Entity Contrastive Learning in a Large-Scale Virtual Assistant System

Jonathan Rubin, Jason Crowley, George Leung, Morteza Ziyadi, Maria Minakova

Amazon Alexa, Cambridge MA, United States

Introduction

- The performance of a virtual assistant is heavily dependent upon how well named entity recognition (NER) tasks are handled.
- Mistaken slot predictions result in propagating incorrect information to downstream modules, causing sub-optimal interactions with users of the system.
- Contrastive learning can be used to improve NER model training by attempting to cluster similar inputs closer together in representation space and repelling dissimilar inputs apart.
- Token contrastive learning (Das et al., 2023) attracts and repels representations at the token level.

Virtual Assistant System Overview

- In this work we apply contrastive learning to improve the performance of a ubiquitous virtual assistant system.
- We first train a common encoder using contrastive sentence embedding (Gao et al., 2021).
- Next, we incorporate ***entity contrastive learning***, based on (Das et al., 2023) to better cluster similar entities together in representation space.
- We train and evaluate joint intent classifiers and named entity recognition models for 11 virtual assistant domains, including music, video, shopping, knowledge, books, sports, calendar etc...

Joint IC and NER Training

- Joint IC-NER models are trained separately for each domain. The model encodes a sequence of (sub-word) utterance tokens through a transformer encoder architecture: $[h_1, h_2, \dots, h_n] = T_{Encoder}([x_1, x_2, \dots, x_n])$.
- In addition to sub-words that are fed to the encoder, each input token is also flagged as either being recognized or unrecognized via lookup in a large gazetteer, $\phi(\cdot) \in \{0,1\}$, which further undergoes a separate gazetteer-based embedding, $[g_1, g_2, \dots, g_n] = G_{Embedding}([\phi(x_1), \phi(x_2), \dots, \phi(x_n)])$.
- Gazetteer embeddings are then combined with the output embeddings of the encoder, $[t_1, t_2, \dots, t_n] = [h_1 \otimes g_1, h_2 \otimes g_2, \dots, h_n \otimes g_n]$, where $\otimes$ is the element-wise product.
- These embeddings are then used by both the IC and NER model heads. The intent classification head accepts a single aggregated embedding that it processes through a collection of linear layers. Its loss function is the standard categorical cross entropy loss (ℓ_{CE}).
- The NER head accepts all embeddings and performs per token classification. Our NER model employs a conditional random field (CRF) to optimize the sequence labeling task (ℓ_{CRF}).

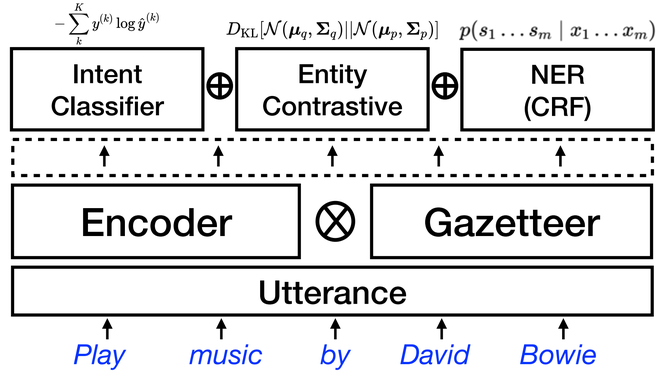


Fig. 1. A schematic overview of a jointly trained IC and NER model with a gazetteer feature and optional entity contrastive learning.

Entity Contrastive Training

- When employing entity contrastive training, a third loss component is added to model training. Diagonal Gaussian embeddings, $\mathcal{N}(\mu_i, \Sigma_i)$, are created. Gaussian embeddings map tokens to densities rather than point vectors and have been shown to better capture representation uncertainty.
- The KL divergence between two diagonal Gaussian distributions is used to evaluate a pair of tokens from a collection of utterances:

$$D_{\mathrm{KL}}[\mathcal{N}(\mu_q, \Sigma_q) \| \mathcal{N}(\mu_p, \Sigma_p)] = \frac{1}{2}\left(\mathrm{Tr}\left(\Sigma_p^{-1}\Sigma_q\right) - l + \log\frac{|\Sigma_p|}{|\Sigma_q|} + (\mu_p - \mu_q)^T \Sigma_p^{-1}(\mu_p - \mu_q)\right)$$

- Given a collection of entities and their labels within a batch, $(x_q, y_q) \in \mathcal{X}$, a set of in-batch matching entities, $\mathcal{X}_p$, can be constructed by locating different tokens that share the same entity label ($y_p = y_q, p \neq q$). The final ℓ_{ENT} loss is constructed for each entity, p, in a batch, $\mathcal{X}$, as follows (where $d(\cdot,\cdot)$ is the mean of both both forward and reverse KL.

$$\ell_{ENT} = -\frac{1}{|\mathcal{X}|}\sum_{p\in\mathcal{X}} \log \frac{\sum_{(x_q,y_q)\in\mathcal{X}_p} \exp(-d(p,q))/\,|\mathcal{X}_p|}{\sum_{(x_q,y_q)\in\mathcal{X}, p\neq q} \exp(-d(p,q))}$$

- The final loss function is a linear combination of:

$$\mathcal{L}_{overall} = w_1 \cdot \ell_{CE} + w_2 \cdot \ell_{CRF} + w_3 \cdot \ell_{ENT}$$

Results

Offline (full system) Results

Profile 1	SEMER ↓	ICER ↓	IRER ↓
Contrastive Encoder	-10.7%	-16.2%	7.9%
Entity Contrastive Training	**-12.7%**	**-17.5%**	**-10.7%**
Profile 2	SEMER ↓	ICER ↓	IRER ↓
Contrastive Encoder	-9.2%	14.6%	6.6%
Entity Contrastive Training	**-11.0%**	**-16.2%**	**-9.0%**

Table 1. Error results compared to a baseline model. Lower is better. Contrastive encoder only training is compared to full entity contrastive learning.

Offline (per domain) Results

↓ Lower is better	**Profile 1**		**Profile 2**	
Domain	Contrastive Encoder (%)	Entity Contrastive (%)	Contrastive Encoder (%)	Entity Contrastive (%)
Global	-19.43	**-19.91**	-17.55	**-18.19**
Music	-7.79	**-11.77**	-8.11	**-11.71**
Notifications	-14.38	**-17.20**	-12.37	**-16.32**
Video	-14.18	**-17.02**	-6.23	**-9.24**
Shopping	**-14.29**	-7.19	**-11.63**	-8.08
Local Search	-15.34	**-23.94**	-16.42	**-25.17**
General Media	-17.30	**-17.63**	-18.23	**-18.28**
Calendar	**-3.21**	-0.96	**-6.76**	-4.50
Books	-11.93	**-17.19**	-8.34	**-14.76**
Cinema Show Times	**-1.78**	+17.08	**-13.87**	+13.87
Sports	**-0.02**	**-0.02**	**-12.00**	-11.97

Table 2. Relative improvement (SEMER) results compared to a baseline model. Lower is better. **Contrastive Encoder** contrastively fine-tunes a common encoder. **Entity Contrastive** further adds an entity contrastive loss function. Results are shown for two virtual assistant profiles.

Online (A/B test) Results

	D_{rules} ↓	D_{stat} ↓	$D_{\text{stat-tail}}$ ↓
Global	0.03	1.97	1.10
Music	-1.85†	-0.01†	-0.06†
Shopping	-13.09†	-8.27†	-8.72†
Video	7.48†	1.89†	2.40†
Overall	**-0.79†**	**-0.55**	**-0.68†**

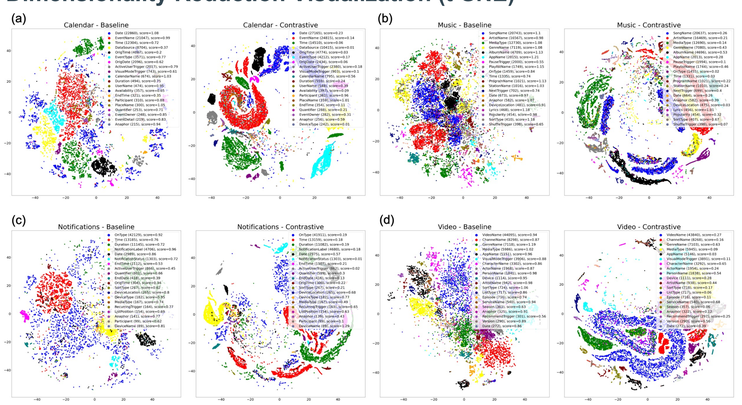
Table 3. A/B test results on live traffic. Shows relative percentage change of user dissatisfaction against the control inferred using behavioral rules and a statistical model applied to all traffic and tail-distribution traffic only. Lower is better.

Embeddings Analysis

Domain	Baseline ↓	Contrastive ↓
Video	0.95	**0.28**
Sports	**0.41**	0.54
Shopping	0.85	**0.14**
Notifications	0.84	**0.21**
Music	1.03	**0.27**
Local Search	1.00	**0.30**
Global	0.77	**0.28**
General Media	0.89	**0.18**
Cinema Show Times	0.71	**0.28**
Calendar	0.83	**0.15**
Books	0.85	**0.15**
Average	0.83	**0.25**

Table 4. Embedding Alignment scores per domain. Lower is better

Dimensionality Reduction Visualization (t-SNE)



References

Das et al. (2022). **"CONTaiNER: Few-shot named entity recognition via contrastive learning".** In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Gao et al. (2021). "SimCSE: Simple contrastive learning of sentence embeddings". In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.