates through several computational steps. First, it utilizes a multi-head attention mechanism, described by:

$$\mathit{Attention}(Q, K, V) = \mathit{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \quad (1)$$

where Q, K, and V stand for the query, key, and value matrices respectively, and d_k is the dimension of the key.

It implements layer normalization, represented by:

$$\mathit{Normalization}(x) = \gamma \times \frac{x - \mu}{\sigma} + \beta \quad (2)$$

where μ and σ denote the mean and standard deviation of the layer's output, and γ and β are trainable parameters.

Finally, it encompasses a feed-forward neural network expressed as:

$$h = \mathit{ReLU}(x \cdot W_1 + b_1) \cdot W_2 + b_2 \quad (3)$$

where W_1 and W_2 are weight matrices and b_1 and b_2 are bias vectors.

From this, the model output is further connected to a fully connected layer, which tailors it for outputs suitable for our specific "Tā" classification task:

$$o = W \cdot h + b \quad (4)$$

Lastly, by focusing particularly on the outputs of characters tagged as '[MASK]' (where the masked character represents "Tā") when inputted into the model, we achieve more accurate detection of "他/她/它 (he/she/it)" within sentences derived from speech-to-text conversions.

2.2. LSTM Fusion of Sound and Semantic

To integrate the phonetic and semantic characteristics of Chinese, we develop a model based on BiLSTM networks and the attention mechanism, named the phonosemantic fusion "Tā" classification model. Beyond conventional text information, this model, namely, **TaL**(STM) also incorporates *part-of-speech* (POS) (Seo et al., 2023) and pinyin (Duan et al., 2019; Xu et al., 2022; Tan et al., 2022) (Fig. 3) data to achieve enhanced performance. The model architecture is similar to Fig. 2, but the inputs should be replaced by Fig. 3.

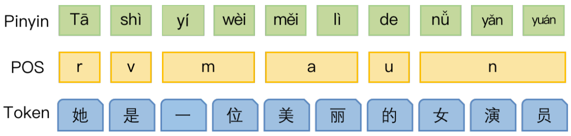


Figure 3: Pinyin and POS of Chinese

The Chinese language contains homophones, words that sound the same but have different meanings. The character "Tā" is a prime example, where it can mean "他" (he), "她" (she), or "它" (it) but they all share the same pronunciation in Mandarin. Therefore, using pinyin alone might not be sufficient to differentiate among them. However, combining pinyin with POS can be very effective for "Tā" classification.

For instance, certain POS and pinyin patterns in a sentence might suggest gender. Adjectives or actions associated with a particular gender can hint towards whether "他" (he) or "她" (she). When analyzing the surrounding words and their POS, the context can be better understood. On the other hand, pinyin provides phonetic insight, which can be crucial, especially when dealing with spoken language data, where tone or pronunciation nuances might give away the intended meaning of "Tā".

Let x_t be the input feature vector at time step t, comprising text information x_t^{text}, POS data x_t^{pos}, and pinyin details x_t^{pinyin}. The composite can be expressed as:

$$x_t = [x_t^{text}; x_t^{pos}; x_t^{pinyin}] \quad (5)$$

where $[;]$ symbolizes vector concatenation.

During the encoding phase, a bidirectional LSTM is employed to process the concatenated embeddings:

$$h_t = \mathit{BiLSTM}(x_t) \quad (6)$$

where h_t signifies the hidden state at moment t.

Subsequently, in the decoding phase, we apply the attention mechanism to weight the input features. The attention weight a is computed based on the current decoder's hidden state s_{t-1} and the relative importance of each encoder hidden state h_t:

$$\begin{aligned} e_{t,i} &= \mathit{align}(s_{t-1}, h_i) = V^T \tanh(W_s s_{t-1} + W_h h_i) \\ a_{t,i} &= \mathit{Softmax}(e_{t,i}) = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})} \end{aligned} \quad (7)$$

here W_s, W_h, and V are the parameters of the attention mechanism, which are learned during the training phase.

Through the attention weights, the context vector c_t is computed as:

$$c_t = \sum_i a_{t,i} h_i \quad (8)$$

This context vector c_t, in conjunction with the decoder's hidden state s_{t-1}, predicts the current output y_t:

$$y_t = g(s_{t-1}, c_t) = \text{softmax}(W_o[s_{t-1}, c_t]) \quad (9)$$

where W_o stands for the weight parameters of the output layer.

2.3. Rule-based Assisted Ngram Model

Given that the specific use of speech recognition can have resource-constrained application environments, such as offline mobile devices, we consider a certain level of accurate recognition of "Tā" with this limitation. To achieve this, we propose an Ngram-based model within rule-based matching, namely, **TaN**(gram). We use a character-level Ngram-based model to calculate initial probabilities and introduce a factor α to adjust the original probabilities based on rules to obtain the final results. Specifically, we harness Ngrams as features and employ a *Multi-Layer Perceptron* (MLP) architecture for classification in Fig. 4.

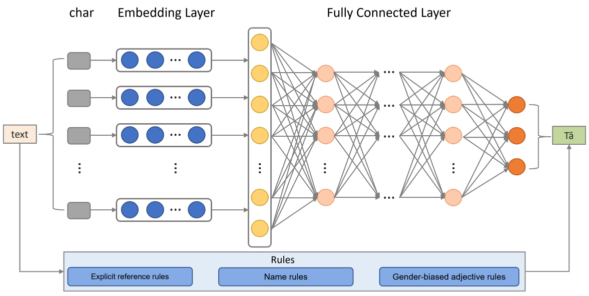


Figure 4: Architectural illustration for **TaN**

Ngram Extraction: For any given input sentence s, Ngrams are generated by sliding a window of size n over the text. These extracted Ngrams serve as features that encapsulate local context around each "Tā" occurrence.

Feature Vector Construction: Unlike the sparse one-hot encoding, embeddings provide a dense representation that captures semantic nuances. Given an Ngram of length N leading up to "Tā" in a sentence s, we represent each character c_i with its embedding:

$$\textit{Embed}(c_i) = E \cdot c_i \tag{10}$$

here $E \in \mathbb{R}^{|V| \times d}$, $|V|$ signifies the size of the character vocabulary, c_i is the one-hot encoded vector of the character.

To form the feature vector F for the segment Ngram leading up to "Tā", we compute the average of the embeddings of its characters:

$$F = \left[\textit{Embed}(c_1); ...; \textit{Embed}(c_N)\right] \tag{11}$$

MLP Classifier: The feature vectors are then fed into an MLP, a neural network model with:

- *Input Layer:* Accepts the Ngram feature vectors F from Eq. 11.
- *Hidden Layers:* Comprises multiple layers of neurons with activation functions (e.g., ReLU) that introduce non-linearity and enable the model to learn complex patterns.
- *Output Layer:* Produces probabilities for each form of "Tā". A softmax activation ensures that the outputs sum to one, representing probability distributions.

Prediction: For a new input sentence s', Ngram features are extracted and transformed into a feature vector. This vector is input into the trained MLP classifier, yielding:

$$\hat{\mathbf{C}} = \textit{MLP}(s') \tag{12}$$

where $\hat{\mathbf{C}} \in \mathbb{R}^3$ denotes the predicted probabilities of "他/她/它 (he/she/it)".

Rule-based Matching Correction: Certain words or characters in the speech text can assist the model in better handling the "Tā" classification problem. To further improve classification accuracy, we design a lightweight rule engine specifically for handling the "Tā" classification, calculating the adjustment factor α (grid search).

We employ three types of rules to correct the original probabilities of Ngrams, including explicit reference rules, name rules, and gender-biased adjective rules. **Explicit reference rules** involve words or characters that explicitly distinguish "他" (he), "她" (she), and "它" (it), such as "男" (male) for "他" (he), "女" (female) for "她" (she), and "兔子" (rabbit) for "它" (it).

Additionally, the distinction between "他" (he) and "她" (she) often involves gender differences. In the **name rules**, some commonly used characters in male and female names can help distinguish gendered "Tā" (Du et al., 2020). For example, "伟" (great) is commonly found in male names, while "丽" (beautiful) frequently appears in female names.

Gender-biased Adjective Rules: Adjectives with gender bias also carry implicit meanings (Zhu and Liu, 2020). Word like "漂亮" (beautiful) often describes "她" (she), while "英俊" (handsome) commonly describes "他" (he).

By analyzing sentences based on rules $[\alpha_1, \alpha_2, \alpha_3]$ and making a second adjustment to the original Ngram-based model output probabilities in Eq. 12, we predict "Tā" as follows:

$$\textit{Ta} = \arg\max_i(\alpha_i \times \hat{C}_i) \tag{13}$$

This strategy effectively classifies "他" (he), "她" (she), and "它" (it) in resource-constrained environments such as mobile devices while maintaining a certain level of recognition accuracy.

3. Experiments

3.1. Experimental Setup

We use some self-constructed multi-source datasets primarily comprised of sentences containing the pronoun "Tā". The dataset includes three subsets:

- **Weibo**[1]: Sourced from the Sina Weibo social media platform, this subset comprises daily opinions and viewpoints posted by ordinary users, totaling over 360,000 sentences, of which 57,384 sentences contain "Tā".
- **Smp**[2]: It consists of Weibo data covering various topics related to the COVID-19 pandemic

[1] https://github.com/SophonPlus
[2] https://smp2020ewect.github.io

and is approximately 5,000 sentences long, of which 3,020 sentences contain "Tā".

- **Tieba**[3]: Derived from forum posts, this subset includes multiturn conversation data with some noise, over 2,320,000 sentences, of which 946,969 sentences contain "Tā".

To enhance the diversity and complexity of the experiment, we generate synthetic speech data for these sentences using machine-generated pronunciation through speech synthesis techniques. We divide the train and test set in 7:3 ratio.

3.2. Baselines

While evaluating the performance of our model, we select two main categories of benchmark models for comparative experiments:

- Open-source speech recognition models: including PaddleSpeech (Bai et al., 2022; Zheng et al., 2021; Zhang et al., 2022) and Whisper (Radford et al., 2023).

- Commercial input methods: including Baidu Speech Recognition[4] (Amodei et al., 2016) and Xunfei Automatic Speech Recognition[5].

These benchmark models cover a variety of popular methods, enabling us to comprehensively evaluate the superiority of our model.

3.3. Evaluation Protocol

We define the calculation formulas for the following three evaluation mertics.

3.3.1. In-Sentence Accuracy (ISA)

For this metric, we first calculate the accuracy for each sentence containing "Tā". Given N such sentences, for the i^{th} sentence S_i with n_i occurrences of "Tā", and if the model correctly predicts c_i of them, the accuracy Acc_i of the sentence is:

$$Acc_i = \frac{c_i}{n_i} \tag{14}$$

Then, we calculate the In-Sentence Accuracy as the average of all Acc_i where $i = 1, 2, \ldots, N$:

$$Acc_{isa} = \sum_{i=1}^{N} Acc_i \times 100 \tag{15}$$

3.3.2. Whole Sentence Accuracy (WSA)

This metric calculates the probability that all "Tā" within a sentence are correctly predicted. Assuming there are M sentences and the number of sentences where all "Tā" are predicted correctly is C, the Whole Sentence Accuracy Acc_{wsa} is:

$$Acc_{wsa} = \frac{C}{M} \times 100 \tag{16}$$

3.3.3. "Tā" Conversion Accuracy (TCA)

This metric measures the proportion of words predicted as "Tā" that are predicted correctly. If the total number of predictions is N and the number of correct predictions is T, Acc_{tca} is:

$$Acc_{tca} = \frac{T}{N} \times 100 \tag{17}$$

3.4. Experimental Analysis

Model Performance. The results in Table 1 unequivocally demonstrate substantial performance improvements of our models in diverse datasets. TaR achieves the most significant performance boost with its superior natural language comprehension and nuanced contextual processing, reaching a peak TCA of 0.91 on Weibo and Tieba. Compared to baselines, all three metrics on average exceed 0.8 in every dataset.

TaL incorporates both speech and semantic features. While slightly behind TaR in performance enhancement, it still exhibits notable improvements. In particular, it achieves a performance gain of 0.1 to 0.2 on Weibo, compared to baselines. Impressively, TaL accomplishes this while requiring less computational and storage resources (details in Appendices 11).

The performance of TaN is similar to that of TaL, but pleasantly exceeds that of TaL in terms of computational speed. Furthermore, although TaL lags behind baselines on certain metrics within the Tieba and Smp datasets, its overall performance still outperforms the baselines. This could be attributed to the fact that our models are primarily trained on Weibo data. The Ngram model excels at capturing local text information, but may have weaker generalization capabilities.

Data Discrepancy. In terms of performance across different datasets, we observe that our model exhibits relatively poorer performance on Weibo, while it excels on the Tieba. The results are correlated with sentence lengths in the datasets, where the Tieba dataset primarily consists of medium to short sentences, while the Weibo dataset contains a higher proportion of longer sentences. This observation is further substantiated by our ablation analysis in Sec. 3.6.1.

[3] https://github.com/codemayq
[4] https://ai.baidu.com/tech/speech
[5] https://www.xfyun.cn/services

Model	Weibo			Tieba			Smp		
	ISA	WSA	TCA	ISA	WSA	TCA	ISA	WSA	TCA
Paddle + ∅	0.66	0.59	0.66	0.82	0.76	0.80	0.70	0.63	0.66
+ TaR	0.87 ↑ 0.21	0.84 ↑ 0.25	0.90 ↑ 0.24	0.87 ↑ 0.05	0.85 ↑ 0.09	0.90 ↑ 0.10	0.85 ↑ 0.15	0.84 ↑ 0.21	0.88 ↑ 0.22
+ TaL	0.80 ↑ 0.14	0.76 ↑ 0.17	0.84 ↑ 0.18	0.81 ↓ -0.01	0.80 ↑ 0.04	0.86 ↑ 0.06	0.75 ↑ 0.05	0.72 ↑ 0.09	0.80 ↑ 0.14
+ TaN	0.78 ↑ 0.12	0.73 ↑ 0.14	0.81 ↑ 0.15	0.81 ↓ -0.01	0.76 -	0.82 ↑ 0.02	0.73 ↑ 0.03	0.67 ↑ 0.04	0.73 ↑ 0.07
Whisper + ∅	0.69	0.62	0.67	0.81	0.78	0.82	0.71	0.66	0.78
+ TaR	0.87 ↑ 0.18	0.83 ↑ 0.21	0.90 ↑ 0.23	0.87 ↑ 0.06	0.85 ↑ 0.07	0.90 ↑ 0.08	0.82 ↑ 0.11	0.79 ↑ 0.13	0.85 ↑ 0.07
+ TaL	0.78 ↑ 0.09	0.73 ↑ 0.11	0.82 ↑ 0.15	0.81 -	0.80 ↑ 0.02	0.86 ↑ 0.04	0.75 ↑ 0.04	0.73 ↑ 0.07	0.81 ↑ 0.03
+ TaN	0.75 ↑ 0.06	0.71 ↑ 0.09	0.81 ↑ 0.14	0.80 ↓ -0.01	0.74 ↓ -0.04	0.81 ↓ -0.01	0.73 ↑ 0.02	0.68 ↑ 0.02	0.74 ↓ -0.04
Baidu + ∅	0.66	0.51	0.64	0.74	0.62	0.73	0.66	0.60	0.62
+ TaR	0.86 ↑ 0.20	0.82 ↑ 0.31	0.89 ↑ 0.25	0.88 ↑ 0.14	0.86 ↑ 0.24	0.91 ↑ 0.18	0.82 ↑ 0.16	0.81 ↑ 0.21	0.86 ↑ 0.24
+ TaL	0.79 ↑ 0.13	0.74 ↑ 0.23	0.82 ↑ 0.18	0.82 ↑ 0.08	0.80 ↑ 0.18	0.86 ↑ 0.13	0.73 ↑ 0.07	0.70 ↑ 0.10	0.78 ↑ 0.16
+ TaN	0.77 ↑ 0.11	0.73 ↑ 0.22	0.82 ↑ 0.18	0.81 ↑ 0.07	0.75 ↑ 0.13	0.82 ↑ 0.09	0.71 ↑ 0.05	0.65 ↑ 0.05	0.72 ↑ 0.10
Xunfei + ∅	0.71	0.57	0.72	0.82	0.74	0.81	0.69	0.58	0.65
+ TaR	0.88 ↑ 0.17	0.84 ↑ 0.27	0.91 ↑ 0.19	0.88 ↑ 0.06	0.87 ↑ 0.13	0.91 ↑ 0.10	0.81 ↑ 0.12	0.79 ↑ 0.21	0.85 ↑ 0.20
+ TaL	0.81 ↑ 0.10	0.77 ↑ 0.20	0.84 ↑ 0.12	0.86 ↑ 0.04	0.80 ↑ 0.06	0.85 ↑ 0.04	0.72 ↑ 0.03	0.69 ↑ 0.11	0.78 ↑ 0.13
+ TaN	0.79 ↑ 0.08	0.75 ↑ 0.18	0.86 ↑ 0.14	0.80 ↓ -0.02	0.76 ↑ 0.02	0.83 ↑ 0.02	0.72 ↑ 0.03	0.65 ↑ 0.07	0.73 ↑ 0.08

Table 1: The performance of our models and baselines

It should be noted that our models not only improve the performance of recognizing the word "Tā" within single sentences but also significantly enhance the overall accuracy of recognizing "Tā" in the text.

Experimental Conclusion. Finally, considering various performance metrics, our model has successfully enhanced the accuracy of TCA, leading to further increases in ISA and WSA. It is worth noting that, due to the complexity of these metrics, the improvement in ISA is relatively smaller when compared to WSA and TCA. This suggests that it may be a promising avenue for further research and optimization.

3.5. Ablation Analysis

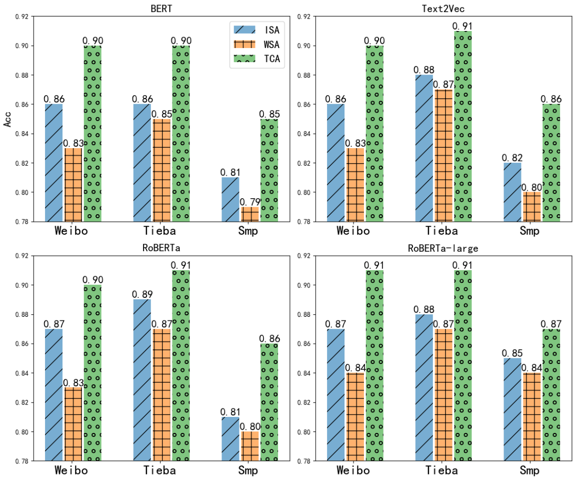


Figure 5: Role of different pre-training models

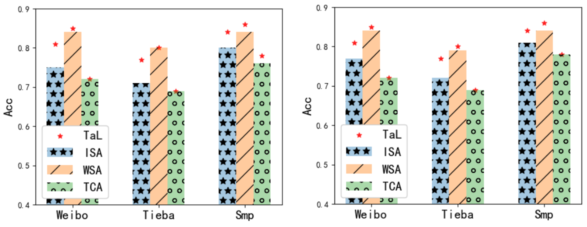
3.5.1. Role of BERT Model

Here, we use different versions of the BERT model (BERT-base (Devlin et al., 2019), Text2vec (Xu, 2023), RoBERTa (Liu et al., 2019), RoBERTa-large) as the core architecture to explore the specific impact of internal language models on the "Tā" classification task.

The experimental results in Fig. 5 indicate that BERT models can achieve relatively good performance. Among them, the RoBERTa-Large model performs well on all metrics and exhibits the most stable performance across various datasets, especially for ISA and WSA on Smp.

3.5.2. Role of Linguistic Features

We conduct ablation experiments on the LSTM model, removing linguistic features such as pinyin, part-of-speech, and other semantic features, to better understand the contribution of these features to the model's performance on the "Tā" classification task.



(a) TaL (POS)

(b) TaL (Pinyin)

Figure 6: Ablation of TaL on Xunfei

Both pinyin and POS play an important role, but in comparison, pinyin plays a more prominent role than POS, which is reflected in Fig. 6(a). It loses more accuracy on both ISA and TCA metrics.

3.5.3. Effect of the *N* and Rules

For the Ngram model, a critical variable is the length of its sliding window *N*. We use different values of *N* to search for the optimal parameter.

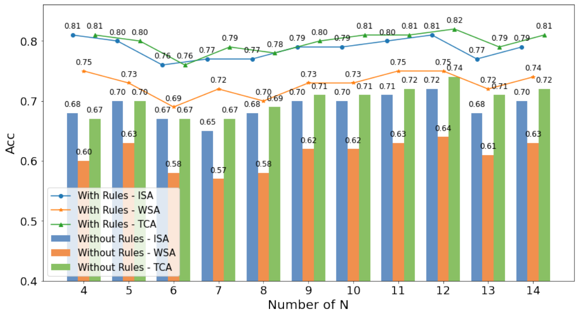


Figure 7: *N* and rules on Baidu-Tieba

As shown in Fig. 7, smaller values of N initially perform better. However, as *N* increases, there is a trend where the performance decreases first and then increases again, with the best results observed around *N*=12. For the "Tā" classification, shorter values of *N* can retrieve some fixed paradigms, but larger *N* can effectively integrate potential relationships in the context, resulting in improved performance.

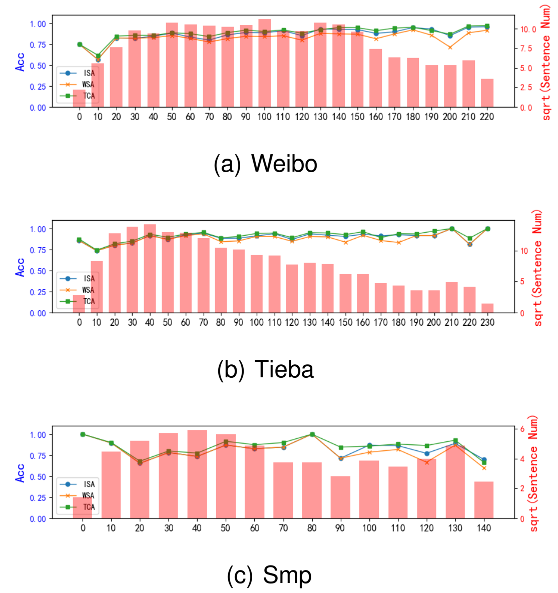
3.6. Quantitative Analysis

It is crucial to explore the changes caused by the differences in the datasets. Therefore, we define four distinct factors of datasets and analyze their relationship with performance.

3.6.1. Sentence Length

Sentence (Utterance, here we also use sentence to represent it) length impacts the model's contextual understanding. While longer sentences offer more contextual hints for "Tā" classification, excessively long ones can introduce noise or risk overfitting. Thus, measuring the correlation between sentence length and performance is vital.

Fig. 8 clearly indicates that sentence lengths are distributed roughly between 0 to 250 characters, with medium-length sentences being more frequent. Sentences of different lengths show varying stability when tested with the model. Specifically, medium-length sentences, give their richness in contextual information and moderate uncertainty, generally exhibit robust and competitive performance. This observation aligns with our prior assumption that sentences of extreme (too long or too short) lengths might perform inconsistently due to noise or insufficient information.



(a) Weibo

(b) Tieba

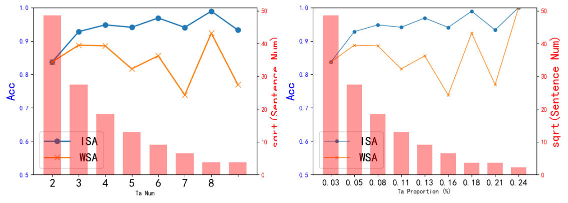
(c) Smp

Figure 8: Sentence Length

3.6.2. Quantity of "Tā"

In situations where multiple "Tā" instances appear in a single sentence, the model needs to search for contextual cues to determine the referents for each "Tā" and establish the corresponding associations within the sentence. Undoubtedly, this presents a significant challenge to the model in terms of computational complexity and its ability to comprehend the context.

Fig. 9(a) showcases that in most sentences, the count of "Tā" usually lies between 1 to 6. Overall, the model's ISA is fairly steady, and even with an increasing count of "Tā", the accuracy slightly improves. This underlines the stability and robustness of our model. However, for WSA, as the count of "Tā" rises, the performance exhibits a declining trend due to accumulating uncertainties. This means the need for further optimization in the model's contextual linkage and overall understanding when handling sentences with multiple "Tā".



(a) Quantity of "Tā"

(b) Proportion of "Tā"

Figure 9: Effect of "Tā" within a single sentence

Wanted Sentence	Recognized Sentence (Xunfei)	Corrected Sentence
她是一位美丽的女演员 She is a beautiful actress	他是一位美丽的女演员	她是一位美丽的女演员
妹妹最爱喝果汁，所以我为她留了一瓶 My sister loves juice best, so I saved a bottle for her	妹妹最爱喝果汁，所以我为她留了一瓶	妹妹最爱喝果汁，所以我为她留了一瓶
我养了一只小白兔，它最爱吃萝卜 I raised a small white rabbit, and it likes to eat radish	我养了一只小白兔他最爱吃萝卜	我养了一只小白兔它最爱吃萝卜

Table 2: Case predicted results

3.6.3. Proportion of "Tā"

The proportion of "Tā" within the entire sentence is a highly important consideration for sentence structure and clarity, enabling a more comprehensive and accurate assessment of the model's ability to understand context.

As depicted in Fig. 9(b), there is an overall upward trend, indicating that as the proportion of "Tā" in the sentence gradually increases, the model's understanding of the sentence improves. However, with an increasing proportion of "Tā", the WSA fluctuates more significantly, suggesting that when dealing with complex sentence structures, the presence of multiple "Tā" instances in a sentence may require the model to access more contextual information to accurately determine reference relationships.

3.6.4. Distribution of "Tā"

The distribution pattern of "Tā" is another key variable, helping us to understand the complexity and challenges of the classification task in a more comprehensive way. If the distribution of "Tā" is highly imbalanced, such that one category (like "他") is more frequent than others (like "她" or "它"), the model might lean towards the predominant category, leading to reduced classification accuracy. Therefore, by analyzing the distribution of "Tā", we can assess the performance of the model more accurately and make the necessary adjustments.

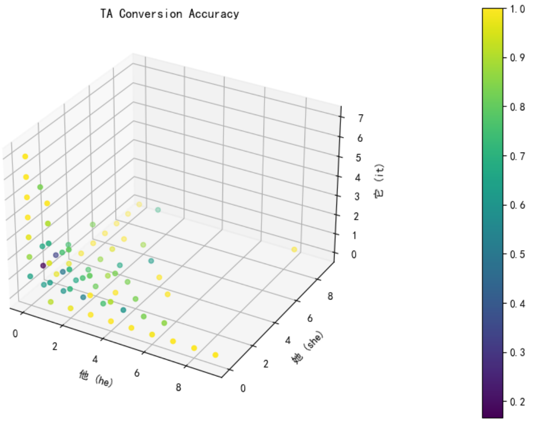


Figure 10: Distribution of "Tā"

As shown in Fig. 10, when a sentence contains only one category of "Tā", accuracy increases with the increase in the number of "Tā" because both the context information points to the same entity. However, as the proportion of mixed usage of "他" (he) and "她" (she) in the sentence increases and the quantity remains limited, gender implication information becomes ambiguous, leading to confusion and a decrease in accuracy.

3.7. Case Study

Table 2 reveals the outstanding capabilities of our model in the "Tā" classification task, demonstrating its extraordinary ability to thoroughly consider contextual information. In the first case, when "Tā" at the beginning of a sentence, it infers successfully that "Tā" refers to "她" (she) based on the object "女演员" (actress) and the modifiers "美丽的" (beautiful). In the second scenario, when the explicit differentiating word "妹妹" (sister) is present, the model accurately employs "她" (her) for reference. In case 3, where "Tā" serves as the subject of the latter half of the sentence, the model successfully determines that "Tā" refers to a rabbit by comprehensively considering the contextual references to "兔子" (rabbit) in the preceding text and the description of "吃萝卜" (eating carrots) in the subsequent text, using "它" (it) to represent the reference.

4. Related Works

4.1. Automatic Speech Recognition

End-to-end ASR systems using DNNs directly map audio to text and show enhanced performance compared to traditional models (Chiu et al., 2018). Saon (Saon et al., 2021) improved RNN Transducers' word error rate using novel integration techniques. Le (Le et al., 2020) proposed a dual-decoder Transformer for ASR tasks. Deep Speech employs a CNN+RNN design for character segmentation in speech recognition.

For Chinese speech recognition, it faces unique challenges due to its phonetic characteristics and lack of clear word boundaries. PLOME (Liu et al., 2021) is a pre-trained model tailored for the Chinese language, while SpeechTransformer

(Zhao et al., 2019) optimizes performance through attention-based mechanisms.

4.2. Homophone Pronoun Correction

In linguistics, homophones often have unique grammatical and semantic traits. For instance, English third-person pronouns (he/she/it) differ in pronunciation. In contrast, Chinese third-person pronouns share pronunciation, causing communication barriers.

Prior research has employed pronoun resolution to discern the specific form of "Tā" by understanding its context. CorefDPR (Yang et al., 2022) predicts the distribution of the types of pronouns by examining the mentions of entities and inferring the references of the omitted pronouns. Yu et al. (Yu and Zhao, 2022) provided a quantitative analysis of the frequent errors in analyzing the functional word "Tā". Lin (Lin and Yang, 2020) proposed the HAN-PL model, which uses a bidirectional attention mechanism and a new paired loss function, ensuring a more distinct learning outcome.

5. Conclusion

This paper aims to solve the problem of differentiating homophonic pronouns in Chinese, with a particular focus on third-person pronouns. For different application scenarios, we design three models with different costs to address distinct scenarios and achieve enhanced recognition of the third-person reference "Tā" in speech-to-text applications. Experimental results indicate our models exhibit commendable performance, potentially enhancing accuracy by approximately 20% when integrated with current input methods.

In the future, we will investigate more rules and techniques for more pronouns to promote the rapid development of Chinese spelling correction, with the intention of developing more portable models to benefit a broader range of applications and improving the user experience.

6. Acknowledgment

The research is supported by The National Natural Science Foundation of China under Grant Nos. 61932004 and 62072205.

7. Optional Supplementary Materials

7.1. Appendices

Here we provide some resource usage of the three models at training progress.

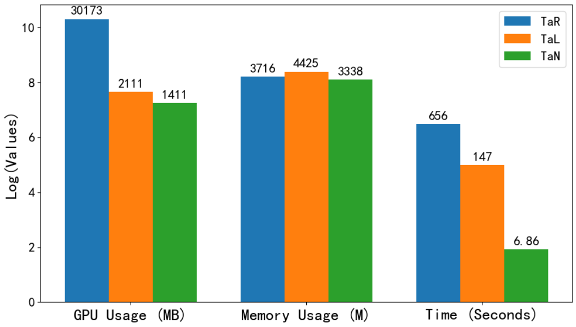


Figure 11: Comparison of Models Across Three Metrics

7.2. Ethical Considerations

Following the guidelines from *General ethical issues in Machine Learning*[6], *The Ethical Considerations of NLP*[7], *Social & Ethical Considerations in NLP Systems*[8], and *Ethics Best Practices for Natural Language Processing* (Leidner and Plachouras, 2017), we briefly summarize our ethical considerations:

Our work mainly focuses on correcting pronoun homophones in the Chinese domain, without **Bias, Fairness, Security, Privacy, Environmental Impact**, and **Discrimination** against a group or collective. Our method combines external professional Chinese linguistic knowledge and has good **Interpretability**. Moreover, we conduct multiple experiments to analyze the reasons for the effectiveness of the model and the contributions of various designs, to improve overall **Transparency** and **Credibility**.

Our data is sourced from publicly available resources on the internet and all references are cited in the paper, such as Github and other data that follows open-source licenses. Our data does not involve any personal **Privacy** or inference of personal information. Our work is dedicated to researching the potential subtle semantic between pronouns in Chinese and does not have any **Police or Military Applications**.

We promise not to use **Non-Ethical Research Methods**, and our goal is to improve the efficiency of human analysts through automation, rather than aiming to reduce labor costs.

It is imperative to highlight our commitment to inclusivity, especially concerning **individuals with visual impairments**. Ensuring our content is accessible to all, regardless of their physical abilities, is not just a moral responsibility but a testament to our dedication to universal usability. We have made concerted efforts to ensure our materials

[6] General ethical issues in Machine Learning
[7] The Ethical Considerations of NLP
[8] Social & Ethical Considerations in NLP Systems

are compatible with screen readers and have also incorporated alternative text descriptions or discriminative shape for most visual content. We understand that an inclusive approach enriches the overall user experience and demonstrates respect for the diversity of our audience.

7.3. Limitation

The motivation for our research stems from the user experience with commercial input methods. By highlighting the existence of this issue, we aim to draw attention from our peers and contribute to the improvement of Chinese speech recognition for the benefit of a wider user base. Therefore our model functions like a versatile add-on, applicable to various scenarios and compatible with any speech recognition API to enhance recognition results.

However, our approach currently has certain limitations. Firstly, its applicability needs to be expanded. This study primarily focuses on distinguishing third-person pronouns in Chinese, but the concepts presented here can be extended to a broader range of pronouns and parts of speech, an aspect we plan to explore in future work. Secondly, the models primarily adapt from generic models, lacking customization based on the specifics of the issue at hand. This is because the study aims to highlight the problem and provide a widely accessible model for improvement.

8. Bibliographical References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse H. Engel, Linxi Fan, Christopher Fougner, Awni Y. Hannun, Billy Jun, Tony Han, Patrick LeGresley, Xiangang Li, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Sheng Qian, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Chong Wang, Yi Wang, Zhiqian Wang, Bo Xiao, Yan Xie, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 173–182.

He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, and Liang Huang. 2022. A^{3}T: Alignment-aware acoustic and text pretraining for speech synthesis and editing. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1399–1411.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4774–4778.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yesim Dokuz and Zekeriya Tufekci. 2021. Mini-batch sample selection strategies for deep learning based speech recognition. *Applied Acoustics*, 171:107573.

Bingjie Du, Pengyuan Liu, and Yongsheng Tian. 2020. A quantified research on gender characteristics of chinese names in a century. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 20–30.

Dagao Duan, Shaohu Liang, Zhongming Han, and Weijie Yang. 2019. Pinyin as a feature of neural machine translation for chinese speech recognition error correction. In *Proceedings of the Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019,*, volume 11856 of *Lecture Notes in Computer Science*, pages 651–663.

Ryo Fukuda, Yuta Nishikawa, Yasumasa Kano, Yuka Ko, Tomoya Yanagita, Kosuke Doi, Mana Makinae, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. 2023. NAIST simultaneous speech-to-speech translation system for IWSLT 2023. In *Proceedings of the 20th International Conference on Spoken Language Translation, IWSLT@ACL 2023, Toronto, Canada (in-person and online), 13-14 July, 2023*, pages 330–340.

Shengxiang Gao, Dewei Kong, Zhengtao Yu, Yang Luo, Jianyi Guo, and Yantuan Xian. 2019. Chinese question speech recognition integrated with domain characteristics. *Int. J. Comput. Sci. Eng.*, 19(3):325–333.

Minglun Han, Linhao Dong, Zhenlin Liang, Meng Cai, Shiyu Zhou, Zejun Ma, and Bo Xu. 2022. Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8532–8536.

Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani. 2019. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1408–1412.

Hang Le, Juan Miguel Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3520–3533.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, EthNLP@EACL, Valencia, Spain, April 4, 2017*, pages 30–40.

Peiqin Lin and Meng Yang. 2020. Hierarchical attention network with pairwise loss for chinese zero pronoun resolution. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8352–8359.

Shulin Liu, Tao Yang, Tianchi Yue, Feng Zhang, and Di Wang. 2021. PLOME: pre-training with misspelled knowledge for chinese spelling correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2991–3000.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ali Bou Nassif, Ismail Shahin, Imtinan Basem Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518.

George Saon, Zoltán Tüske, Daniel Bolaños, and Brian Kingsbury. 2021. Advancing RNN transducer technology for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 5654–5658.

Kyusung Seo, Joonhyung Park, Jaeyun Song, and Eunho Yang. 2023. Weavspeech: Data augmentation strategy for automatic speech recognition via semantic-aware weaving. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, pages 1–5.

Minghuan Tan, Yong Dai, Duyu Tang, Zhangyin Feng, Guoping Huang, Jing Jiang, Jiwei Li, and Shuming Shi. 2022. Exploring and adapting chinese GPT to pinyin input method. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1899–1909.

Anchong Xu, Wenxin Hu, Jing Yang, and Jun Zheng. 2022. Mandarin speech recognition based on chinese syllable similarity for children. In *Proceedings of the 7th International Conference on Signal and Image Processing (ICSIP)*, pages 131–136.

Ming Xu. 2023. Text2vec: Text to vector toolkit.

Jingxuan Yang, Si Li, Sheng Gao, and Jun Guo. 2022. Corefdpr: A joint model for coreference resolution and dropped pronoun recovery in chinese conversations. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:571–581.

Yue Yu and Qiurong Zhao. 2022. Expletive ta in mandarin chinese: A quantitative study. *Complex.*, 2022:2156978:1–2156978:8.

Hui Zhang, Tian Yuan, Junkun Chen, Xintong Li, Renjie Zheng, Yuxin Huang, Xiaojie Chen, Enlei Gong, Zeyu Chen, Xiaoguang Hu, dianhai yu, Yanjun Ma, and Liang Huang. 2022. Paddlespeech: An easy-to-use all-in-one speech toolkit. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*.

Ruiqing Zhang, Chao Pang, Chuanqiang Zhang, Shuohuan Wang, Zhongjun He, Yu Sun, Hua Wu, and Haifeng Wang. 2021. Correcting chinese spelling errors with phonetic pre-training. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2250–2261.

Yuanyuan Zhao, Jie Li, Xiaorui Wang, and Yan Li. 2019. The speechtransformer for large-scale mandarin chinese speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7095–7099.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *Proceedings of the 2021 International Conference on Machine Learning*, pages 12736–12746.

Shucheng Zhu and Pengyuan Liu. 2020. Great males and stubborn females: A diachronic study of corpus-based gendered skewness in chinese adjectives. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 31–42.