國立臺灣師範大學
NATIONAL TAIWAN NORMAL UNIVERSITY

💃 DANCER: Entity **D**escription **A**ugmented **N**amed Entity **C**orr**E**cto**R** for Automatic Speech Recognition

Yi-Cheng Wang, Hsin-Wei Wang, Bi-Cheng Yan, Chi-Han Lin, Berlin Chen

National Taiwan Normal University

SMIL

Quick Summary

- In this research,
  - Focus on named entity correction (NEC) for automatic speech recognition (ASR)
  - Present a novel NEC framework, dubbed DANCER
  - DANCER leverages an efficient entity description augmented masked language model (EDA-MLM) to alleviate the phonetic confusion for NEC on ASR transcription
  - Our EDA-MLM comprised a dense retrieval model, enabling MLM to adapt swiftly to domain-specific entities for the NEC task
  - Experiments are conducted on a public benchmark dataset (AISHELL-1) evaluated the feasibility of our proposed method

LREC-COLING 2024

National Taiwan Normal University
Speech and Machine Intelligence Laboratory

Introduction

- Background

ASR Mistranscription

Play bad romance by lady gaga

User

ASR System

ASR Transcription

Introduction

- Background

ASR Mistranscription

ASR Error Correction on domain-specific phrase

1 → **ASR with an external language models**

Shallow Fusion, WFST, On-the-fly Rescoring

2 → **Inject domain-specific contexts into ASR**

Deep Contextual Biasing

3 → **Post-correction on ASR transcripts** ✓

Named Entity Correction

LREC-COLING 2024

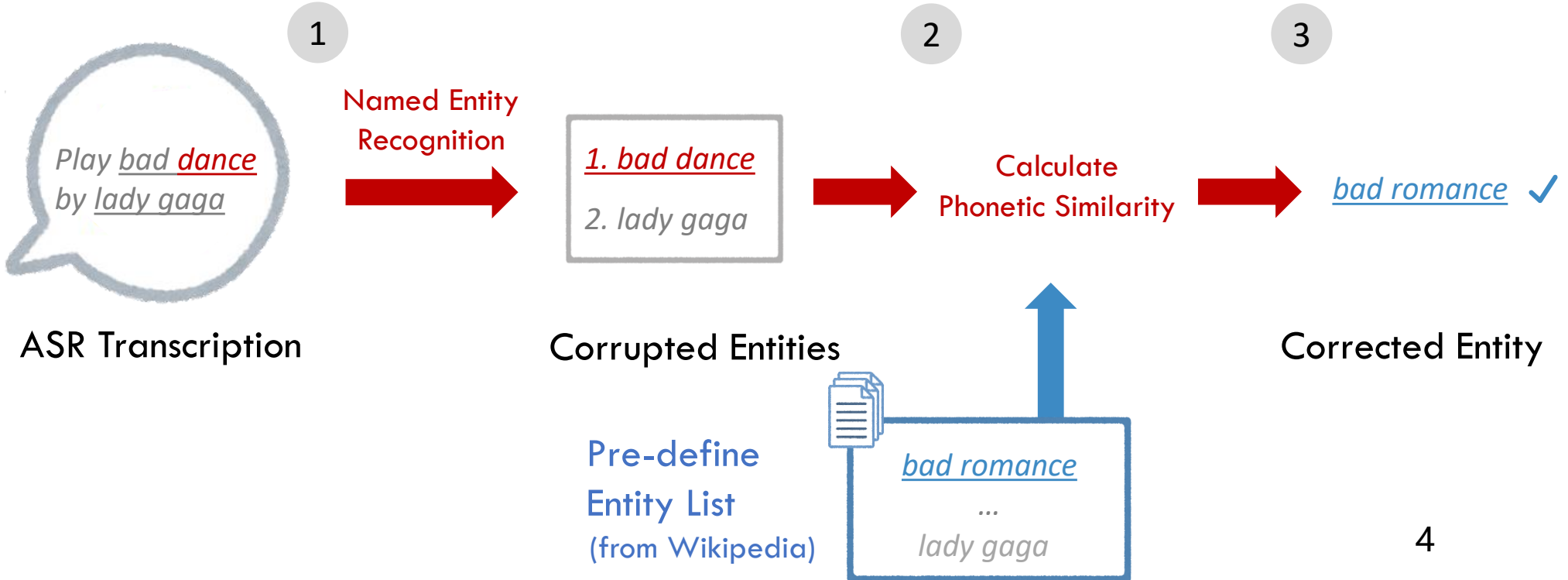
National Taiwan Normal University
Speech and Machine Intelligence Laboratory

Introduction

- Named Entity Correction (NEC)

Prevailing Approaches Are Phonetic Edit-distance-based (PED) Methods

- ASR systems typically mistranscribe entities to acoustically similar words
- A simple phonetic matching mechanism is effective enough to correct the corrupted entity

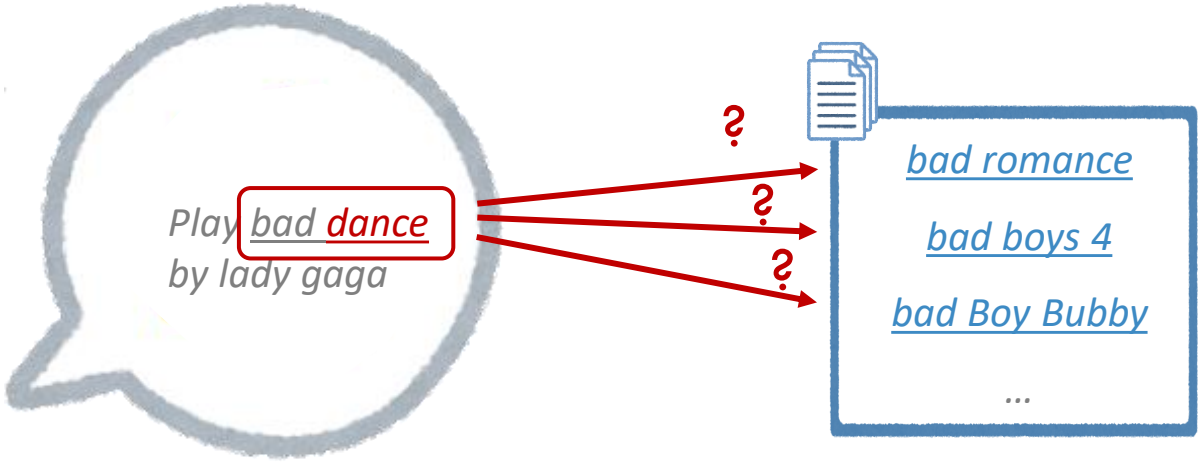


Introduction

- Named Entity Correction (NEC)

  Prevailing Approaches Are Phonetic Edit-distance-based (PED) Methods

  - However, as the named entity (NE) list grows, the problems of phonetic confusion in the NE list are exacerbated
  - For example, homophone ambiguities increase substantially



ASR Transcription

Pre-define Entity List
(from Wikipedia)

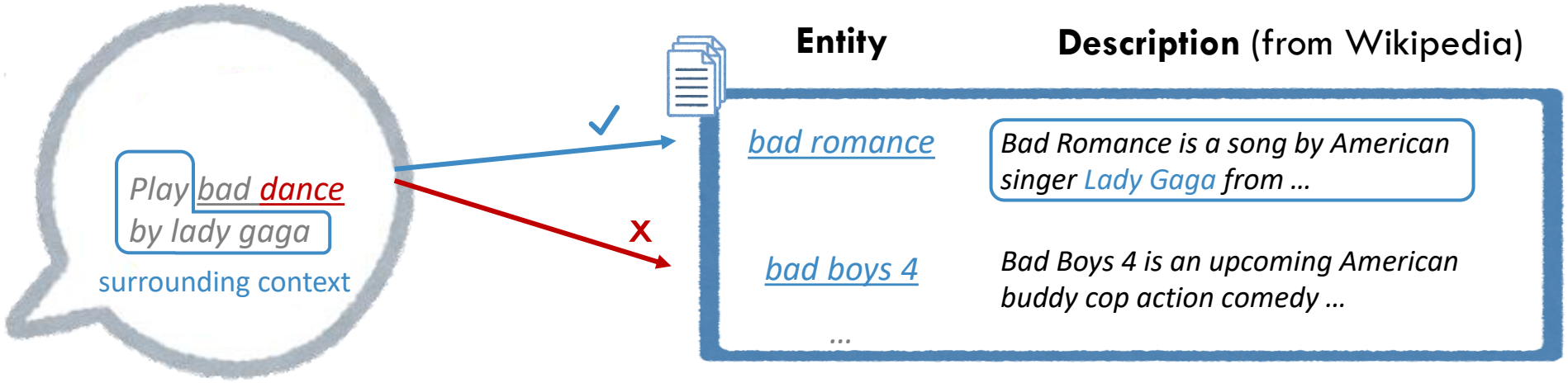
Introduction

- Named Entity Correction (NEC)

Goal (concepts of the proposed model)

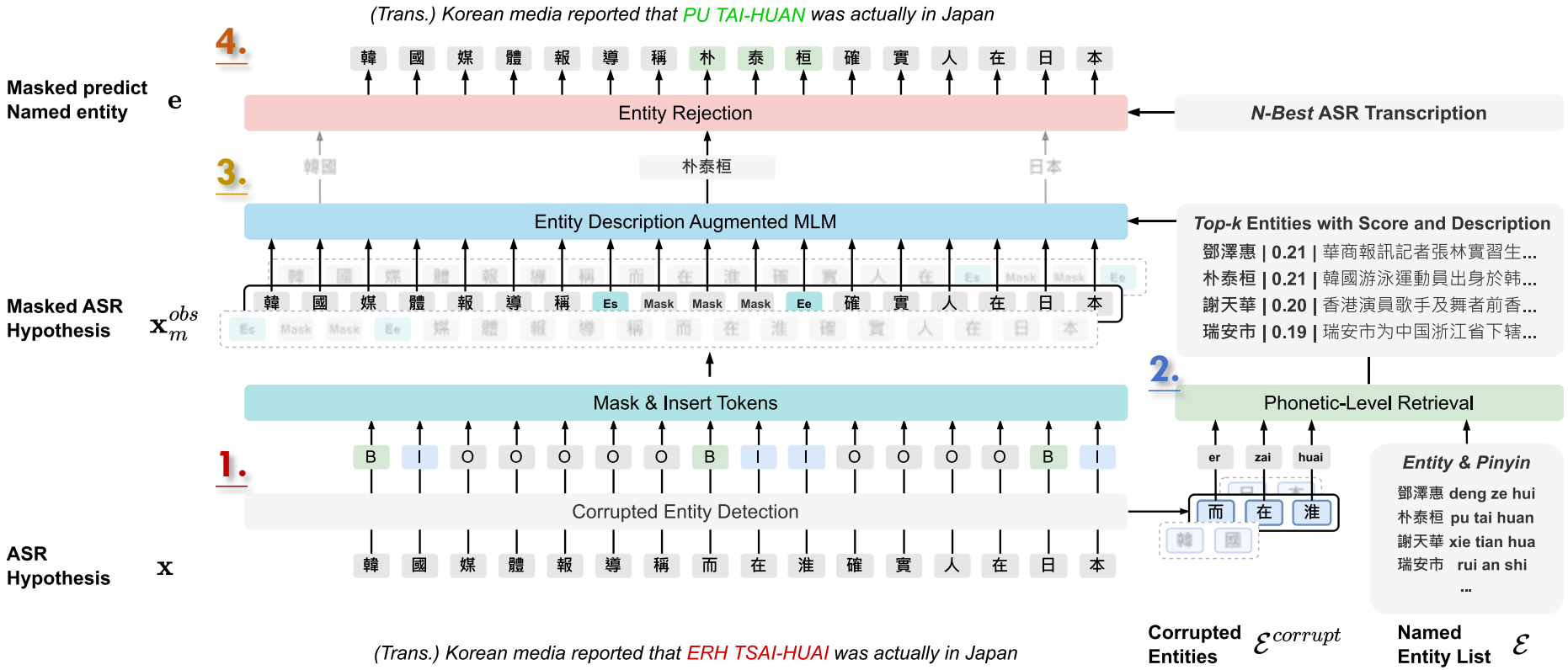
In this work, we propose a novel NEC framework, dubbed DANCER

- In additional to only use the phonetic information like PED-NEC methods does:
  - ➢ We judge the semantic relation between ASR transcription and the entity descriptions



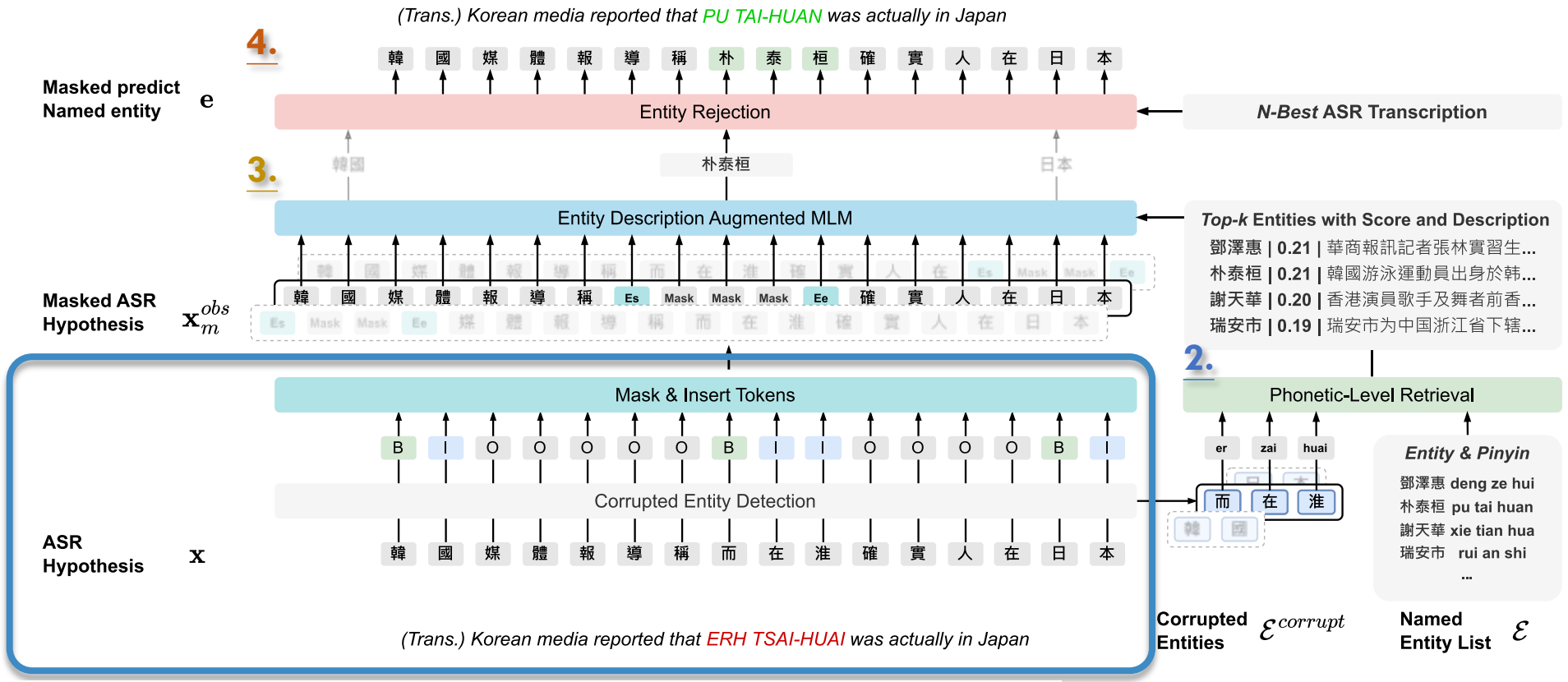
DANCER

Overview of DANCER



DANCER

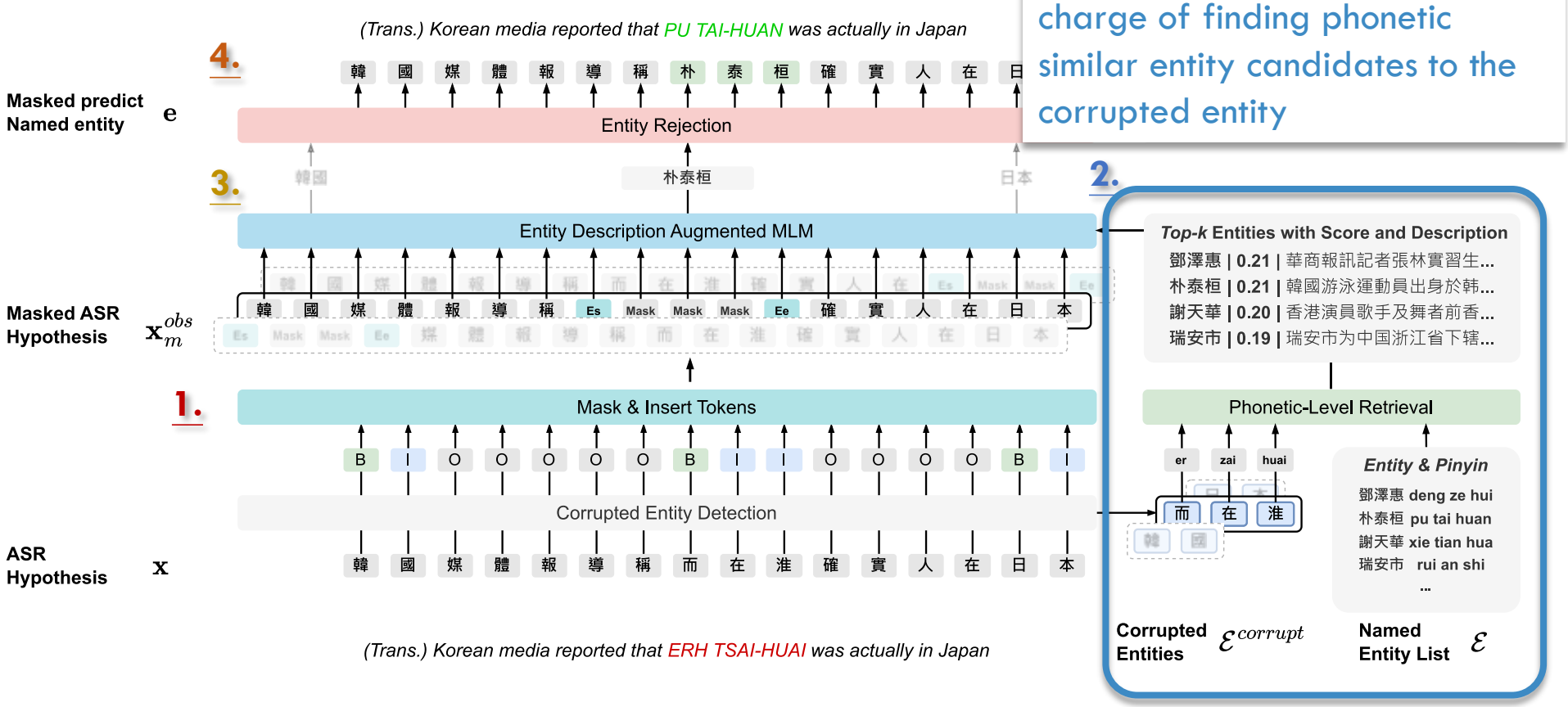
Overview of DANCER



1. Corrupted Entity Detection is in charge of detecting mistranscribe entities

DANCER

Overview of DANCER



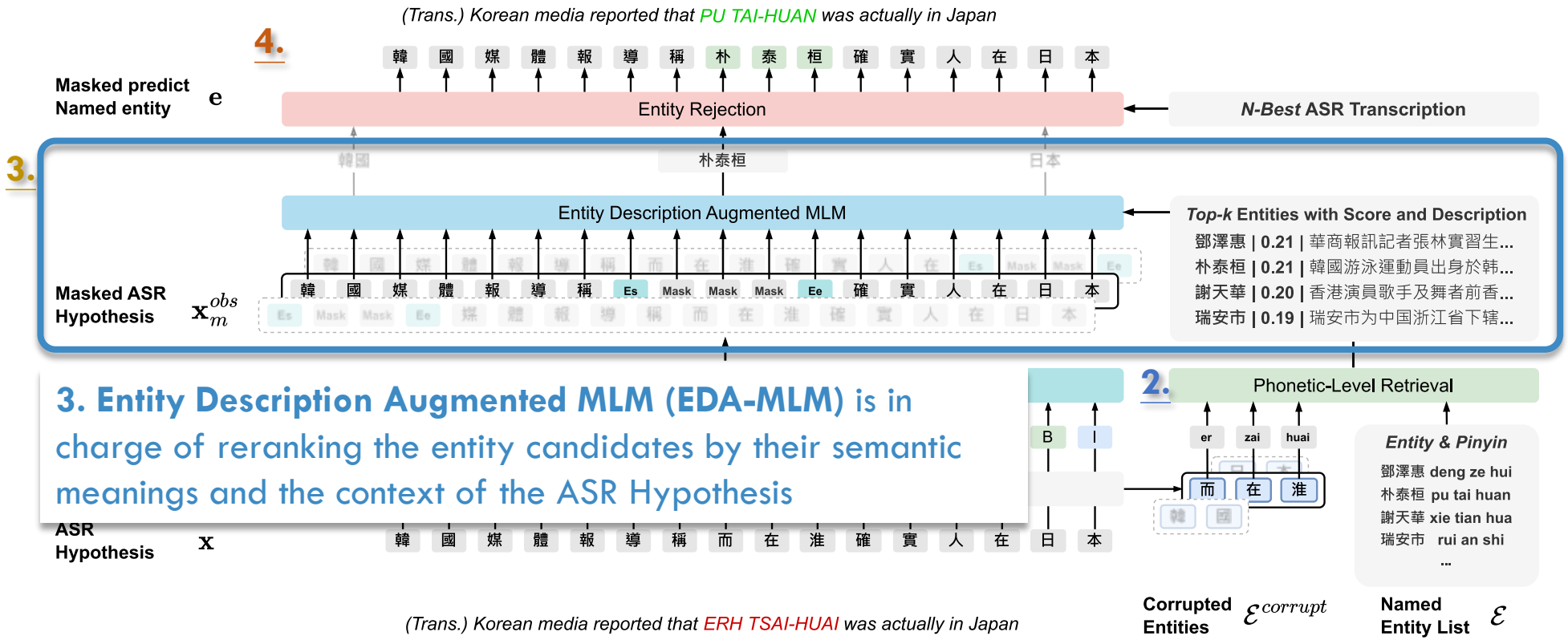
2. Phonetic-level Retriever is in charge of finding phonetic similar entity candidates to the corrupted entity

LREC-COLING 2024

National Taiwan Normal University
Speech and Machine Intelligence Laboratory

DANCER

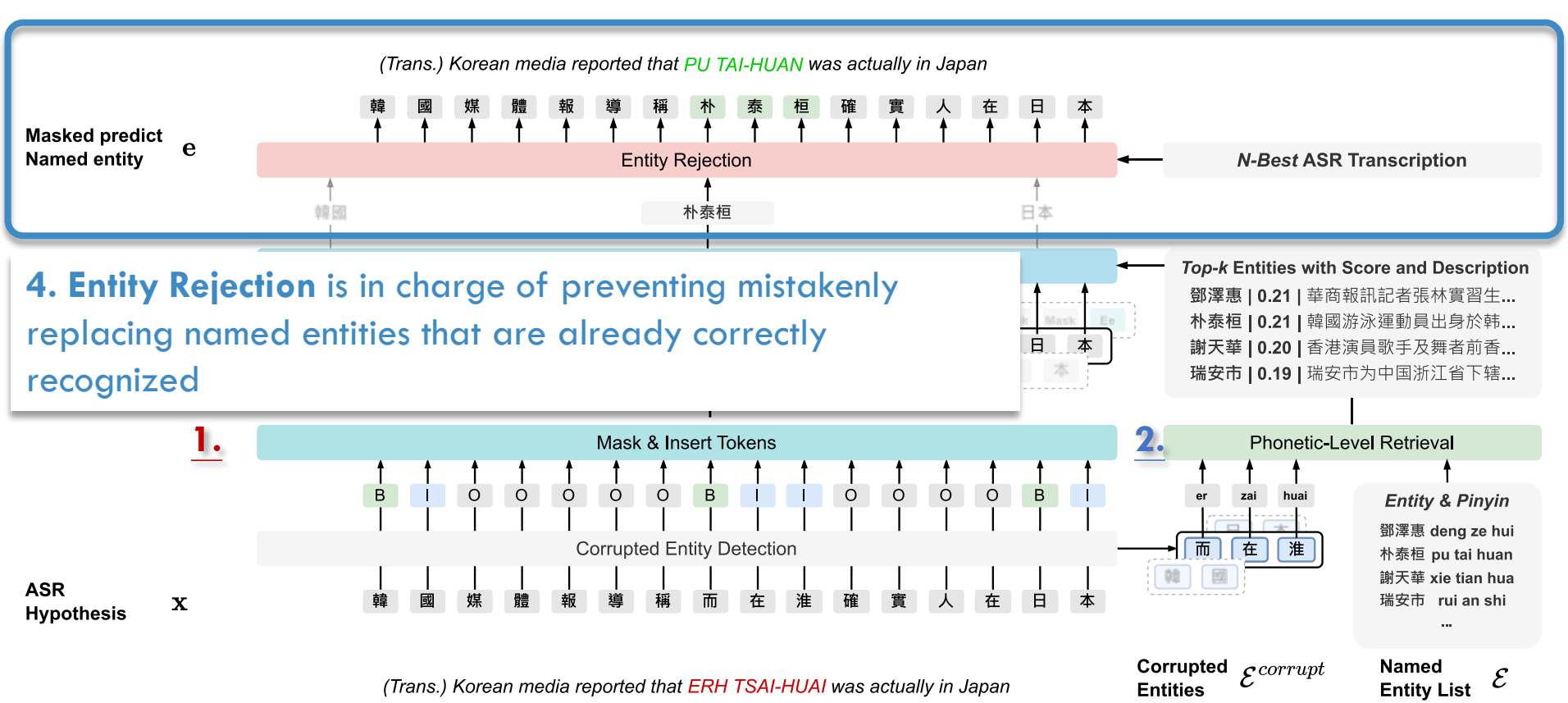
Overview of DANCER



3. Entity Description Augmented MLM (EDA-MLM) is in charge of reranking the entity candidates by their semantic meanings and the context of the ASR Hypothesis

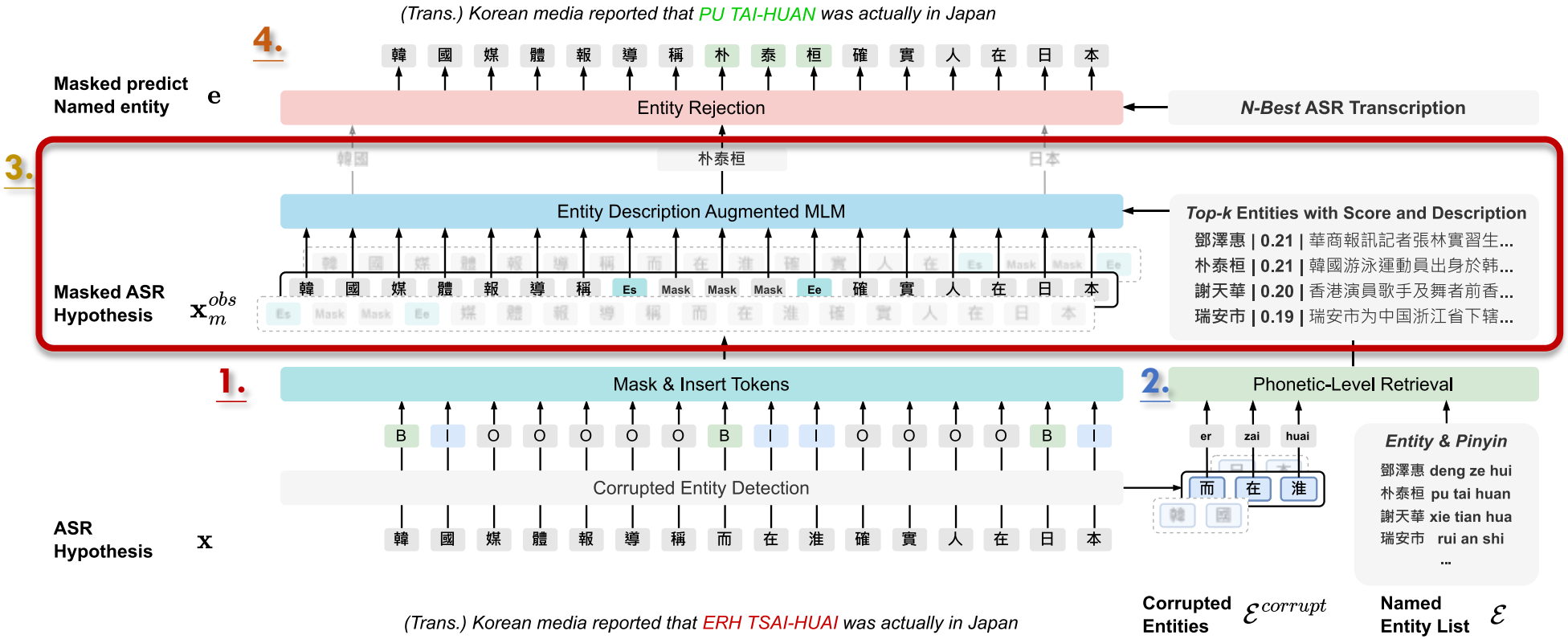
DANCER

Overview of DANCER



DANCER

Overview of DANCER

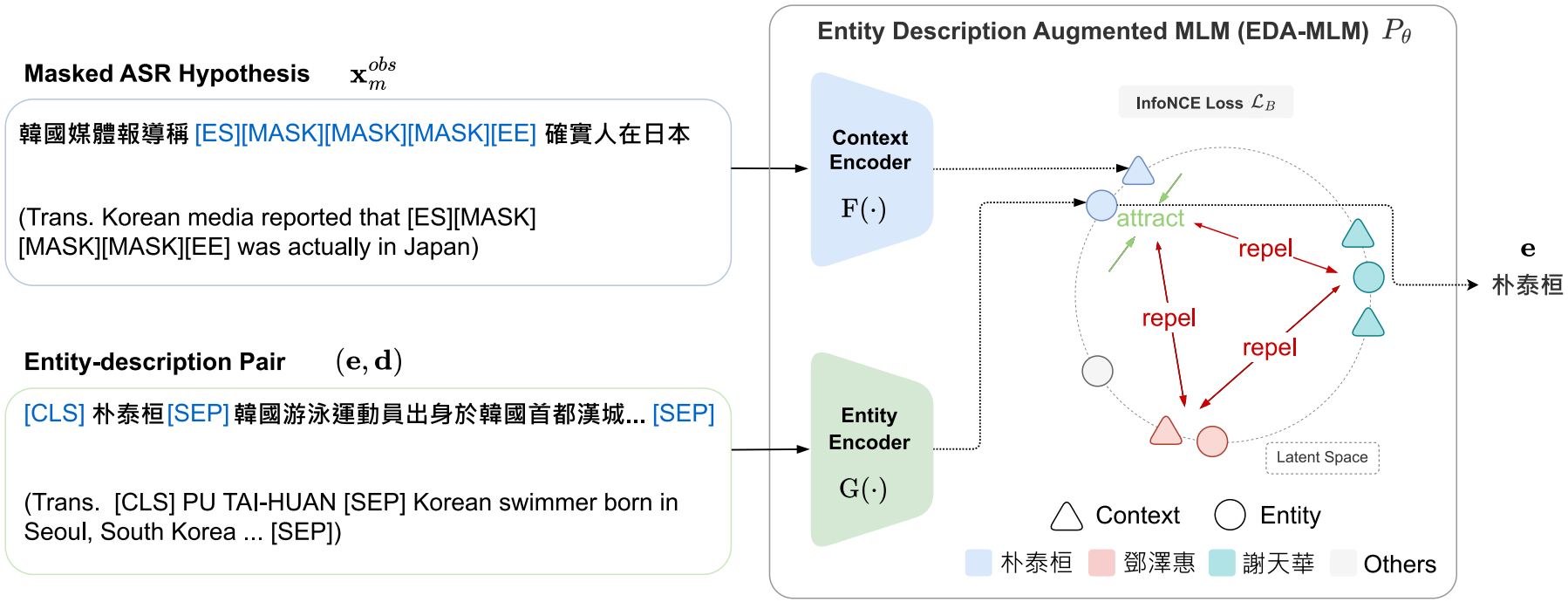


DANCER

- Masked Language Model (MLM)
  - The non-autoregressive MLM is renowned for its efficiency and capacity to derive rich contextual representations for an anchor mask
  - However, one drawback that MLM faced with is its insufficient capacity of adapting and accommodating to unseen phrases

DANCER

- Entity Description Augmented Masked Language Model (EDA-MLM)
  - Inspired by dense retrieval methods, we employ a dual-encoder architecture to aid MLM in retrieving external knowledge
  - EDA-MLM is trained by using Contrastive Learning



Experiments

Dataset – AISHELL-1 & Homophone

- **AISHELL-1**
  - A commonly used open-source speech corpus for evaluating Chinese ASR systems
  - Containing over 170 hours of Mandarin speech data
  - Involved diverse domains, such as:
    - Finance, Science and Technology, Sports, Entertainment, and News
- **Homophone**
  - Homophone test set contains 115 highly phonetically confusing speech utterances, which are curated from the test set of AISHELL-1

Experiments

- Evaluation Metrics:
  - Character Error Rate (**CER**) ↓
  - Named Entity Character Error Rate (**NE-CER**) ↓
  - Non-Named Entity Character Error Rate (**NNE-CER**) ↓
  - Named Entity Recall Rate (**NE-Recall**) ↑
- Baseline methods:
  - Conformer ASR model
  - Phonetic Edit-distance-based Named Entity Corrector (**PED-NEC**)

Experiments

- Entity Description Construction:
  - **Entity List:**
    - We utilized the publicly available AISHELL-NER dataset to obtain the tagging information of named entities in AISHELL-1
  - **Entity Descriptions**
    - We utilized Chinese Wikipedia as our source of knowledge
    - We used a given entity as the query to search for the most relevant article from Wiki
    - We applied a text normalization to eliminate semi-structured data from the acquired article
    - The entity description was then formed by extracting the first 100 words from the article

Experiments

Main Results

- *Main results of our DANCER model on the AISHELL-1 and Homophone test set*

Model	AISHELL Test Set (%)				Homophone Test Set (%)			
	CER	NNE CER	NE CER	NE Recall	CER	NNE CER	NE CER	NE Recall
Conformer	4.62	**4.00**	11.12	78.36	8.41	**5.27**	15.58	70.25
PED-NEC	4.34	4.01	8.14	84.63	10.08	5.35	20.88	56.72
- w/o rejection	4.90	4.65	8.22	85.61	10.67	6.05	21.42	56.14
DANCER	**4.29**	4.00	**7.57**	85.85	**7.17**	5.30	**11.33**	**79.84**
- w/o rejection	4.84	4.64	7.63	**86.81**	7.87	6.12	12.04	78.68

➢ Main Results:

- Our approach, utilizing entity semantics, leads to better CER reduction for both datasets
- Our DANCER model achieved an additional CER reduction of 28.87% relatively over the PED-NEC model

➢ Impact of Incorporating Entity Rejection Mechanism:

- Slight decrease in NE recall rate
- However, this cautious process significantly reduces CER related to non-entity parts of test utterances

Experiments

- Few-shot Generalization
  - *Analysis of few-shot generalization ability on the AISHELL-1 test set*

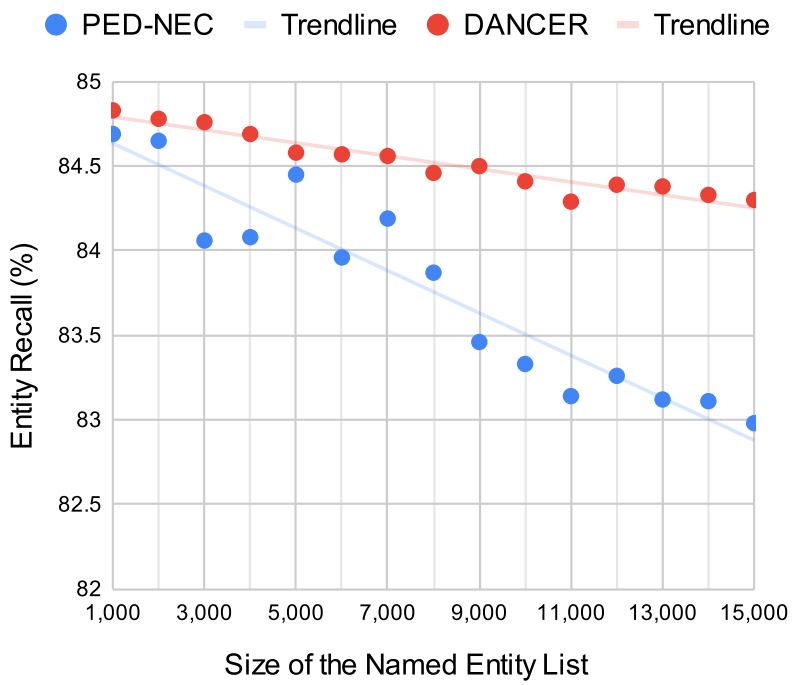
Model	Test Set NE-Recall (%)		
	≤ 0-shot	*≤ 5-shot*	*≤ 100-shot*
Conformer	38.83	50.32	70.55
PED-NEC	61.73	66.92	79.82
DANCER	**61.77**	**68.39**	**80.86**

- DANCER consistently outperforms PED-NEC across all shot categories
- Zero-Shot Ability:
  - EDA-MLM can adapts to unseen entities and demonstrates promising zero-shot ability

Experiments

Impact of the Entity List Size

- *Impact of the named entity list size on AISHELL-1 test set*



- Our proposed DANCER model can effectively leverage entity semantics to alleviate this problem
- There is a sizable performance gap between DANCER and the phonetic edit-distance-based NEC method (PED-NEC) as the entity sizes increase

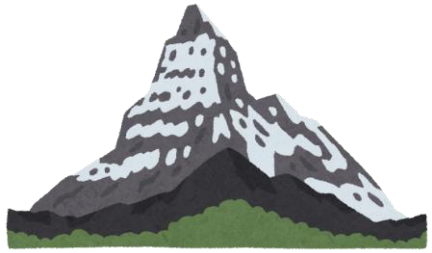
Conclusion and Future Work

- In this research,
  - Proposed a novel method (i.e., DANCER) for NEC
  - We leverages entity descriptions to provide additional information that helps mitigate the problems of phonetic confusion incurred by ASR
- As to future work,
  - We plan to explore alternative entity modeling regimes, such as graph-based modeling
  - Incorporate the NE list ahead of time into the corrupted entity detector to reduce the search space

LREC-COLING 2024

National Taiwan Normal University
Speech and Machine Intelligence Laboratory

Thank you for listening!



LREC-COLING 2024

National Taiwan Normal University
Speech and Machine Intelligence Laboratory