Contextual Biasing of Named-Entities with Large Language Models

Chuanneng Sun[1], Zeeshan Ahmed[2], Yingyi Ma[2],
Zhe Liu[2], Lucas Kabela[2], Yutong Pang[2], and Ozlem Kalinli[2]

[1] Rutgers University—New Brunswick, Piscataway, NJ, USA
[2] Meta AI, Menlo Park, CA, USA

Introduction & Motivation

Language Models (LMs) in Automatic Speech Recognition (ASR) systems suffer from a drastic reduction in quality when recognizing uncommon words, e.g., Named Entities (NE) that appear infrequently in training data. Generally, these uncommon words can be inferred from the context in which the model is being used.

There have been many works studying the contextual biasing problem. Traditional solutions like n-gram based class LM [1], dynamic weighted finite state transducer (WFST) [2], and Neural Network LMs (NNLMs) [3] have been proposed but the Word Error Rate (WER) is still to be improved. Other methods like Personalized LM (PLM) [4] or Multi-Head Attention-based approaches [5] have also been proposed, but the ability of neural networks have not been fully exploited.

Having witnessed the success of Large Language Models (LLMs), in this work, we propose to leverage the ability of LLMs for contextual biasing to further improve the WER.

Contributions

- We propose contextual biasing prompts when calculating the second-pass score. We propose a simple yet effective format of prompts to incorporate the lists of biasing entities. Moreover, we propose a few-shot learning method by providing examples in the prompts.
- We introduce a multi-task training framework. This involves augmenting the Large Language Model (LLM) with a dedicated tag head that predicts the entity class tag (e.g., person, location) of the next tokens. The entire model is jointly trained with two distinct losses: the entity tag prediction loss and the token prediction loss.
- We propose dynamic prompting to improve attention efficiency and to reduce the input sequence length. For each token, we first obtain the entity class prediction from the class head (top 1). Then, we subset the contextual information in the prompt based on the class for token prediction.

Our Approach

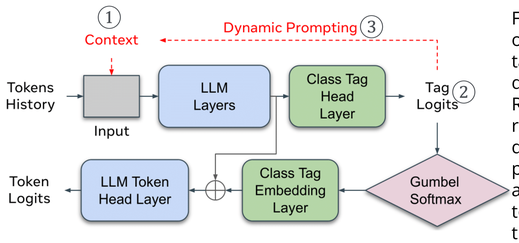
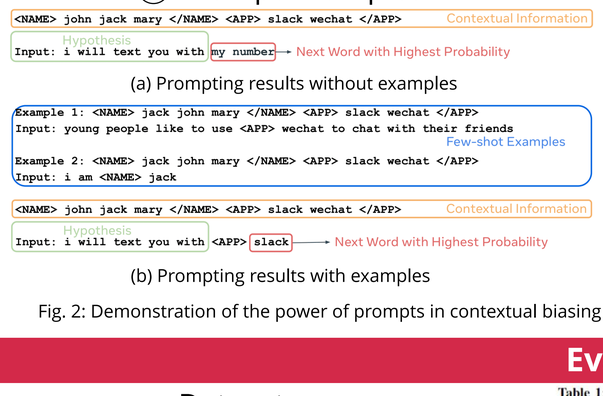


Fig. 1: The flow chart of the proposed multi-task model with dynamic prompting. Red dashed arrows represent the dynamic prompting process. Black solid arrows represent token prediction given the selected prompts.

① Prompt Examples



(a) Prompting results without examples

(b) Prompting results with examples

Fig. 2: Demonstration of the power of prompts in contextual biasing

② Multi-Task Training

We introduce an enhancement by transforming the model into a multi-task architecture. Alongside the primary task of next-token prediction, we incorporate a secondary task: next-token class prediction. LoRA is applied to fine-tune the LLM. The loss is:

$$L = \alpha L_{CE}(f_{token}(x), y_{token}) + (1-\alpha)L_{CE}(f_{class}, (x)\ y_{class})$$

③ Dynamic Prompting

We propose dynamic prompting, where, instead of having every entity class in the biasing list, we only provide the biasing list corresponding to the next token's entity class. With the help of multi-task training and the class tag head, the log-likelihood for the sentence becomes a conditional log-likelihood, which can be written as,

$$log\ P(w_T, \ldots, w_0|C) = \sum_{t=1}^{T} log\ P(w_t|h_w, c(h_w))$$

Evaluation Results

Dataset:

- An in-house dataset of calling messaging and dictation (CMD). An in-house audio model, with 13 transformer layers as encoder and 1 LSTM model as decoder is used.

- Reformatted SLUE-VoxPopuli dataset by treating ground truth entities as context information. Wave2vec2 is used as the audio model.

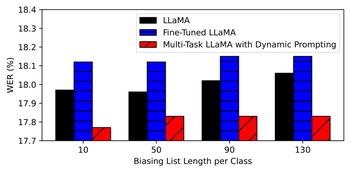


Fig. 3: Model performance when the length of the biasing list varies. We have three classes of entities in total.

Table 1: Second-pass WER evaluation results on different variations of un-fine-tuned LLMs.

Model	CMD	SLUE
First Pass Baseline	7.10	20.03
Oracle	4.37 (38.5%)	16.84 (16.0%)
LLaMA	7.07 (0.4%)	18.35 (8.4%)
+Biasing in Inference	5.92 (16.6%)	18.11 (9.6%)
+Few-shot Examples	**5.84 (17.8%)**	**18.10 (9.6%)**
RoBERTa	7.10 (0%)	19.60 (2.1%)
+ Biasing in Inference	7.10 (0%)	19.73 (1.5%)

Table 2: Second-pass WER evaluation results on different variations of fine-tuned LLMs.

Model	CMD	SLUE
Fine-tuned LLaMA	7.03 (1.0%)	18.00 (10.1%)
+Biasing in Inference	5.93 (16.5%)	17.98 (10.2%)
+Biasing in Training	5.75 (19.0%)	18.02 (10.0%)
Fine-tuned RoBERTa	7.06 (0.6%)	19.11 (4.6%)
+Biasing in Inference	7.07 (0.4%)	19.15 (4.4%)
+Basing in Training	7.06 (0.6%)	19.16 (4.3%)
Multi-Task LLaMA	6.99 (1.6%)	17.95 (10.4%)
+Biasing in Both	5.71 (19.6%)	17.85 (10.9%)
+Dynamic Prompting	**5.68 (20.0%)**	**17.77 (11.3%)**

Table 3: Ablation study on the impact of the appearance of ground truth entities in the biasing list in terms of WER.

Model	CMD-GT	CMD-NGT
LLaMA	-	-
+Biasing in Inference	5.92 (16.6%)	6.91 (2.7%)
Fine-tuned LLaMA	-	-
+Biasing in Inference	5.93 (16.5%)	6.95 (2.1%)
+Biasing in Training	5.75 (19.0%)	7.02 (1.1%)
+Dynamic Prompting	5.71 (19.6%)	7.02 (1.1%)

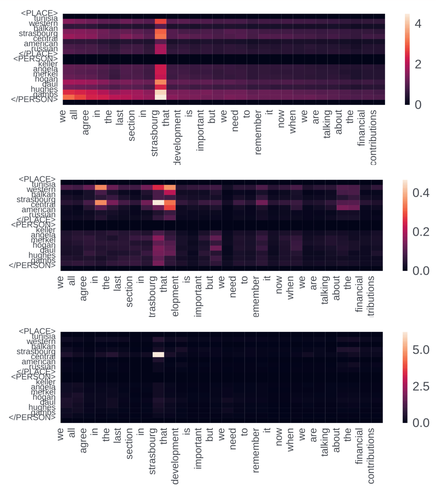


Fig. 4: The attention weights for different layers in the LLaMA model we fine-tuned. "Strasbourg" is in the input sentence, and the attention layers are able to attend to it in the biasing list.

References

[1] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer, "Class-based n-gram models of natural language," Computational linguistics, vol. 18, no. 4, pp. 467–480, 1992
[2] Antoine Bruguier, Duc Le, Rohit Prabhavalkar, Dangna Li, Zhe Liu, Bo Wang, Eun Chang, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer, "Neural-fst class language model for end-to-end speech recognition," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 6107–6111.
[3] Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L. Seltzer, "Deep shallow fusion for rnn-t personalization," in 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 251–257.
[4] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani, "Lamp: When large language models meet personalization," arXiv preprint arXiv:2304.11406, 2023.
[5] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen, "Joint grapheme and phoneme embeddings for contextual end-to-end asr.," in Interspeech, 2019, pp. 3490–3494