Axis Tour: Word Tour Determines the Order of Axes in ICA-transformed Embeddings

Hiroaki Yamagiwa[1], Yusuke Takase[1], Hidetoshi Shimodaira[1,2]
[1]Kyoto University [2]RIKEN AIP
EMNLP 2024 Findings

Background

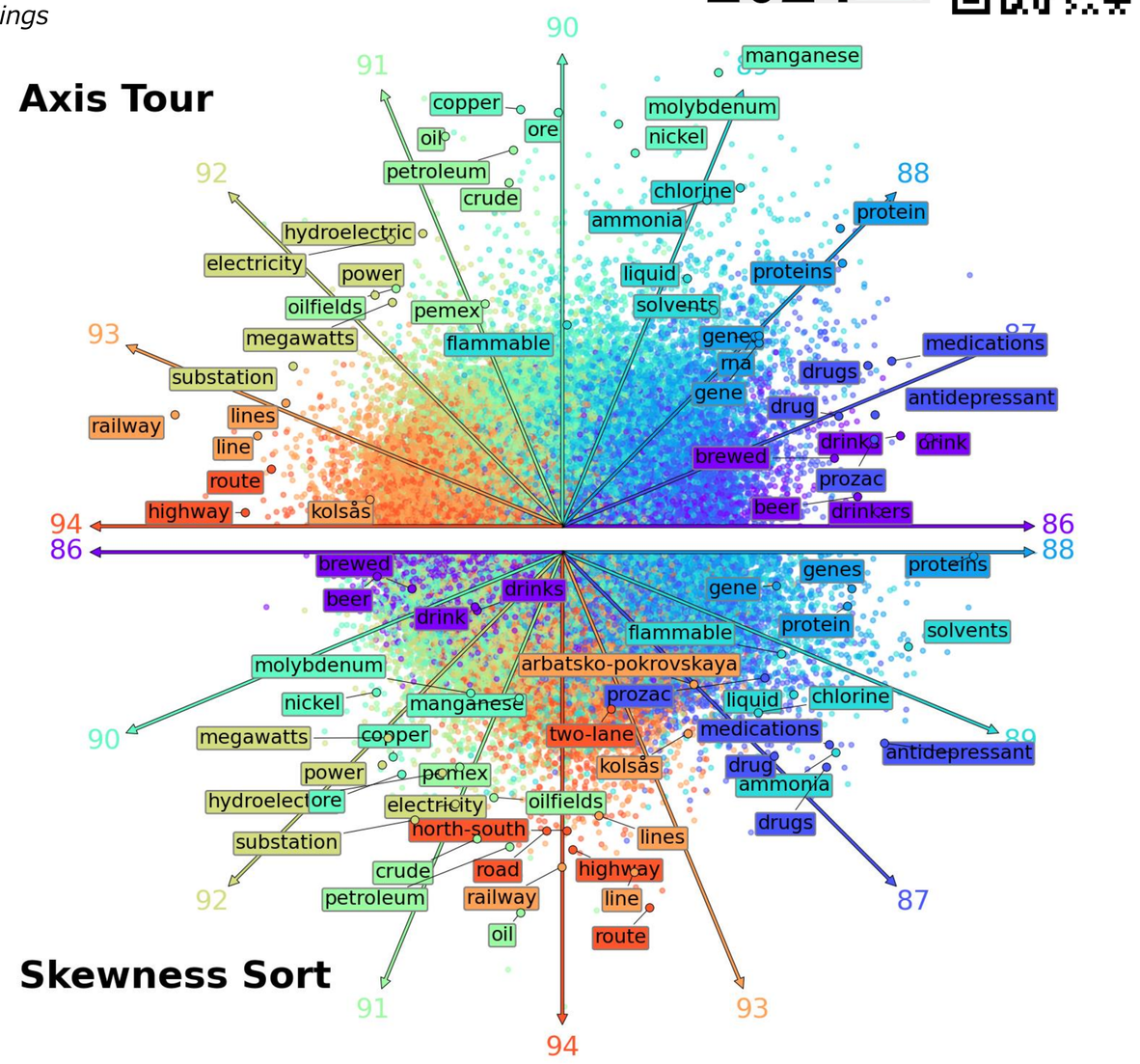
- ICA-transformed embeddings $\boldsymbol{S} \in \mathbb{R}^{n \times d}$ reveal interpretable and distinct semantic axes [1].
- However, the order of these axes is arbitrary [2].
- Word Tour [3] uses the Traveling Salesman Problem (TSP) for high-dimensional embeddings and computes 1D embeddings that represent continuous meanings.

Approach

- Define the **axis embedding**, which represents the meaning of an axis.
- The proposed **Axis Tour** uses the embeddings to perform Word Tour and determines the order of the axes in $\boldsymbol{S}$.

Results

- The meanings of the axes change continuously.
- The scatterplot of the embeddings is more interpretable.
- Merging consecutive axes to create low-dimensional embeddings leads to improved performance.



Word Tour [3]

- For the embeddings $\{\boldsymbol{x}_i\}_{i=1}^{n} \subset \mathbb{R}^d$, solve the following TSP to obtain 1D embeddings:

$$\min_{\tau \in \mathcal{P}([n])} \|\boldsymbol{x}_{\tau_1} - \boldsymbol{x}_{\tau_n}\| + \sum_{i=1}^{n-1} \|\boldsymbol{x}_{\tau_i} - \boldsymbol{x}_{\tau_{i+1}}\|.$$

- The 1D embeddings show semantic continuity.

Axis Embedding (Ours)

- Define $\widehat{\boldsymbol{S}} \in \mathbb{R}^{n \times d}$ as the normalization of $\boldsymbol{S}$, where the row vectors are given by $\hat{\boldsymbol{s}}_i = \boldsymbol{s}_i / \|\boldsymbol{s}_i\| \in \mathbb{R}^d$.
- The ℓ-th **axis embedding** is given

$$\boldsymbol{v}_\ell \coloneqq \frac{1}{k} \sum_{i \in \mathrm{Top}_k^\ell} \hat{\boldsymbol{s}}_i \in \mathbb{R}^d,$$

where Top_k^ℓ is the indices of the top k components when comparing the values along the ℓ-th axis of $\widehat{\boldsymbol{S}}$.

Axis Tour (Ours)

Formulation

- Perform Word Tour using axis embeddings with cosine distance:

$$\max_{\tau \in \mathcal{P}([d])} \cos(\boldsymbol{v}_{\tau_1}, \boldsymbol{v}_{\tau_d}) + \sum_{\ell=1}^{d-1} \cos(\boldsymbol{v}_{\tau_\ell}, \boldsymbol{v}_{\tau_{\ell+1}}).$$

- Then $\boldsymbol{T} \in \mathbb{R}^{n \times d}$ is obtained by reordering the axes of $\boldsymbol{S}$ to make their meanings. continuous

Dimensionality Reduction

- Divide the interval $[1, \ldots, d]$ into p equal parts.
- For the r-th subinterval I_r, consider the 1D projection $\boldsymbol{T}\boldsymbol{f}_r \in \mathbb{R}^n$, where $\boldsymbol{f}_r \in \mathbb{R}^d$ is a unit vector whose ℓ-th component is based on the skewness γ_ℓ of the ℓ-th axis ($\ell \in I_r$).
- Compute $\boldsymbol{T}\boldsymbol{F} \in \mathbb{R}^{n \times p}$ by $\boldsymbol{F} \coloneqq [\boldsymbol{f}_1, \cdots, \boldsymbol{f}_p] \in \mathbb{R}^{d \times p}$.

Semantic Continuity of Axes

Qualitative Observation

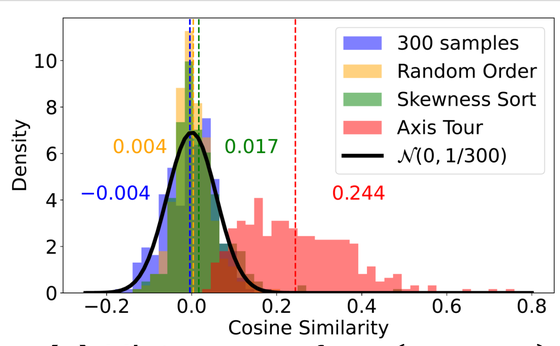
- After Axis Tour with GloVe, subintervals of the axes are extracted, and the top 5 words for each are shown.
- The semantic continuity of the axes is observed.

23	24	25	26	27	28	29	30	31
serb	russian	czech	germany	france	canada	australia	wiltshire	liga
bosnian	russia	prague	german	french	canadian	australian	shrewsbury	relegated
croatia	moscow	poland	berlin	le	ontario	queensland	lincolnshire	fc
croatian	sergei	polish	von	paris	quebec	brisbane	peterborough	f.c.
serbian	aleksandr	warsaw	cologne	du	saskatchewan	perth	croydon	serie
101	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**
pay	land	laws	court	lawsuits	charges	camp	corpses	remain
fees	property	regulations	judge	lawsuit	alleged	prison	corpse	remained
payments	lands	enacted	appellate	litigation	prosecutors	buchenwald	exhumed	stayed
payment	estate	law	appeals	suits	indicted	camps	dismembered	stubbornly
paid	bergisches	provisions	supreme	suit	convicted	inmates	bodies	stays
237	**238**	**239**	**240**	**241**	**242**	**243**	**244**	**245**
award	film	superhero	album	piano	paintings	manuscript	language	name
awards	films	marvel	albums	violin	painting	biographies	languages	names
awarded	movie	spin-off	band	cello	art	pages	pashto	surname
prize	starring	superheroes	self-titled	percussion	sculpture	book	colloquial	phrase
emmy	directed	characters	ep	orchestral	watercolor	handwritten	dialect	misspelling

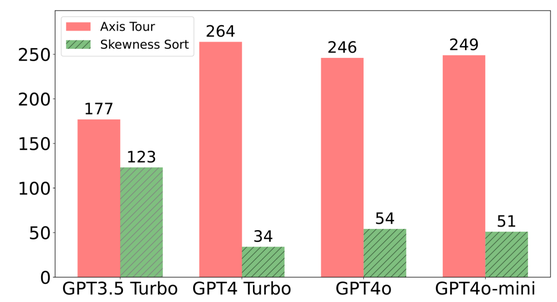
Quantitative Evaluation

Comparison of the axes of $\boldsymbol{S}$ reordered by randomness, descending skewness, and Axis Tour:

- In (a), the cosine similarities by Axis Tour deviate from 0, indicating semantic continuity between the axes.
- In (b), Axis Tour results in more pairs of semantically related adjacent axes for GPT models.



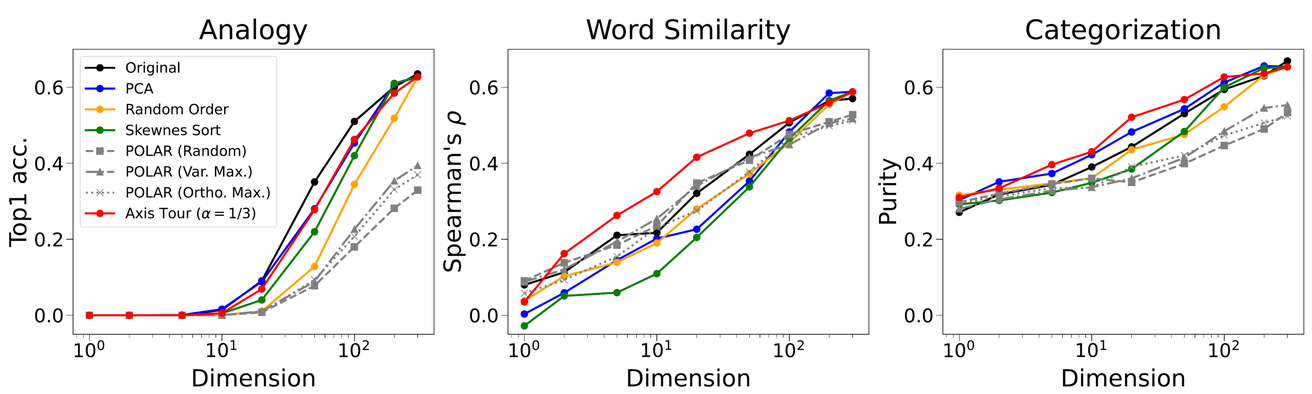
(a) Histogram of $\cos(\boldsymbol{v}_\ell, \boldsymbol{v}_{\ell+1})$



(b) Comparison of the number of related axes in the GPT models.

Dimensionality Reduction

- Comparison of low-dimensional embeddings from Axis Tour and baseline methods on analogy, word similarity, and categorization tasks.
- Axis Tour is better or comparable.



References

[1] Yamagiwa, et al. Discovering Universal Geometry in Embeddings with ICA. EMNLP 2023.
[2] Hyvärinen et al. Independent Component Analysis. Wiley 2001.
[3] Sato. Word tour: One-dimensional word embeddings via the traveling salesman problem. NAACL 2022.