Towards the Identification of Latent Structures in Language Embeddings

Ryunouske Abe, Takatomi Kubo, Kazushi Ikeda

Nara Institute of Science and Technology

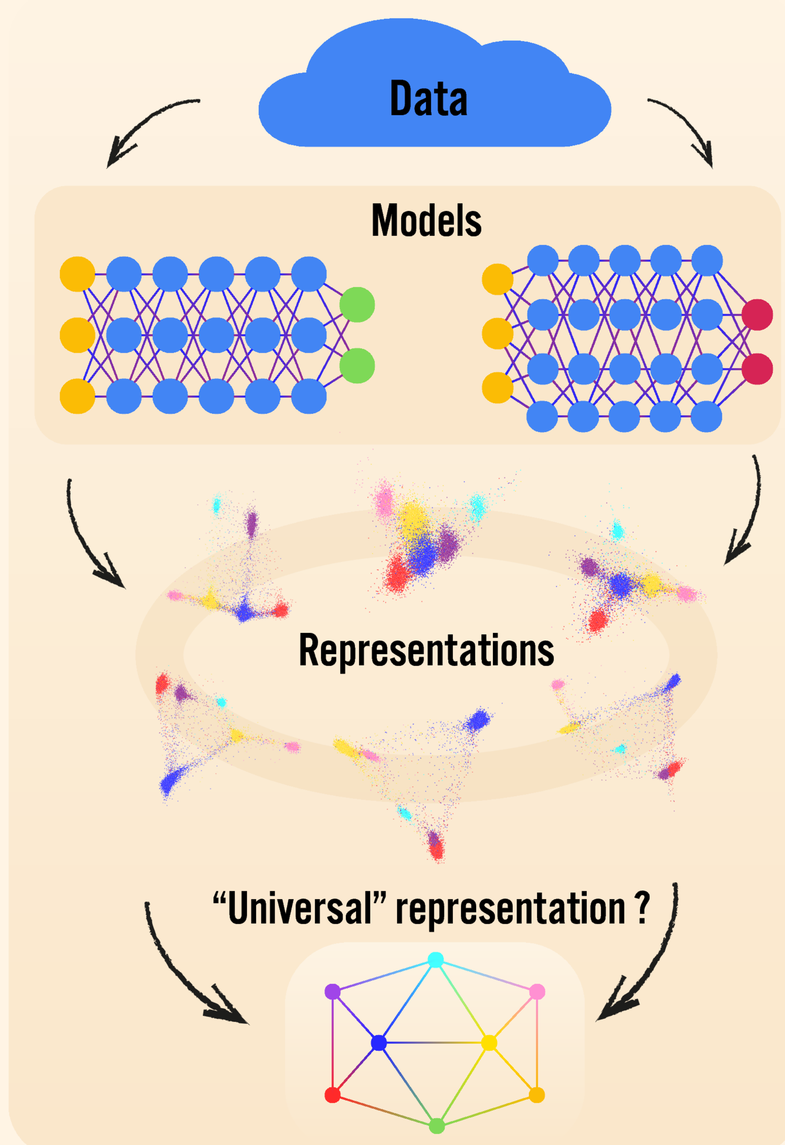
NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Background: Intrinsic universality in latent representations

Language models often share structural patterns, suggesting the presence of universal regularities in the data.

"Platonic representation hypothesis" (M. Huh et al.)



Aims: Identify the latent structure

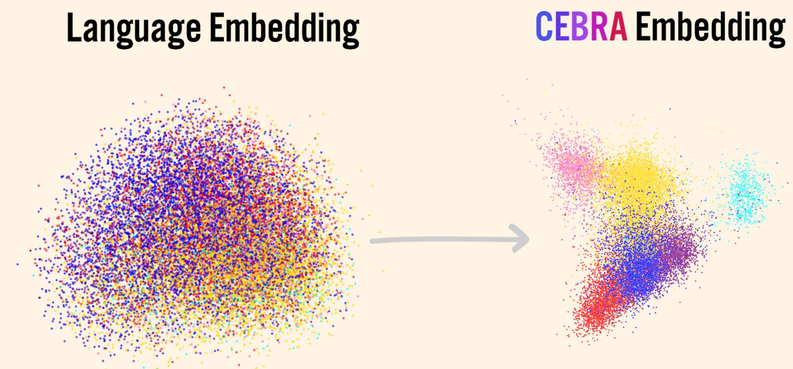
Low-dimensional structure inherent in high-dimensional language embeddings.

Approach: CEBRA

A contrastive learning framework, grounded in the theory of nonlinear ICA.

CEBRA

CEBRA: a contrastive learning framework, origially developed for neural data. Ensures Identifiability up to linear transformation, by leveraging auxiliary variables. (e.g., time, labels).



Objective function

generalized InfoNCE loss:

$$\mathcal{L}[\psi]_n = \underset{\substack{\mathbf{x}\sim p_{\mathcal{D}}(\mathbf{x}),\mathbf{y}_+\sim p(\mathbf{y}|\mathbf{x}) \\ \mathbf{y}_1,\dots,\mathbf{y}_n\sim q(\mathbf{y}|\mathbf{x})}}{\mathbb{E}}\left[-\psi(\mathbf{x},\mathbf{y}_+) + \log\sum_{i=1}^{n} e^{\psi(\mathbf{x},\mathbf{y}_i)}\right]$$

Minimizer of generalized InfoNCE loss:

$$\psi^*(\mathbf{x},\mathbf{y}) = \log\frac{p(\mathbf{y}\mid\mathbf{x})}{q(\mathbf{y}\mid\mathbf{x})} + C(\mathbf{x}),$$

The representations matche:

$$\mathrm{f(x)} = \mathrm{L\tilde{f}(x)}$$

Materials & Methods

Datasets

- dair-ai/emotion
- GoEmotions
- AG-news
- 20 news group

Language Embedding Models

- BERT-base-uncased
- RoBERTa-base
- all-mpnet-base-v2
- all-MiniLM-L6-v2
- EmbeddingGemma-300M
- Qwen3-Embedding-8B
- jina-embeddings-v2-base-en
- granite-embedding-english-r2

sentence-transformer

CEBRA

- offset1-model-mse: Similarity in the latent space: Euclidian
- offset1-model-v4: Similarity in the latent space: Cosine

Pipeline:

1. Extract high-dimensional embeddings from datasets.
2. Train CEBRA with discrete emotion labels as a AUX.

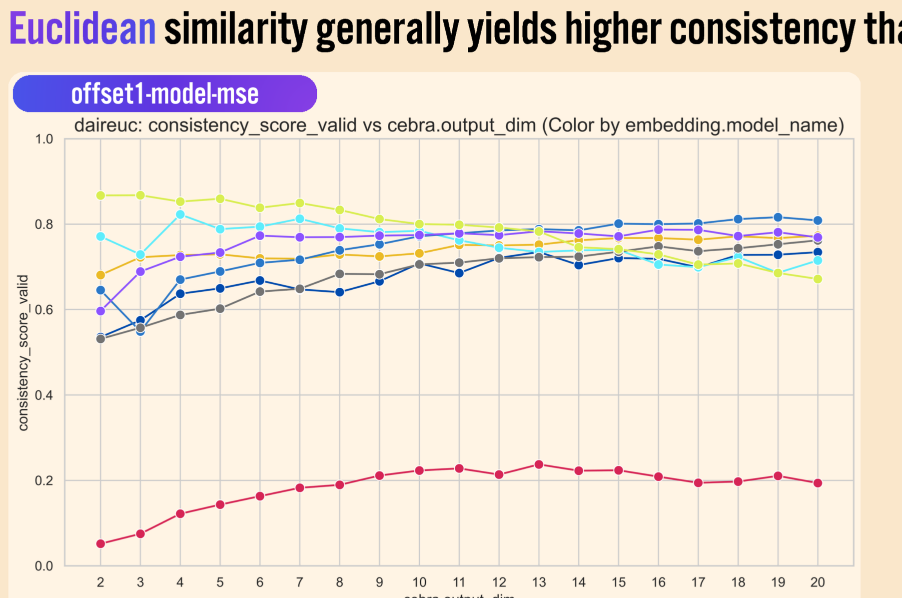
Evaluation:

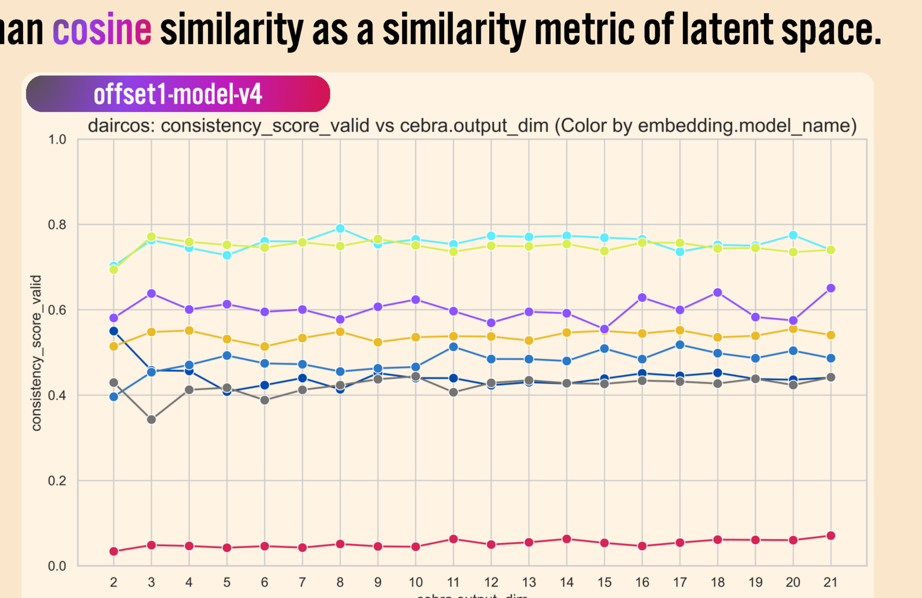
We measure Consistency across 5 random seeds using R^2 scores of linear regressions between learned embeddings.

Results

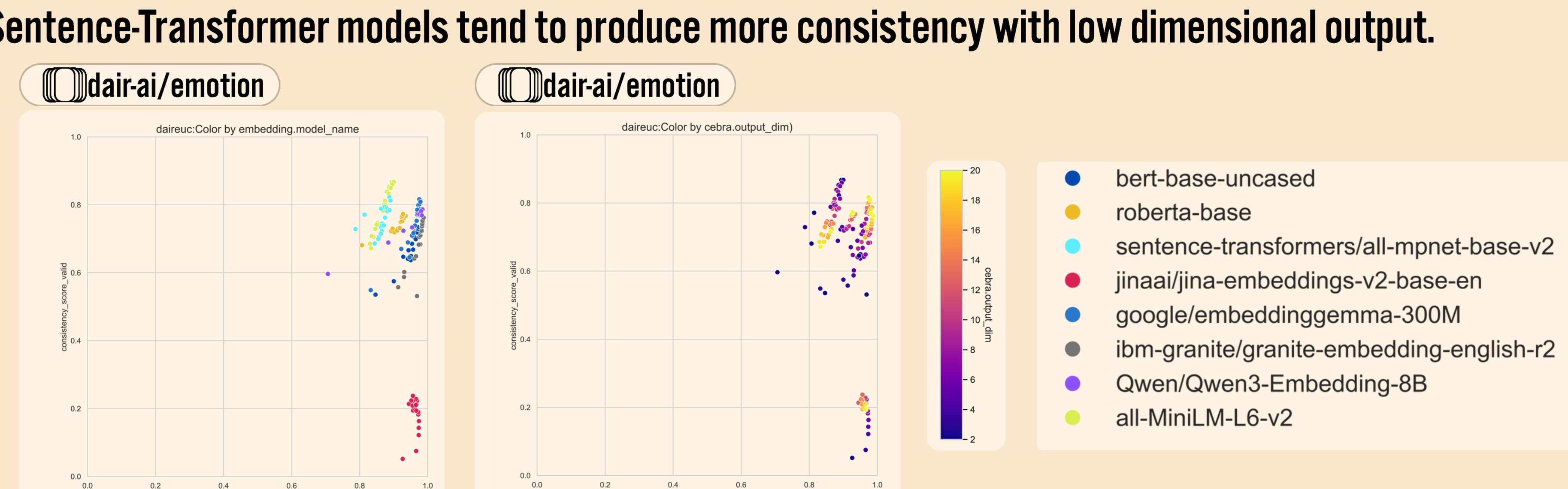
CEBRA learns low-dimensional latent structures highly consistent across random initializations

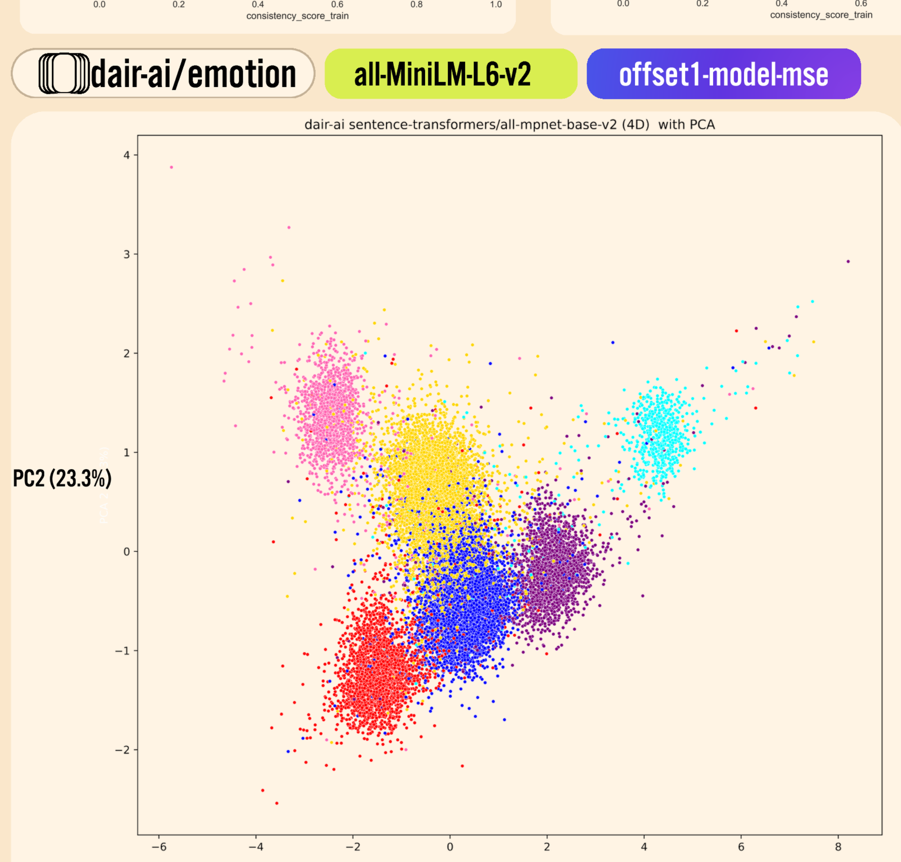
Euclidean similarity generally yields higher consistency than cosine similarity as a similarity metric of latent space.

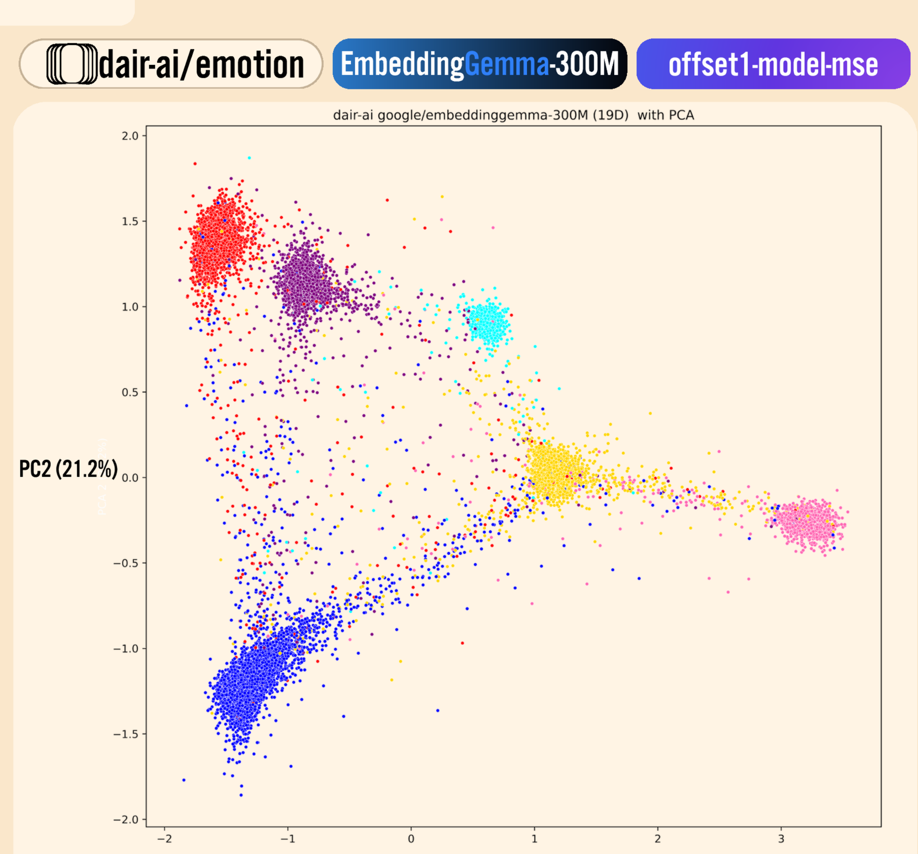




The choice of language embedding model strongly influences the the learned latent structure; Sentence-Transformer models tend to produce more consistency with low dimensional output.







Discussions

CEBRA reflect stable, label-relevant factors inherent in the data:

CEBRA produces latent embeddings with clear and well-separated cluster structures aligned with the emotion labels.

The arrangement of clusters are highly consistent :

Label-related organization may reflect a stable latent structure associated with the data-generating process.

Choice of auxiliary variables with minimal inductive bias

Can we move beyond explicit labels to achieve identifiability in a more unsupervised manner?

Interpretation of indeterminacy up to a linear transformation:

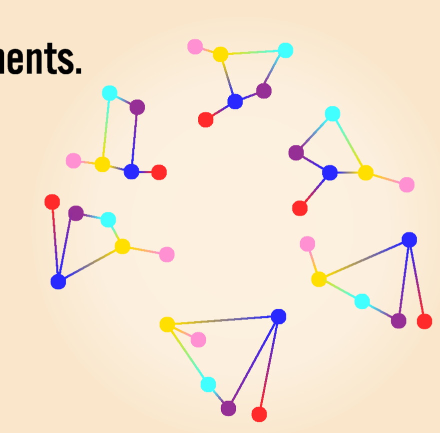
How can we define robust and meaningful metrics on the relative geometric structure.

Geometrical structure of latent space:

Analyzing distance/similarity patterns and relative geometric arrangements.

Topological structure of latent space:

Examining cluster relationships, connectivity patterns, and manifold-level organization.



Extensions to Multimodal and Neural Domains:

Future work will apply this framework to multimodal data (image,audio) and to neural data, probing whether similar identifiable structures emerge in biological brains.

Acknowledgements

This work was supported by JST, CREST Grant Number JPMJCR22P2, Japan.

References

M. Huh et al. (2024), arXiv:2405.07987. URL: https://arxiv.org/abs/2405.07987

A. Hyvärinen et al. (2019), AISTATS, PMLR. DOI: 10.48550/arXiv.1805.08651. URL: https://proceedings.mlr.press/v89/hyvarinen19a.html

S. Schneider et al. (2023), Nature. DOI: 10.1038/s41586-023-06031-6. URL: https://www.nature.com/articles/s41586-023-06031-6

H. Yamagiwa et al. (2023), EMNLP 2023. DOI: 10.18653/v1/2023.emnlp-main.283. URL: https://aclanthology.org/2023.emnlp-main.283/

R. Li et al. (2024), EMNLP 2024. DOI: 10.18653/v1/2024.emnlp-main.1065. URL: https://aclanthology.org/2024.emnlp-main.1065/

E. Saravia et al. (2018), EMNLP 2018. DOI: 10.18653/v1/D18-1404. URL: https://aclanthology.org/D18-1404/

D. Demszky et al. (2020), arXiv:2005.00547. URL: https://arxiv.org/abs/2005.00547

J. Devlin et al. (2019), NAACL-HLT 2019 (Minneapolis), DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/

Y. Liu et al. (2019), arXiv:1907.11692. URL: https://arxiv.org/abs/1907.11692

W. Wang et al. (2020), URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

G. Roeder et al (2020), arXiv:2007.00810, URL: https://arxiv.org/abs/2007.00810