Entity Contrastive Learning in a Large-Scale Virtual Assistant System

Jonathan Rubin
Amazon Alexa AI, USA
jonrubin@amazon.com

Jason Crowley
Amazon Alexa AI, USA
jascrowl@amazon.com

George Leung*
Novo Nordisk A/S, USA
kwlg@novonordisk.com

Morteza Ziyadi
Amazon Alexa AI, USA
mziyadi@amazon.com

Maria Minakova
Amazon Alexa AI, USA
minakova@amazon.com

Abstract

Conversational agents are typically made up of domain (DC) and intent classifiers (IC) that identify the general subject an utterance belongs to and the specific action a user wishes to achieve. In addition, named entity recognition (NER) performs per token labeling to identify specific entities of interest in a spoken utterance. We investigate improving joint IC and NER models using entity contrastive learning that attempts to cluster similar entities together in a learned representation space. We compare a full virtual assistant system trained using entity contrastive learning to a baseline system that does not use contrastive learning. We present both offline results, using retrospective test sets, as well as online results from an A/B test that compared the two systems. In both the offline and online settings, entity contrastive training improved overall performance against baseline systems. Furthermore, we provide a detailed analysis of learned entity embeddings, including both qualitative analysis via dimensionality-reduced visualizations and quantitative analysis by computing alignment and uniformity metrics. We show that entity contrastive learning improves alignment metrics and produces well-formed embedding clusters in representation space.

1 Introduction

Named Entity Recognition (NER) is a well-studied and fundamental task within Natural Language Understanding (NLU). The performance of a virtual assistant is heavily dependent upon how well 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. Contrastive learning attempts to cluster similar inputs closer together in their representation space and repel dissimilar inputs apart. Token contrastive learning attracts and repels representations at the token level and was introduced in (Das et al., 2022) for improving performance in few-shot NER tasks.

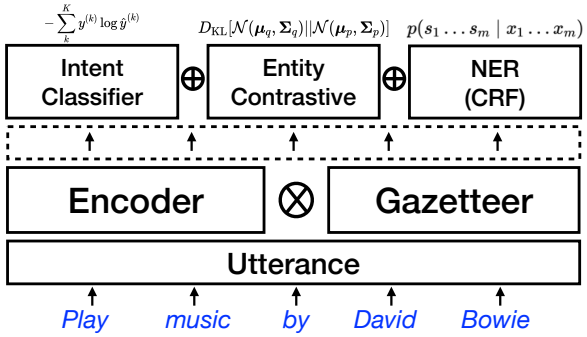

Figure 1: A schematic overview of a jointly trained IC and NER model with a gazetteer feature and optional entity contrastive learning.

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., 2022), to better cluster similar entities together in representation space. We train and evaluate joint IC and NER models in 11 domains. For each domain, we evaluate performance with and without an additional entity contrastive loss. We further provide results of an online A/B test that measures user satisfaction and show improved performance when using entity contrastive training. Furthermore, we perform a detailed embeddings analysis to determine the effect that the entity contrastive loss function has on entity representations. In particular, we compute alignment and uniformity metrics (Wang and Isola, 2020) of learned entity representations. Finally, we also present qualitative results in the form of t-SNE visualizations comparing models with entity contrastive training vs. without. We show that entity contrastive learning improves alignment metrics as

*Work done during the author's tenure at Amazon.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 159–171

well as clustering behavior in representation space.

2 Virtual Assistant System Overview

Fig. 1 shows a schematic overview of a jointly trained IC and NER model that makes up part of the NLU component of a full virtual assistant system. Joint IC-NER models are trained separately for each domain. The IC-NER model encodes a sequence of (sub-word) utterance tokens, $x_1, x_2, \ldots, x_n$, through a transformer encoder architecture, $[h_1, h_2, \ldots, h_n] = T_{Encoder}([x_1, x_2, \ldots, x_n])$. 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, $[g_1, g_2, \ldots, g_n] = G_{Embedding}([\phi(x_1), \phi(x_2), \ldots, \phi(x_n)])$. Gazetteer embeddings are then combined with the output embeddings of the encoder, $[t_1, t_2, \ldots, t_n] = [h_1 \otimes g_1, h_2 \otimes g_2, \ldots, h_n \otimes g_n]$, where $\otimes$ is the element-wise product. These embeddings are then used by both the IC and NER model heads.

2.1 Joint IC and NER Training

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, $\ell_{CE} = -\sum_k^K y^{(k)} \log \hat{y}^{(k)}$, where K is the total number of intent classes per domain, $y^{(k)}$ is 0 or 1 ground truth for intent class, k, and $\hat{y}^{(k)}$ is the predicted value for that intent.

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:

$$p(s_1 \ldots s_n \mid t_1 \ldots t_n; \boldsymbol{w}) = \frac{\exp(\boldsymbol{w} \cdot \Phi(t_1 \ldots t_n, s_1 \ldots s_n))}{\sum_{s'_1 \ldots s'_n \in \mathcal{S}^m} \exp(\boldsymbol{w} \cdot \Phi(t_1 \ldots t_n, s'_1 \ldots s'_n)}$$

$$\ell_{CRF} = -\sum_{i=1}^{M} \log p_i(s_1 \ldots s_n \mid t_1 \ldots t_n; \boldsymbol{w})$$

where $\boldsymbol{w}$ are learnable weights, M is the number of utterances, S_m is the space of all possible sequences and $\Phi(t_{i\ldots}, s_{i\ldots})$ is the product of selected potential functions that reflects the plausibility score of a given labeling, see (Lafferty et al., 2001) for further details.

2.2 Entity Contrastive Training

When employing entity contrastive training, a third loss component is added to model training, as described in (Das et al., 2022). Diagonal Gaussian embeddings, $\mathcal{N}(\mu_i, \Sigma_i)$, are created by passing each encoded token representation, t_i, through separate networks, $\mu_i = f_\mu(t_i)$ and $\Sigma_i = \text{ELU}(f_\Sigma(t_i)) + (1 + \epsilon)$. These networks respectively infer the mean and variance of the Gaussian embeddings. Here, ELU is the Exponential Linear Unit and ϵ is added for numerical stability. Gaussian embeddings map tokens to densities rather than point vectors and have been shown to better capture representation uncertainty (Vilnis and McCallum, 2015). As the KL divergence between two diagonal Gaussian distributions has a closed form solution, a pair of tokens from a collection of utterances can be evaluated as follows (note that l is the embedding dimension):

$$\begin{aligned} & D_{\text{KL}}\left[\mathcal{N}\left(\mu_q, \Sigma_q\right) || \, \mathcal{N}\left(\mu_p, \Sigma_p\right)\right] \\ &= \frac{1}{2}\left(\text{Tr}\left(\Sigma_p^{-1}\Sigma_q\right) - l + \log\frac{|\Sigma_p|}{|\Sigma_q|}\right. \\ &\quad \left. + \left(\mu_p - \mu_q\right)^T \Sigma_p^{-1}\left(\mu_p - \mu_q\right)\right) \end{aligned} \quad (1)$$

Further, as the KL divergence is not symmetric, both forward and reverse directions are considered: $d(p, q) = \frac{1}{2}\left(D_{\text{KL}}\left[\mathcal{N}_q \| \mathcal{N}_p\right] + D_{\text{KL}}\left[\mathcal{N}_p \| \mathcal{N}_q\right]\right)$.

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 ($y_p = y_q$, where $p \neq q$). The final ℓ_{ENT} loss is constructed for each entity, p, in a batch, $\mathcal{X}$, as follows:

$$\ell_{ENT} = -\frac{1}{|\mathcal{X}|}\sum_{p \in \mathcal{X}} \log \frac{\sum_{(x_q, y_q) \in \mathcal{X}_p} \exp(-d(p, q)) / \, |\mathcal{X}_p|}{\sum_{(x_q, y_q) \in \mathcal{X}, p \neq q} \exp(-d(p, q))} \quad (2)$$

2.3 Overall Loss Function

The final loss function is a linear combination of the cross entropy loss of the intent classifier, the CRF loss given by the NER output and the entity contrastive loss:

$$\mathcal{L}_{overall} = w_1 \cdot \ell_{CE} + w_2 \cdot \ell_{CRF} + w_3 \cdot \ell_{ENT} \quad (3)$$

where $w_1 \ldots w_3$ are hyper-parameters that weight each of the individual loss components. In our experiments we set each $w_i = 1$.

↓ 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 1: 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.

2.4 Implementation Details

We use a BERT (Devlin et al., 2019) style encoder with embedding dimension 768 and Gaussian embedding dimension 128. The encoder is made up of 4 hidden layers with 16 attention heads. The encoder's weights are first initialized via a task-specific model distillation procedure (Citation anonymized due to self-reference). Encoder weights are further fine-tuned using contrastive sentence embedding (Gao et al., 2021), where a single positive utterance is contrasted with 10 negative utterances. The fine-tuned encoder is common and shared between domains. Each domain's IC-NER model is then further trained for a maximum of 60 epochs and early stopping was invoked if there was no improvement in validation error rate for 4 epochs.

3 Experimental Results

We provide experimental results in the following three settings:

Offline (per domain): We compare 11 domain models trained using entity contrastive learning vs. baseline models without entity contrastive training. All domains that utilize gazetteers are included.

Offline (full system): We compare a full virtual assistant system trained using entity contrastive learning against a baseline system on a collection of static test-sets.

Online: We conduct an A/B test using live traffic to compare a full virtual assistant system trained using entity contrastive learning vs. a baseline model that does not.

Full descriptions of each error metric used for (offline) evaluation are given in Appendix A. We provide brief summaries here:

SEMER: Semantic Error Rate reflects the proportion of incorrectly labeled entities and intents.

ICER: Intent Classification Error Rate measures the proportion of misclassified intents

IRER: Intent Recognition Error Rate measures how often predictions contain **any** mistakes in either entities or intent.

3.1 Offline (per domain) Results

Table 1 shows per domain relative improvement SEMER results compared to a live baseline model that doesn't utilize entity contrastive training. Lower results are better. Two candidate models are compared: 1) **Contrastive Encoder**, where only the encoder was pre-trained using supervised sentence contrastive learning based on (Gao et al., 2021) and 2) **Entity Contrastive**, which builds on top of 1), and further trains using the entity contrastive loss function from Section 2.2. Results are shown for two virtual assistant profiles. Profile 1 is a voice only system, whereas Profile 2 is an assistant that has a display monitor.

Cinema Show Times was the only domain that did worse than the baseline when using entity contrastive training. This may be due to the relatively large number of entity types (33) and the smaller training and validation dataset size (30,311 and 3,368, respectively). Appendix B lists the total

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 2: Error results compared to a baseline model. ↓ Lower is better. Contrastive encoder only training is compared to full entity contrastive learning.

	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 comparing an experimental virtual assistant system that employs entity contrastive learning against a baseline control system. Measurements show relative percentage change of user dissatisfaction against the control inferred using behavioral rules (D_{rules}), a statistical model applied to all traffic (D_{stat}) and tail-distribution traffic only ($D_{\text{stat-tail}}$). ↓ Lower is better. †Indicates statistically significant results at a 95% confidence level.

number of utterances in both training and validation datasets, as well as the number of entities labels for all 11 domains. All other domains improved against the baseline. Overall, **entity contrastive** training out-performed **contrastive encoder** training in 8 out of 11 domains for Profile 1 and 7 out of 11 domains for Profile 2. Furthermore, **entity contrastive** training achieved the best results for the top four highest-traffic domains in both profiles.

3.2 Offline (full system) Results

Table 2 shows overall relative improvement against a baseline system measured using SEMER, ICER and IRER metrics. Once again we compare a virtual assistant system that trained a contrastive encoder only vs. full entity contrastive training. We see that entity contrastive training leads to larger relative improvement, compared to contrastive encoder training only, for all metrics.

3.3 Online (A/B test) Results

The final set of results we present were collected from an A/B test using an experimentation platform to evaluate full virtual assistant systems on live customer traffic. Once again we compare a system that uses entity contrastive training against a baseline model that does not. The experimental (contrastive) and control (baseline) model each received 10% of customer traffic and the A/B test ran for two weeks. As no ground truth is available for online data, we rely on a rule based system (D_{rules}), and a statistical model (D_{stat}) that infers user dissatisfaction given a virtual assistant's response. We also measure user dissatisfaction specifically for tail traffic, i.e. the bottom 40% of frequent utterances ($D_{\text{stat-tail}}$).

Results are presented as relative comparisons to the baseline system in Table 3. Per-domain results are included for domains of special interest, including those with higher traffic volumes. The overall results, in the final row, evaluate the full virtual assistant system on all domains. Lower results are better. Overall the experimental contrastive model improved all user dissatisfaction metrics. Results are statistically significant at the 95% confidence level ($p < 0.05$) for D_{rules} and $D_{\text{stat-tail}}$ and just outside the range for D_{stat} ($p = 0.058$). Per-domain results show that the baseline model outperformed the experimental model for Global (however not statistically significantly, $p > 0.05$), and Video ($p < 0.05$). Further analysis showed that the experimental model likely incorrectly predicted the Video domain on device profiles that didn't have display capability. The largest improvements were observed in the Shopping domain ($p < 0.05$) and there are also improvements in Music (although not statistically significant, $p > 0.05$).

4 Embeddings Analysis

We further provide qualitative and quantitative analysis of the entity representations learned by a baseline and contrastive model. The baseline model differs to the contrastive model only by removing the ℓ_{ENT} component of the loss function in Eqn. (3).

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: Alignment scores per domain comparing baseline vs. contrastive NER learning. ↓ Lower is better.

4.1 Qualitative: Dimensionality Reduction Visualization

For each domain, we derived t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) plots to visualize entity representations learned by the baseline and contrastive model. Embeddings were pulled from a randomized subset of validation data. Dimensionality reduction took place on the μ_i representations learned by each model, $\mathbb{R}^{128} \rightarrow \mathbb{R}^2$. Fig. 2 shows a comparison between the baseline and contrastive model for four domains (a) Calendar, (b) Music, (c) Notifications and (d) Video. Appendix D displays t-SNE plots for the remaining domains. Looking at Fig. 2(a) for the Calendar domain, we can see that points for the most frequent entity type (Date) don't appear to cluster at all and are quite dispersed in the t-SNE plot on the left (baseline). However, in the plot on the right (contrastive) we see a well-formed cluster for Date in the top right. We also notice in Fig. 2(b) for the Music domain, the most frequent entity type (SongName) exhibits some clustering behavior in the baseline, but forms multiple distinct clusters in the contrastive model. We can also easily see points that did not have the SongName label within these clusters. In particular, there are many overlapping points for AlbumName, ArtistName and Lyrics. AlbumName and Lyrics can likely overlap with SongName and cause confusion for the model. Given that the data-set is very large, annotation errors are also frequent and it is possible these overlapping points could potentially identify errors in the labeling process.

4.2 Quantitative: Alignment and Uniformity

We further analyze representation quality using the quantitative metrics of alignment and uniformity introduced in (Wang and Isola, 2020). The alignment metric assumes a distribution of positive pairs and calculates expected distance between representations of these pairs. Positive pairs should lie closer together in representation space and produce lower values. Conversely, uniformity measures how well learned representations are distributed uniformly on a unit hyper-sphere for instances from *all* classes.

Given that we do not rely on positive pairs, but instead wish to align token representations belonging to the same class (i.e. has the same entity label), we slightly alter the original alignment metric to consider all non self-referential, pairwise comparisons between instances that belong to the same class, $p_{cls|x \neq y}$. The uniformity metric remains the same as in (Wang and Isola, 2020). We set hyperparameters as follows, $\alpha = 2$ and $t = 2$.

$$\mathcal{M}_{\text{align}}(f; \alpha) = \mathbb{E}_{x,y \sim p_{cls|x \neq y}} \left[\|f(x) - f(y)\|_2^{\alpha} \right] \tag{4}$$

$$\mathcal{M}_{\text{uniform}}(f; t) = \log \mathbb{E}_{x,y \sim p_{\text{data}}} \left[e^{-t\|f(x)-f(y)\|_2^2} \right] \tag{5}$$

Table 4 shows alignment values per domain. The values in Table 4 are computed by taking the average alignment scores for all entities within each domain. Alignment values for each entity type are given in Appendix C. A weighted average is taken that considers the number of tokens with a given entity label. Lower values imply better alignment between representations within the same class.

We can see in Table 4 that all domains have lower alignment values with entity contrastive training, except for the Sports domain. The Sports domain has the least amount of training data and entity types (see Appendix B), which may be the reason that entity contrastive training does not result in improvement over the baseline model.

Finally, we compute uniformity metrics. To reduce computational cost, we randomly sample 10% of the entity embeddings. The uniformity scores for the baseline and contrastive models were -3.54 and -3.11, respectively, indicating that the baseline model produced embeddings that are likely more uniformly distributed than the contrastive model.

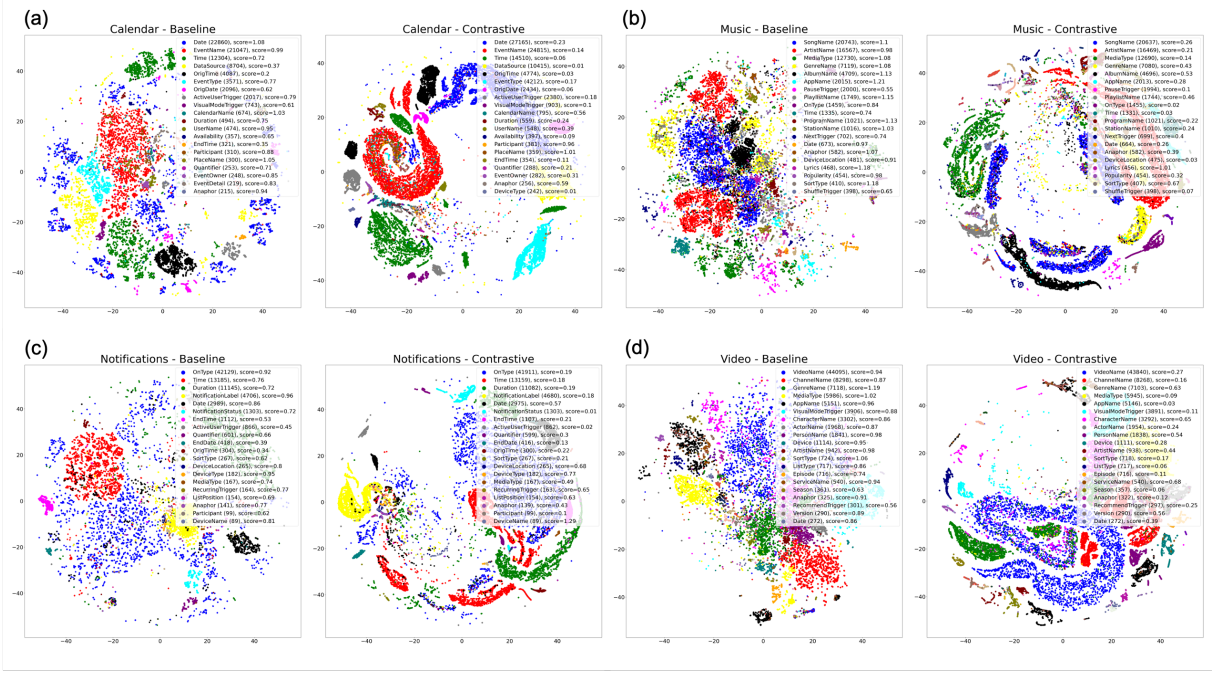

Figure 2: A collection of t-SNE plots comparing embeddings from a baseline (left figure) and contrastive model (right figure) in four domains (a) Calendar, (b) Music, (c) Notifications, (d) Video. Legend entries are restricted to the top 20 most frequent slot labels with counts shown in parentheses. Alignment scores are also shown.

5 Related Work

Contrastive learning has been applied with tremendous success over the last few years in tasks that process data such as audio (Oord et al., 2018), vision (Chen et al., 2020) and natural language (Fang et al., 2020). Contrastive losses, such as InfoNCE (Oord et al., 2018; Hénaff et al., 2019), build on the original idea of noise contrastive estimation (Gutmann and Hyvärinen, 2010; Mnih and Kavukcuoglu, 2013) that learns a data distribution by comparing it against a chosen noise distribution. Contrastive representation learning can either be unsupervised (Chen et al., 2020; He et al., 2020) or supervised (Khosla et al., 2020). Unsupervised or self-supervised approaches have relied upon techniques such as data augmentation (Chen et al., 2020; He et al., 2020) and future self prediction (Oord et al., 2018) as a way of ignoring superfluous information to learn better class representations. Supervised approaches (Khosla et al., 2020) incorporate class label information during learning and were introduced to avoid problems with in-batch false positives. In natural language tasks, contrastive learning approaches based on data augmentation techniques have not fared as well compared to their vision counterparts. SimCSE (Gao et al., 2021) introduced both unsupervised and supervised approaches for learning contrastive sentence embeddings. The unsupervised approach relies solely on varying dropout masks to achieve different representations of the same input sentence, whereas the supervised task uses examples from natural language inference datasets (Conneau et al., 2017). Rather than learning sentence embeddings, (Das et al., 2022) introduced token contrastive learning in the context of improving few-shot learning. Our work does not focus on few-shot learning, but instead seeks to evaluate joint IC-NER models trained with entity contrastive learning for the purpose of improving a large-scale virtual assistant system.

6 Conclusion

We presented jointly trained IC and NER models augmented with entity contrastive learning via an additional loss function that attempts to pull similar entities together in representation space, and repel dissimilar entities apart. We provided a comprehensive evaluation of entity contrastive learning within a full virtual assistant system by comparing to baselines in both offline and online (A/B test) experiments. Results show that employing entity contrastive learning improves overall error and alignment metrics and produces well-formed embedding clusters in representation space.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 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.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. 2019. Data-efficient image recognition with contrastive predictive coding. *CoRR*, abs/1905.09272.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, 26.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Luke Vilnis and Andrew McCallum. 2015. Word representations via gaussian embedding. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

A Performance (Error) Metrics

The error metrics used to assess offline performance are as follows:

SEMER: Semantic Error Rate evaluates slot-filling and intent classification performance jointly, as follows:

$$\frac{\text{\# Deletion + \# Insertion + \# Substitution}}{\text{\# Correct + \# Deletion + \# Substitution}}$$

Deletion occurs when the slot name is present in ground truth but not in the prediction. Insertion is the opposite when extra slot names are included in the prediction. Substitution errors occur when predictions do match ground truth slot labels, but for an incorrect slot value. Correct slots are when both the slot name and slot value match. Intent classification errors are also counted as substitution errors above.

ICER: Intent Classification Error Rate measures the rate at which the intent of utterances are incorrectly predicted:

$$\text{ICER} = \frac{\text{\# Incorrect Intents}}{\text{\# Total Utterances}}$$

IRER: Intent Recognition Error Rate measures how often predictions contain **any** mistake in either slots or intent.

$$\text{IRER} = \frac{\text{\# Incorrect (Slot or Intent)}}{\text{\# Total Utterances}}$$

B Dataset Sizes

Table 5 shows the dataset sizes (training and validation) for 11 gazetteer based domains. Also depicted are the total number of entities per domain. Domains are listed in descending order based on number of utterances. Global is the largest domain and Sports is the smallest.

C Alignment Tables Per Domain

Alignment scores per slot are shown for each domain in Tables 6 to 16. The baseline model includes no entity contrastive training. Results are restricted to the top ten most frequent slots due to display purposes. The missing remaining slots exhibit similar trends to those shown. Size refers to the number of tokens with a given slot label and Score is the alignment score. Lower is better. The final column shows relative change as a percentage. Negative values show improvement of the contrastive model over the baseline.

D t-SNE Visualizations

Figs. 3 and 4 depict the remaining t-SNE plots not shown in the main body of the text. Once again, for each domain, the baseline embeddings are on the left and the contrastive model embeddings are on the right. As in the figures in the main body, non-entity (O) tokens are removed as they are not subject to contrastive training and legend entries are restricted to the top 20 most frequent slot labels with counts shown in parentheses. Alignment scores are also shown. As with the figures in the main body, we see improved clustering behavior in the contrastive embeddings compared to the baseline embeddings in all domains, except for the Sports domain, which is quite sparse. It is also possible that the perplexity value (which depends on dataset size) is not optimal for the sports domain due to the smaller dataset size.

Domain	Training instances	Validation instances	Number of entities
Global	3,165,309	351,702	117
Music	2,160,488	240,055	119
Notifications	818,963	90,996	62
Video	686,520	76,280	63
Shopping	602,748	66,972	54
Local Search	294,098	32,678	75
General Media	167,776	18,642	30
Calendar	137,313	15,258	46
Books	125,139	13,905	50
Cinema Show Times	30,311	3,368	33
Sports	21,347	2,372	13
Total	**8,210,012**	**912,228**	**662**

Table 5: Total number of training and validation utterances for 11 domains that utilize entity contrastive learning in a large-scale virtual assistant system.

Calendar	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
Date	22860	1.08	27165	0.23	-78.21
EventName	21047	0.99	24815	0.14	-85.61
Time	12304	0.72	14510	0.06	-91.89
DataSource	8704	0.37	10415	0.01	-96.92
OrigTime	4087	0.20	4774	0.03	-83.13
EventType	3571	0.77	4212	0.17	-77.91
OrigDate	2096	0.62	2434	0.06	-89.63
ActiveUserTrigger	2017	0.79	2380	0.18	-77.19
VisualModeTrigger	743	0.61	903	0.10	-82.95
CalendarName	674	1.03	795	0.56	-45.65

Table 6: Alignment scores per slot for Calendar domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

Music	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
SongName	20743	1.10	20637	0.26	-76.18
ArtistName	16567	0.98	16469	0.21	-78.55
MediaType	12730	1.08	12690	0.14	-87.52
GenreName	7119	1.08	7080	0.43	-59.86
AlbumName	4709	1.13	4696	0.53	-53.33
AppName	2015	1.21	2013	0.28	-77.14
PauseTrigger	2000	0.55	1994	0.10	-82.58
PlaylistName	1749	1.15	1744	0.46	-59.87
OnType	1459	0.84	1455	0.02	-98.10
Time	1335	0.74	1331	0.03	-96.18

Table 7: Alignment scores per slot for Music domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

Notifications	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
OnType	42129	0.92	41911	0.19	-79.61
Time	13185	0.76	13159	0.18	-76.72
Duration	11145	0.72	11082	0.19	-73.28
NotificationLabel	4706	0.96	4680	0.18	-81.83
Date	2989	0.86	2975	0.57	-33.90
NotificationStatus	1303	0.72	1303	0.01	-98.13
EndTime	1112	0.53	1107	0.21	-60.03
ActiveUserTrigger	866	0.45	862	0.02	-96.29
Quantifier	601	0.66	599	0.30	-54.46
EndDate	418	0.39	416	0.13	-66.83

Table 8: Alignment scores per slot for Notifications domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

Video	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
VideoName	44095	0.94	43840	0.27	-71.26
ChannelName	8298	0.87	8268	0.16	-81.68
GenreName	7118	1.19	7103	0.63	-47.49
MediaType	5986	1.02	5945	0.09	-90.99
AppName	5151	0.96	5146	0.03	-96.84
VisualModeTrigger	3906	0.88	3891	0.11	-87.14
CharacterName	3302	0.86	3292	0.65	-24.85
ActorName	1968	0.87	1954	0.24	-72.31
PersonName	1841	0.98	1838	0.54	-44.64
Device	1114	0.95	1111	0.28	-70.81

Table 9: Alignment scores per slot for Video domain – baseline vs. contrastive. Note that the number slots has been truncated for display purposes.

Shopping	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
ItemName	51246	0.91	53039	0.10	-88.83
ShoppingListType	5685	0.71	5876	0.28	-61.31
ProductSortType	4632	0.91	4797	0.30	-66.96
VisualModeTrigger	2430	0.46	2527	0.04	-92.11
ShoppingServiceName	1179	0.70	1210	0.06	-91.54
RecommendTrigger	1118	0.41	1158	0.23	-44.14
DealType	628	0.54	650	0.25	-54.14
Anaphor	471	0.64	481	0.39	-38.35
Quantifier	451	0.70	470	0.27	-61.93
PurchaseDate	388	0.64	450	0.47	-27.03

Table 10: Alignment scores per slot for Shopping domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

Local Search	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
PlaceName	42865	1.08	39231	0.22	-79.71
PlaceType	10667	1.12	9756	0.35	-68.90
DestinationPlaceName	10162	0.92	9315	0.19	-79.43
LocationSortType	6723	1.11	6196	0.24	-78.34
City	6082	1.09	5594	0.30	-72.69
Location	3745	1.06	3402	0.77	-27.54
DestinationLocation	3575	0.93	3233	0.30	-67.28
Anaphor	2611	0.98	2400	0.71	-27.91
Date	2292	0.91	2137	0.36	-60.37
PlaceFeature	2282	1.20	1999	0.64	-46.65

Table 11: Alignment scores per slot for Local Search domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

General Media	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
AppName	48421	0.91	48231	0.17	-81.31
MediaType	4697	0.88	4685	0.14	-84.14
VisualModeTrigger	1548	0.55	1548	0.05	-90.17
GenreName	657	1.04	651	0.91	-12.72
SettingValue	560	0.68	555	0.43	-36.53
SortType	392	0.69	392	0.12	-83.22
Anaphor	219	0.85	219	0.44	-48.91
DeviceBrand	218	0.75	218	0.20	-72.92
ListPosition	192	0.73	192	0.24	-67.19
DeviceType	89	0.71	89	0.49	-31.13

Table 12: Alignment scores per slot for General Media domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

Global	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
Setting	3591	0.82	2819	0.12	-85.10
MediaType	2536	0.78	2091	0.11	-85.87
DeviceType	2125	0.80	1765	0.22	-71.97
DeviceBrand	1971	0.70	1597	0.06	-91.48
ChannelName	1230	0.73	1072	0.15	-80.21
SearchContent	1038	0.84	858	0.21	-75.25
SettingValue	927	0.85	751	0.54	-36.92
DeviceLocation	558	0.81	471	0.38	-52.94
VisualModeTrigger	544	0.66	420	0.04	-93.17
ServiceName	535	0.82	411	0.33	-60.40

Table 13: Alignment scores per slot for Global domain – baseline vs. contrastive. Note that the number slots has been truncated for display purposes.

Books	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
BookName	34890	1.02	37703	0.16	-84.24
MediaType	23127	0.80	24870	0.08	-89.76
ServiceName	18965	0.77	20496	0.04	-94.93
AuthorName	4714	0.94	5075	0.42	-54.84
ActiveUserTrigger	3837	0.44	4170	0.03	-93.59
GenreName	3056	0.97	3313	0.61	-36.46
SortType	2994	0.56	3271	0.20	-64.82
SectionType	2078	0.90	2321	0.04	-95.76
Narrator	2057	0.60	2265	0.29	-51.25
Anaphor	1728	1.01	1861	0.29	-71.08

Table 14: Alignment scores per slot for Books domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

Cinema Show Times	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
MovieTitle	25800	1.06	7134	0.36	-66.04
EndTime	18795	0.19	5206	0.02	-90.80
MediaType	17210	0.66	4749	0.17	-73.88
PlaceName	9413	0.97	2594	0.31	-67.74
Time	8733	0.25	2413	0.04	-82.28
Date	4810	0.88	1314	0.71	-19.28
PlaceType	2679	0.88	751	0.23	-73.53
SortType	2190	0.92	602	0.30	-67.00
City	1828	0.86	502	0.96	10.82
PostalCode	1708	0.69	478	0.07	-89.20

Table 15: Alignment scores per slot for Cinema Show Times domain – baseline vs. contrastive. Results are restricted to the top ten most frequent slots due to display purposes.

Sports	Baseline		Contrastive		
Slot	Size	Score ↓	Size	Score ↓	% change
Date	739	0.37	721	0.57	55.33
SortType	130	0.62	124	0.43	-31.23
VisualModeTrigger	61	0.42	58	0.63	52.68
SportsRole	19	0.61	19	0.82	34.47
Time	18	0.37	18	0.02	-93.37
Sport	10	0.49	10	0.01	-98.04
League	4	0.04	4	0.00	-98.81
Anaphor	2	0.03	2	0.00	-97.17

Table 16: Alignment scores per slot for Sports domain – baseline vs. contrastive.

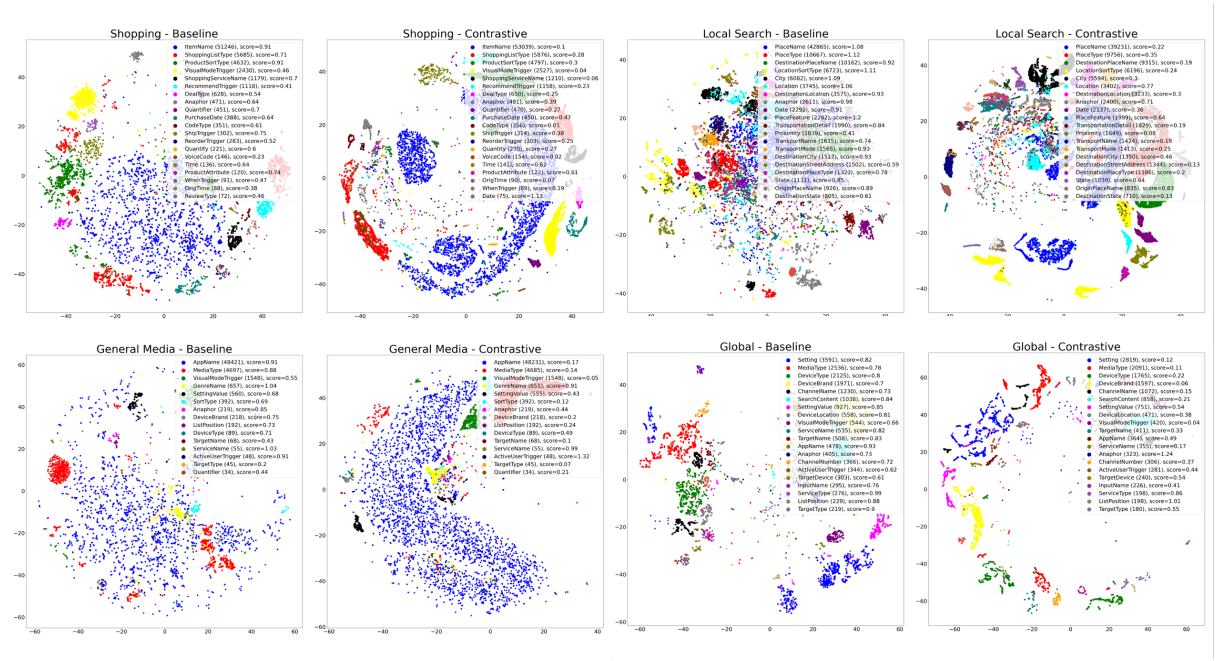

Figure 3: Remaining t-SNE plots for domains: Shopping (top left), Local Search (top right), General Media (bottom left) and Global (bottom right).

Figure 4: Remaining t-SNE plots for domains: Books (top left), Cinema Show Times (top right) and Sports (bottom middle).