System Report for CCL25-Eval Task 12: Surpassing LLMs with a Simple Pipeline for Mandarin Spoken Entity-Relation Extraction

Wuganjing Song
The Hong Kong University of Science and Technology
wsongan@connect.ust.hk

Abstract

We present a strong and practical pipeline system for Mandarin spoken entity and relation extraction (Spoken-ERE), which integrates an industrial-grade ASR module (FireRedASR) with a span-based joint entity-relation extraction model. Unlike recent approaches that rely on large language models (LLMs) for end-to-end spoken information extraction, our method uses a modular pipeline design that is lightweight, interpretable, and easy to deploy. Despite its simplicity, **our system achieves top-tier performance in a recent shared task workshop, outperforming several 5× larger LLM-based systems for 20% on F1-score**. We demonstrate through experiments that with robust ASR and a well-designed span-based model, classical pipelines remain competitive and, in some scenarios, even preferable to LLM-based solutions for spoken information extraction in Mandarin.

Keywords: Spoken ERE , Mandarin , Pipeline System , ASR, Span-based Model , Small Models

1 Introduction

Spoken content understanding is a central challenge in building dialogue systems, voice agents, and domain-specific knowledge extraction tools. While entity and relation extraction (ERE) (Pawar et al., 2021) from text has been well studied, its extension to spoken data—particularly in Mandarin—remains underexplored due to challenges in ASR errors, disfluency, and lack of annotated speech datasets (Dighe et al., 2023).

Recent trends (Chu et al., 2023; Huang et al., 2024; Cui et al., 2024) advocate using large language models (LLMs) for end-to-end spoken information extraction. These methods rely on large-scale pretraining and instruction tuning, but face challenges in industrial deployment due to latency, cost, and interpretability. Moreover, LLMs are often less robust to noisy ASR outputs and domain-specific vocabulary.

In this paper, we revisit the classical pipeline architecture for spoken ERE: an upstream ASR module followed by a dedicated entity-relation extractor. Specifically, we use FireRedASR(Xu et al., 2025), an open-source, industrial-grade Mandarin ASR system, and pair it with a span-based joint ERE model inspired by(Eberts and Ulges, 2020).

While such a pipeline lacks joint optimization or speech-level modeling, we show that it outperforms many LLM-based systems in a recent workshop on spoken information extraction. This highlights that with a strong ASR and task-specific modeling, classical pipelines remain highly competitive and offer an efficient, robust alternative to LLMs.

- We design and evaluate a practical ASR-to-ERE pipeline for Mandarin spoken understanding.
- We show that our method surpasses several LLM-based baselines in a shared benchmark task.

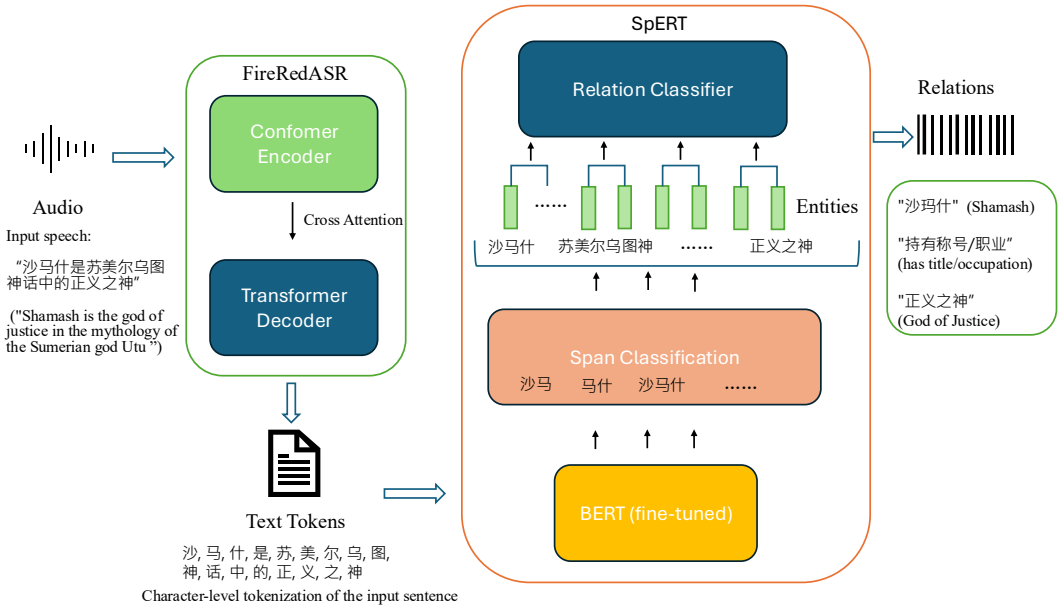

Figure 1: Framework of the Spoken-ERE

2 Method

2.1 Overview

Our pipeline consists of two main components:

1. **ASR Module** – Converts input speech into text.
2. **ERE Module** – Extracts entities and their relations from transcribed text.

This design offers modularity, interpretability, and ease of deployment.(Figure 1)

2.2 ASR Component: FireRedASR

We adopt the AED variant of FireRedASR (Xu et al., 2025), a Transformer-based encoder-decoder model trained on 70,000 hours of industrial-grade Mandarin speech. Its architecture combines a Conformer encoder with a Transformer decoder, enabling robust transcription in noisy conditions.

The encoder begins with convolutional subsampling and follows with Conformer blocks to capture both local and global speech patterns. The decoder generates character sequences with strong contextual awareness.

FireRedASR-AED achieves a 3.18% CER on public benchmarks, outperforming some larger LLM-based models, despite having only 1.1B parameters. This efficiency and robustness make it a reliable ASR front-end for our pipeline.

2.3 Entity and Relation Extraction: SpERT

For the ERE module, we implement a span-based joint extraction model inspired by Eberts(2020).

The model consists of:

- A BERT encoder for token representation.
- A span classifier that enumerates candidate spans for entities.
- A relation classifier that scores span pairs for potential relations.

Unlike sequence labeling methods(Nguyen and Verspoor, 2019; Li et al., 2019), the span-based architecture(Dixit and Al-Onaizan, 2019; Ji et al., 2022; Li et al., 2023) supports overlapping entities and

enables joint modeling of entity and relation extraction, which is well-suited for complex, relation-rich utterances in spoken dialogue.

While the original SpERT was designed for English corpora, we adapt it to Mandarin Chinese by replacing the encoder with **Chinese-Bert-wwm**(Cui et al., 2019), a pretrained Chinese BERT model trained with whole word masking. This substitution enables better tokenization and semantic representation for Chinese texts, especially in the context of spoken-style utterances which often contain informal structures and named entities.

We fine-tune the adapted SpERT model on a manually transcribed and annotated Mandarin speech dataset from CCL2025-Eval Task 12.[1] This adaptation preserves the strengths of span-based joint extraction while ensuring strong compatibility with Chinese linguistic features.

3 Experiments

3.1 Dataset

We trained and evaluate our pipeline system on the **CCL 2025 Evaluation Task 12 dataset**, which focuses on spoken entity and relation extraction from Mandarin speech. The dataset contains transcribed speech segments annotated with relational triples in the form of `(head, relation, tail)`, and covers a wide range of topics and entity types. The task presents challenges related to ASR errors, spoken-style disfluencies, and long-range dependencies.

3.2 Evaluation

We follow the official evaluation protocol defined by the shared task organizers. System outputs are compared against the annotated gold-standard triples on the basis of exact match. Three standard metrics are used:

- **Precision (P)**: The ratio of correctly extracted triples to the total number of triples predicted by the system.

$$\mathrm{P} = \frac{\text{number of correct triples}}{\text{number of predicted triples}} \tag{1}$$

- **Recall (R)** The ratio of correctly extracted triples to the total number of gold triples in the dataset.

$$\mathrm{R} = \frac{\text{number of correct triples}}{\text{number of gold triples}} \tag{2}$$

- **F1 Score**: The harmonic mean of precision and recall, reflecting the overall effectiveness of the system.

$$\mathrm{F1} = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{3}$$

3.3 Result

We report our system's performance in terms of F1 score, on the CCL2025 Task 12 evaluation set. Table 1 summarizes the results of our system in comparison to a strong LLM-based baseline.

System	Param. Size	F1 Score
Qwen2-Audio	7B	0.46
Ours	**1.2B**	**0.56**

Table 1: F1-score comparison on CCL2025-Eval Task 12.

Despite being over 5× smaller in parameters, our pipeline achieves a significantly higher F1 score than the Qwen2-Audio 7B model. This highlights the effectiveness and efficiency of a carefully designed pipeline, especially when high-quality ASR and lightweight span-based extraction are used in tandem.

[1] https://github.com/DMU-ITREC/CSRTE-CCL2025

4 Conclusion & Future Work

We presented a practical pipeline for Mandarin spoken entity and relation extraction, combining a robust ASR system with a span-based joint model. Despite its simplicity and modularity, our approach outperforms several LLM-based baselines in a recent shared benchmark. This demonstrates that well-designed lightweight systems remain competitive for spoken information extraction, especially in industrial or resource-constrained settings.

While our findings highlight the strength of non-LLM methods, we also acknowledge that multimodal speech-language models may surpass current pipelines under more optimal configurations. We hope our results inspire further exploration of both lightweight and hybrid approaches in this evolving field.

Future work includes:

- **Tight ASR-ERE Integration**: Investigating tighter coupling between ASR and the extraction model, such as through shared encoder representations or confidence-aware training that conditions ERE decisions on ASR uncertainty signals.
- **Acoustic-Textual Fusion in Lightweight Models**: Use shallow fusion layers to combine ASR audio features and text representations, improving extraction accuracy without full speech-LMs.
- **Deployment-Oriented Optimization**: Exploring quantization, pruning, and latency-aware design for deploying real-time spoken ERE systems on edge devices or in low-resource environments.

References

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. 2024. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*.

Pranay Dighe, Yi (Siri) Su, Daniel Zheng, Yunshu Liu, Vineet Garg, Xiaochuan Niu, and Ahmed Tewfik. 2023. Leveraging large language models for exploiting asr uncertainty. In *ICASSP*.

Kalpit Dixit and Yaser Al-Onaizan. 2019. Span-level model for relation extraction.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Bin Ji, Hao Xu, Jie Yu, Shasha Li, Jun Ma, Yuke Ji, and Huijun Liu. 2022. A two-phase paradigm for joint entity-relation extraction. *arXiv preprint arXiv:2208.08659*.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. *arXiv preprint arXiv:1905.05529*.

Qibin Li, Nianmin Yao, Nai Zhou, Jian Zhao, and Yanan Zhang. 2023. A joint entity and relation extraction model based on efficient sampling and explicit interaction. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–18.

Dat Quoc Nguyen and Karin Verspoor. 2019. End-to-end neural relation extraction using deep biaffine attention. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 729–738. Springer.

Sachin Pawar, Pushpak Bhattacharyya, and Girish K. Palshikar. 2021. Techniques for jointly extracting entities and relations: A survey.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *arXiv preprint arXiv:2501.14350*.

