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

LREC-COLING 2024

OUTLINES

Motivation	“Ta” classification in Chinese Speech Recognition
Solution	“Ta Correct” scheme & TaR/TaL/TaN Model
Evaluation	Comparison with Baselines & Experimental Results
Conclusion	Summary of Findings & Next Steps
Q&A Session	

01

Motivation

Xunfei Input Method

Wechat Input Method

Tā hi Tā ʃi Tā ɪt

他 (he)，她 (she), 它 (it)

Cause

- English third-person pronouns (he/she/it) differ in pronunciation.
- Chinese third-person pronouns (他/她/它) share pronunciation.
- Chinese pronouns references are more ambiguous
- Void Reference "Ta" in Modern Chinese

Realted Work

- Chinese spelling correction (CSC). Most errors arise from homophones.
- Deep neural networks like RNN and LSTM for audio-to-text processing
- End-to-end structure with pre-trained BERT to differentiate phonetics and characters

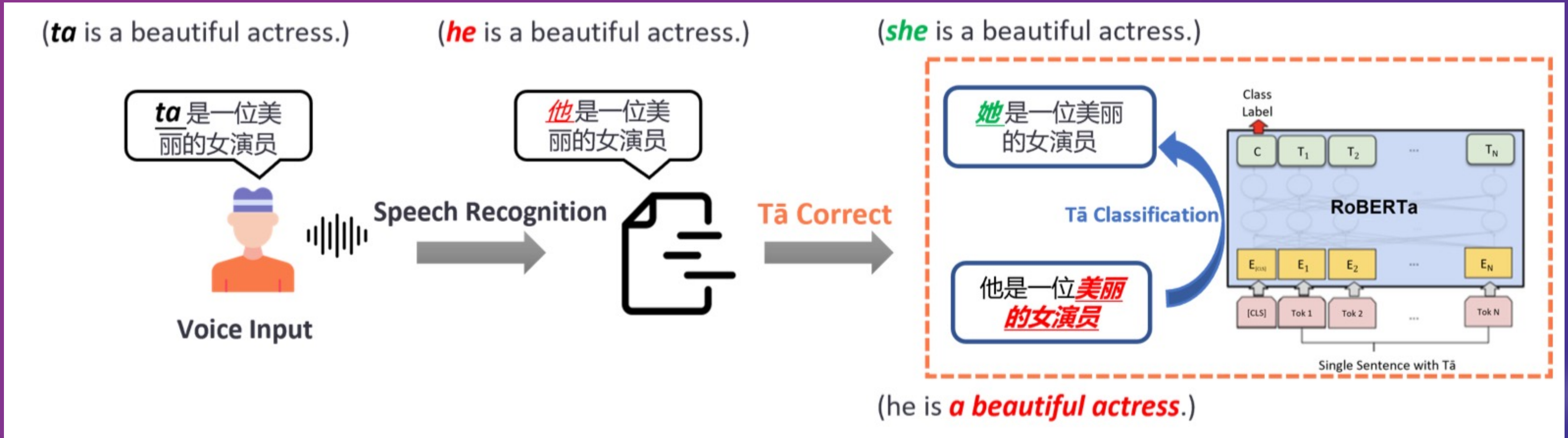
02

Solution

IDEAS

Inspired by Chinese spelling correction, propose the post-processing "Ta Correct" scheme

- <u>Language model</u>
- <u>LSTM model with phonetics and semantic features</u>
- <u>Rule-based assisted Ngram model</u>



IDEAS

high **accuracy** low

Language Model

- Based on Roberta MLM
- High precision
- High consumption

Phonetics and Semantic

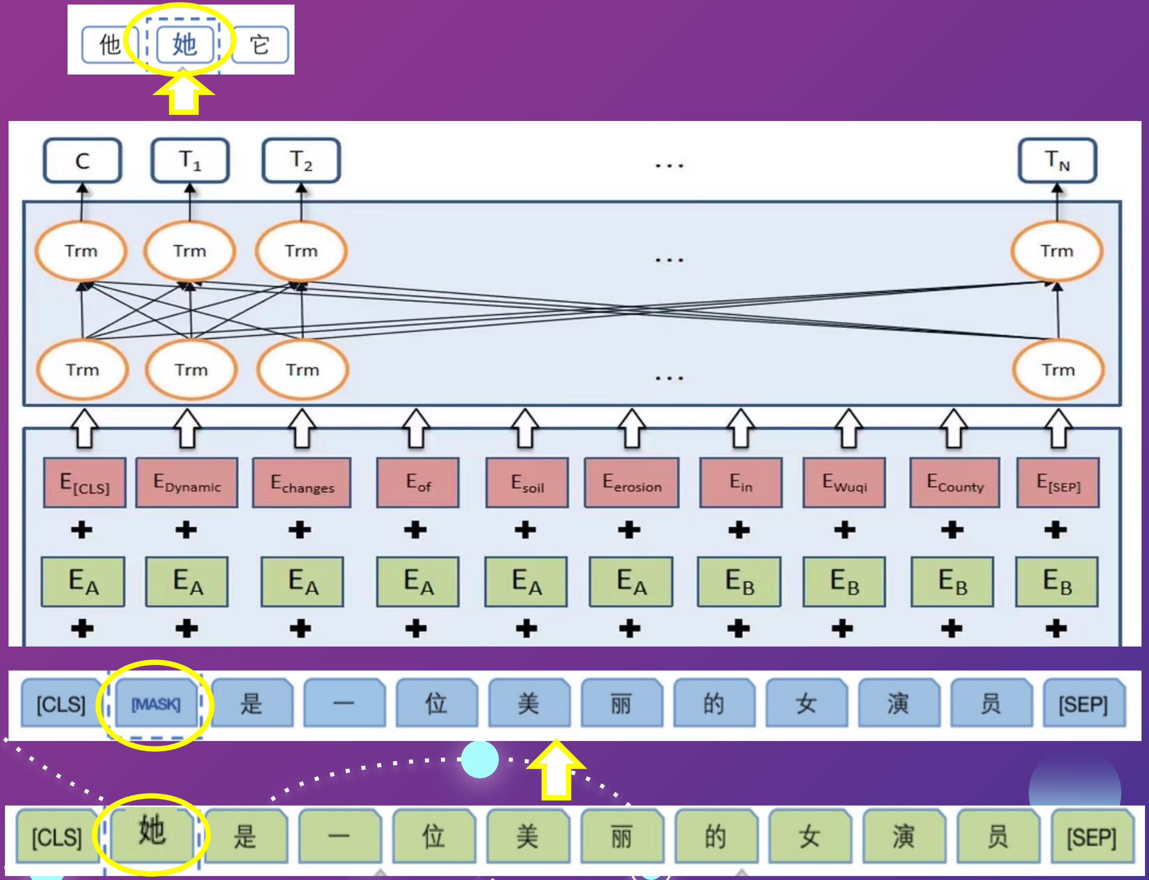
- Fuse LSTM and Attention
- Incorporate linguistic elements

Linguistic Rules

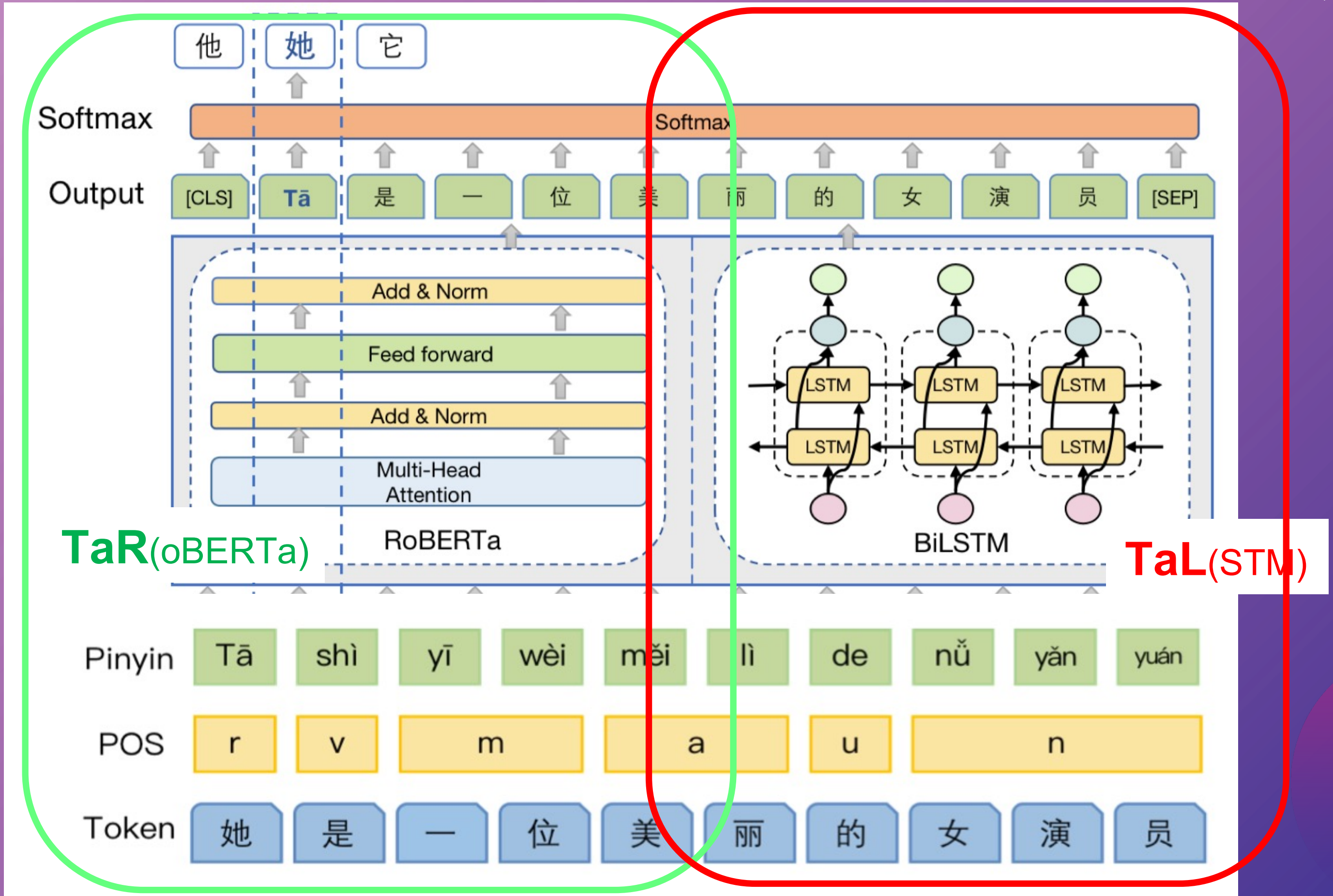
- Based on Ngram model
- Incorporate linguistic rules
- Low consumption.

high **consumption** low

METHODOLOGY



Language Model / LSTM Model



LSTM + Attention

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

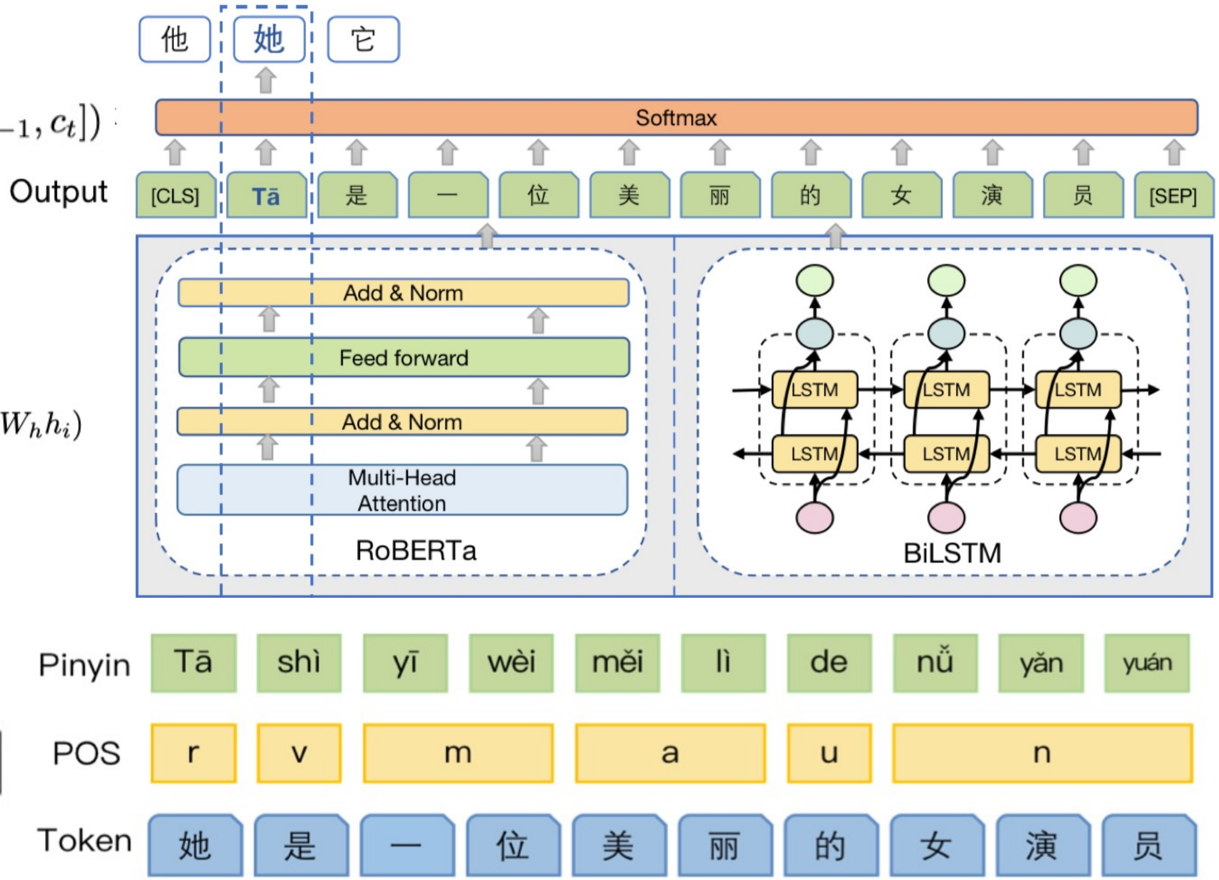
$$c_t = \sum_i a_{t,i} h_i$$

$$e_{t,i} = \textit{align}(s_{t-1}, h_i) = V^T \tanh(W_s s_{t-1} + W_h h_i)$$

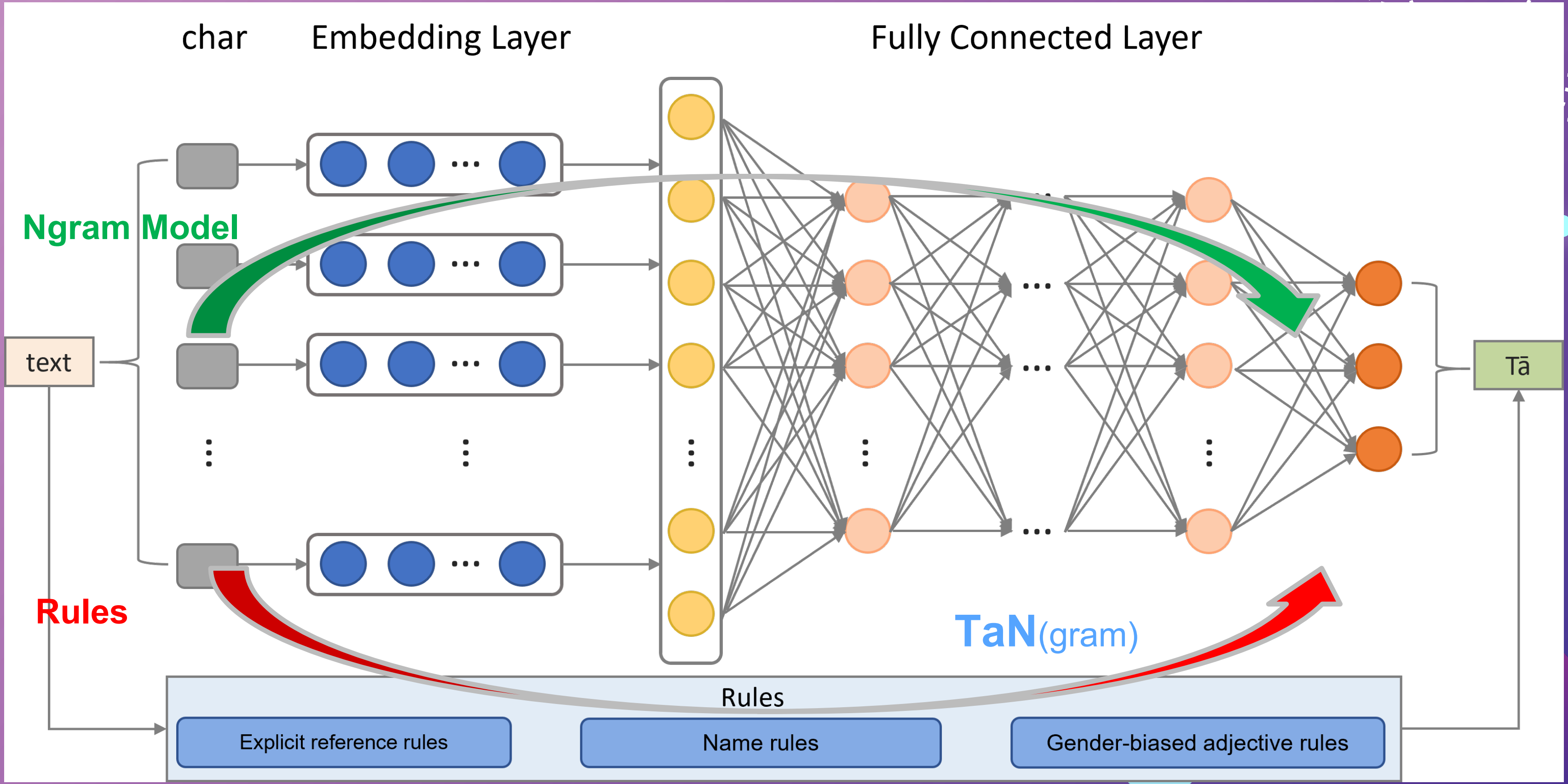
$$a_{t,i} = \textit{Softmax}(e_{t,i}) = \frac{\exp(e_{t,i})}{\sum_j \exp(e_{t,j})}$$

$$h_t = \textit{BiLSTM}(x_t)$$

$$x_t = [x_t^{text}; x_t^{pos}; x_t^{pinyin}]$$



Rule-based Assisted Ngram Model



Rules

Explicit reference rules: 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).

Name rules: commonly used characters in male and female names. For example, “强" (strong) 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. Word like "漂亮" (beautiful) often describes "她" (she), while "英俊" (handsome) commonly describes "他" (he).

03

Evaluation

1. Experimental Settings

Datasets

- **Weibo**: Sourced from the Sina Weibo, totaling over 360,000 sentences, of which 57,384 sentences contain "Ta".
- **Smp**: It consists of Weibo data covering various topics related to the COVID-19, with 5,000 sentences, of which 3,020 sentences contain "Ta".
- **Tieba** : Derived from forum posts, over 2,320,000 sentences, of which 946,969 sentences contain "Ta".

We generate speech data for these sentences using machine-generated pronunciation. We divide the train and test set in 7:3 ratio.

1. Experimental Settings

Baselines

- Open-source speech recognition models: including **PaddleSpeech** and **Whisper**
- Commercial input methods: including **Baidu Speech Recognition** and **Xunfei Automatic Speech Recognition**.

1. Experimental Settings

Evaluation Metrics

- **In-Sentence Accuracy (ISA):** The average of the accuracy of the predicted Ta in each sentence, reflecting how accurately the model identifies "Ta" on a sentence-by-sentence basis
- **Whole Sentence Accuracy (WSA):** Proportion of sentences in which the prediction Ta is all accurate, indicating the model's effectiveness at perfect prediction across entire sentences
- **“Ta” Conversion Accuracy (TCA):** Prediction accuracy of Ta for the entire dataset, assessing the model's comprehensive performance in correctly identifying "Ta."

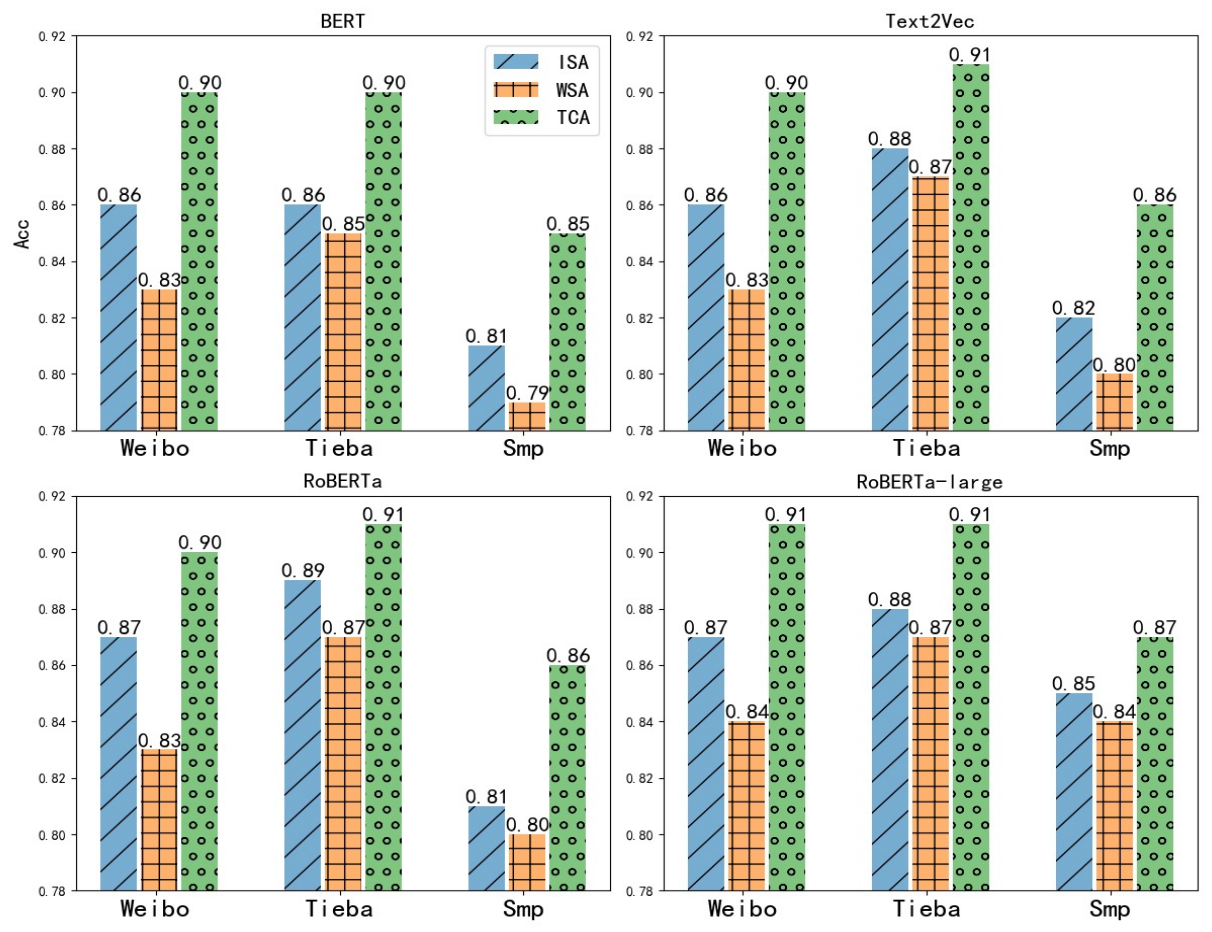
2. Experimental Results

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

3. Other Experiments

➢ Role of BERT Model

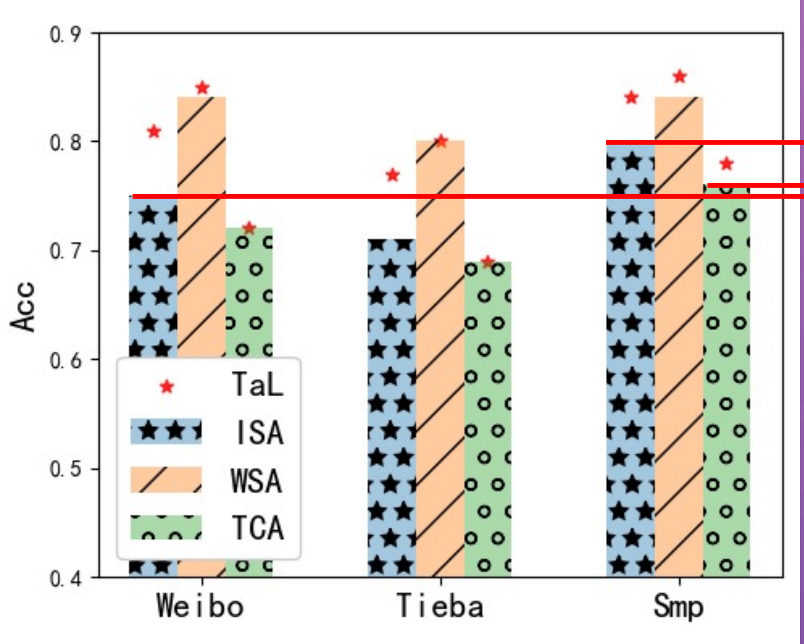
- BERT-base
- Text2vec
- RoBERTa
- RoBERTa-large √



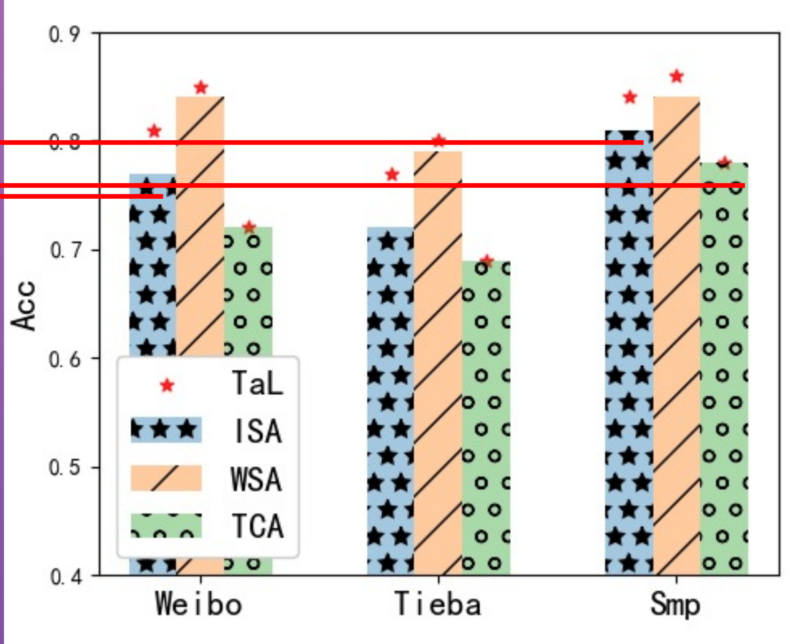
3. Other Experiments

➢ Role of Linguistic Features

Pinyin plays a more prominent role than POS. It loses more accuracy on both ISA and TCA metrics.



TaL (with only POS)

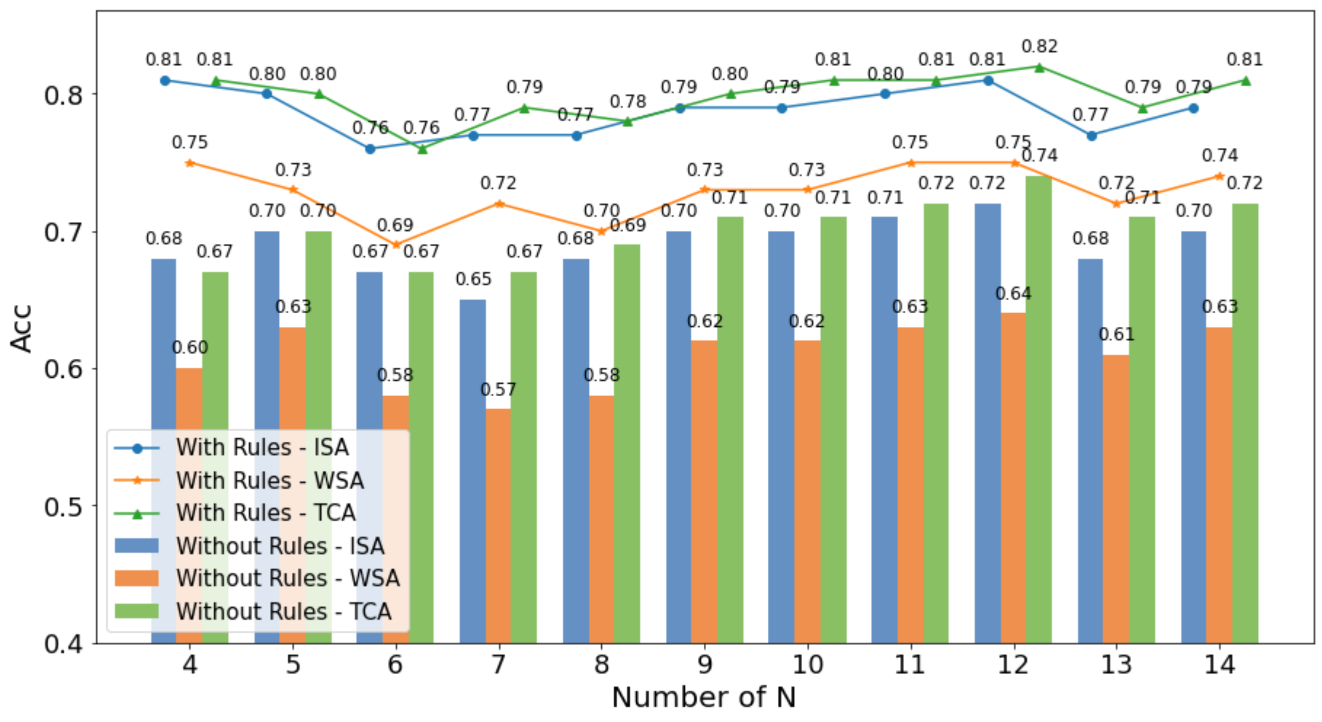


TaL (with only Pinyin)

3. Other Experiments

➢ Effect of the N and Rules

As N increases, there is a trend where the performance decreases first and then increases again with the best results observed around N=12.



4. Case Study

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 He is a beautiful actress	她是一位美丽的女演员 She is a beautiful actress
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 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 My sister loves juice best, so I saved a bottle for her
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 I raised a small white rabbit, and he likes to eat radish	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

04

Conclusion

Conclusion and Future Works

Contribution

- We've created three models, each with different costs, to handle different situations and better recognize the third-person pronoun "Ta" in speech-to-text applications.

Future Works

- We plan to explore additional rules and techniques for more pronouns to help advance Chinese spelling correction quickly. Our goal is to develop more adaptable models to enhance a wider array of applications and improve the user experience.

05

Q&A Session

THANKS!

DO YOU HAVE ANY QUESTIONS?

zhang_zb@hust.edu.cn

BDTS SCTS CGCL

LREC-COLING 2024