Information-Theoretic Generative Clustering of Documents

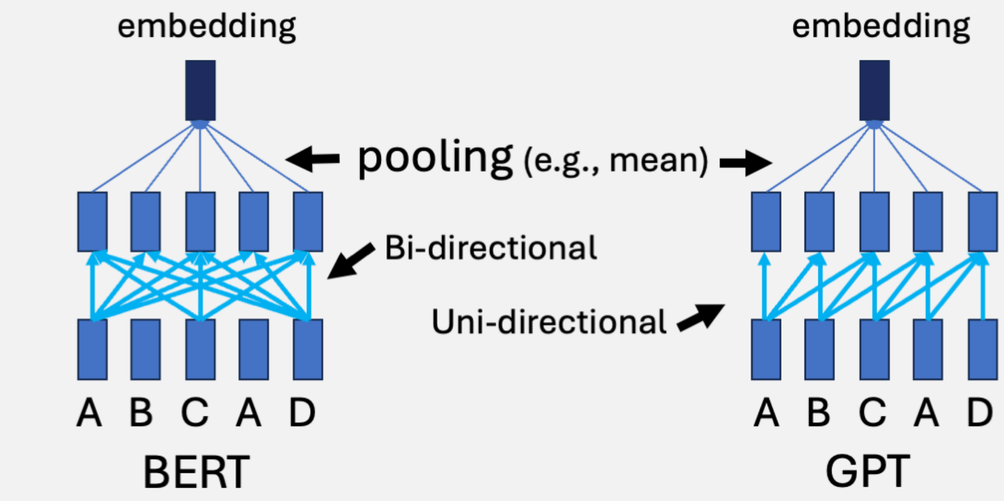
— Clustering on Language Model-Generated Sequence Space

Paper Code AAAI 2025

Xin Du, Kumiko Tanaka-Ishii
Waseda University https://ml-waseda.jp

Document Embedding: A Persistent Challenge for Generative LLMs

Generative LLM research is thriving.
However, most-widely used document embedding models still rely on non-generative, bidirectional architectures like BERT.



Why GPT embeddings underperform is well understood:

- Autoregressive: Cannot use right-to-left context
- Non-stationary: pooling is ineffective …

No effective fix exists yet.

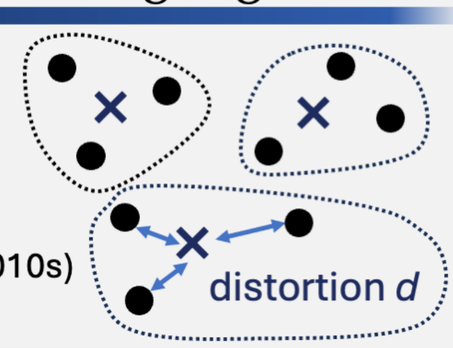
Is there an alternative to embeddings for using generative LMs in document representation?

YES! We propose a new approach for clustering.

Centroid-Based Document Clustering Algorithms

_k_-means (1970s)

d: Squared Euclidean distance
minimization of total point-centroid distortion



k-means + Document Embedding (2010s)
Bidirectional language models: ELMo, BERT, …

Information-Theoretic Clustering (2000s)
d: KL Divergence on V = vocabulary
Bag-of-Words representation of documents

$$\min_{x \mapsto k \text{ assignment}, \text{ centroids } p(Y|k)} \text{distortion} = \sum_{\text{all docs}} \mathrm{KL}\left[p(Y|X) \,\|\, p(Y|k) \right]$$

word random variable over V; doc x; centroid of cluster k

$p(Y=w|X)$ = occurring freq. of word w in document

Generative Clustering (ours)

d: KL Div. on V^∞ = word sequences, defined using a language model

$p(Y=[w_1, \ldots, w_t] \mid X)$ = generation prob. of $[w_1, \ldots, w_t]$ from X

Challenge 1: KL is intractable because of infinitely many possible $[w_1, \ldots, w_t]$
Challenge 2: Needs a computable form for centroids.

Generative Clustering: Distortion Estimation through Regularized Importance Sampling

$$\mathrm{KL}[p(Y|x) \,\|\, p(Y|k)] \approx \frac{1}{N} \sum_{i=1}^{N} \left(\frac{p(y_i|x)}{\phi(y_i)} \right)^{\alpha=0.25} \log \frac{p(y_i|x)}{p(y_i|k)}$$

regularization strength

$\mathcal{Y}=\{y_i\}$ i.i.d.$\sim\phi$

Vanilla importance sampling has $\alpha = 1$

Regularization greatly reduces variance of the estimator, enhancing clustering accuracy

Step 1. Sampling "queries" as $\mathcal{Y}$

Some prior **proposal** distribution ϕ → LLM (doc2query) → DAA, CADB, CABBDA, AAAB, … = $\mathcal{Y}$ "*queries*" (e.g., N=300)

doc2query models are pretrained on retrieval tasks

See the paper for choice of ϕ

Step 2. Estimate KL on $\mathcal{Y}$

Documents $\mathcal{X}$: ABCAD, DACBC, DACAA → LLM (doc2query) → Generation probability $\mathcal{Y}$: DAA, CADB, CABBDA, AAAB, …

p(y|x) documents p(y|k) clusters

Step 3. Solve for minimal total (estimated) KL divergence

ABCAD, DACBC, DACAA

Documents with similar p(y|x) are assigned to the same cluster.

Clustering is performed through a two-step iteration algorithm akin to k-means.

Comparison with Embedding Approaches

$$x \mapsto \left\{ \left(\frac{p(y_i|x)}{\phi(y_i)} \right)^{0.25} \right\}_{y_i \in \mathcal{Y}}$$ is the "embedding" in GC

- Embedding reformulated as a sampling problem (of $\mathcal{Y}$).
- Every dimension is interpretable (by a query text in $\mathcal{Y}$).
- Precision is controllable via size of $\mathcal{Y}$ (typically, 300).
- Potential adaptation via prompting (future work).

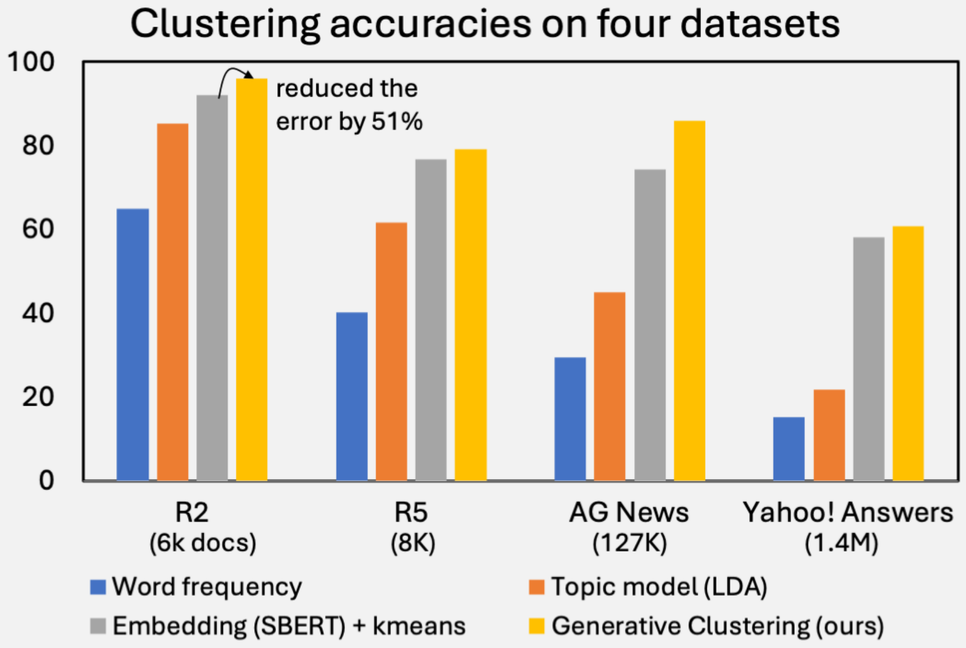
Limitations

More costly than conventional embedding approaches.

Effects of LLM choices & fine-tuning methods is yet to be explored. (we used pre-trained doc2query without fine-tuning)

Experiments

Clustering accuracies on four datasets



GC achieves the state of the art, excelling on all four datasets of varying sizes.
Showcased generative LMs' potential for clustering.