Learning Disentangled Document Representations Based on a Classical Shallow Neural Encoder

Yuro Kanada[1 [0009−0008−0382−3121]], Sumio Fujita[2 [0000−0002−1282−386X]], and Yoshiyuki Shoji[1 [0000−0002−7405−9270]]

[1] Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
kanada.yuro.21@shizuoka.ac.jp, shojiy@inf.shizuoka.ac.jp
[2] LY Corporation, Chiyoda-ku Tokyo 102–0094, Japan
sufujita@lycorp.co.jp

Abstract. This paper proposes a document embedding method designed to obtain disentangled distributed representations. The resulting representations are expected to satisfy two key criteria: independence across dimensions and semantic interpretability of each dimension. We enhanced a classic shallow neural network-based embedding model with two modifications: 1) guidance task integration, where the network is trained to perform both a simple auxiliary metadata prediction task and a surrounding term prediction task simultaneously, and 2) loss regularization for independence, where the loss function includes both prediction accuracy and the independence across dimensions (*i.e.*, the Kullback-Leibler divergence from a multivariate normal distribution). We evaluated the proposed method through both automatic and human-subject experiments using synthetic datasets and movie review texts. Experimental results show that even shallow neural networks can generate disentangled representations when dimensional independence is explicitly promoted.

Keywords: Document Embedding · Distributed Representation · doc2vec · Disentangled Representation.

1 Introduction

Driven by the growing demand for machine learning and data-driven services, embedding techniques have become increasingly central in modern data processing. Converting various data, such as documents, into compact, dense vectors is a common practice for inputting data into deep neural networks. Various methods have been proposed for this purpose: Word2vec [16] vectorizes words, doc2vec [14] vectorizes documents, node2vec [8] vectorizes graphs, and CLIP [19] vectorizes images.

Among the various approaches to embedding, shallow neural network-based methods are traditionally used, particularly for text embedding. Models such as word2vec and doc2vec demonstrated that even simple architectures can capture meaningful semantic relationships by predicting the surrounding context

of words or documents. In recent years, the emergence of large language models (LLMs) has enabled the generation of higher-dimensional embeddings. For instance, methods based on Transformer architectures are increasingly utilizing encoders such as BERT and SentenceBERT to embed textual data. However, the embedding methods based on shallow neural networks continue to play a critical role in practical applications due to their efficiency, simplicity, and effectiveness in various real-world settings.

Although embedding-based distributed representations are practical tools in machine learning, they often lack interpretability for human users. Semantic meanings are typically encoded across multiple dimensions, and individual dimensions may capture a mixture of unrelated concepts. As a result, while such representations can preserve semantic relationships (*e.g.*, placing "dog" closer to "cat" than to "snake"), they do not offer clear, disentangled dimensions that correspond to interpretable attributes.

This limitation leads to two significant problems. First, embedding vectors are difficult to interpret from a human perspective. Given a document embedding, it is generally unclear which dimensions correspond to which properties of the document. This ambiguity hinders the ability to compare documents meaningfully or to understand how semantic structure is encoded during training. While research on interpretability for large language models (LLMs) is advancing, even the internal mechanisms of shallow models remain poorly understood.

Second, conventional embeddings do not support fine-grained semantic operations. For example, retrieving documents that are more humorous or more emotionally negative than a given document is infeasible without knowing which dimensions reflect these attributes. This lack of alignment with interpretable semantics makes it challenging to perform a similarity search based on specific conceptual criteria. Supporting such targeted operations could improve both the interpretability and the downstream effectiveness of models that use these embeddings.

To address these issues, there is growing interest in disentangled representations, which capture each dimension as an independent and semantically meaningful factor of variation. Disentanglement has recently gained attention in the context of explainable AI, particularly in the design of encoder architectures for generative models. For instance, Zhang *et al.* [24] introduced a disentangled architecture for Transformer-based encoders, while Colombo *et al.* [6] proposed regularizers to promote disentanglement. Such efforts toward disentanglement have been extensively explored in the field of image generation, where a variety of methods, including β-VAE [11], have been proposed. In this work, we explore the application of disentanglement techniques to classical text encoders constructed with shallow neural networks.

This study proposes a method for learning document embeddings in which each vector dimension is both independent and semantically meaningful. The method is designed to address the two limitations discussed above. To this end, we extend a classical embedding approach based on a shallow neural network trained on a context word prediction task.

We modified the traditional doc2vec-based network with two improvements in its training process. First, a guidance task is incorporated during training, where the model is required to predict metadata in addition to surrounding words. Second, during parameter updates, the model assesses the degree to which the batch of vectors approximates a multivariate standard normal distribution, and this evaluation is incorporated into the loss function.

The document embeddings obtained through this training procedure are referred to in this paper as Disentangled Representations. In this paper, a Disentangled Representation is defined as a vector that satisfies the following three conditions. First, the representation reflects real-world relational distances in the latent space. Second, each dimension of the vector is statistically independent. Third, each dimension carries an interpretable semantic meaning.

The main contribution of this study is to demonstrate that dimensional disentanglement techniques, which are commonly used in the image generation domain, are also effective when applied to simple shallow neural encoders; Specifically, during training, the KL divergence between the encoded vector and a multivariate normal distribution is computed and incorporated into the loss function. Importantly, our aim is not to achieve highly optimized or state-of-the-art Disentangled Representations (such as those obtained using models like BERT), but rather to demonstrate that meaningful disentanglement can be achieved effectively with low-cost shallow neural networks.

2 Related Work

This study focuses on the training process of existing embedding methods. It introduces modifications that enable the acquisition of Disentangled Representations, in which each dimension is both independent and semantically meaningful. In this respect, the study is closely related to major embedding approaches. It is also relevant to the field of Disentangled Representation Learning (DRL). This is because the aim is to reduce dependency across vector dimensions and to obtain representations that can be interpreted more easily.

2.1 Document Embedding Methods

Word and document embeddings are foundational for representing textual data as continuous vectors in machine learning models. word2vec [16] captures semantic relationships by predicting surrounding words from a target word or vice versa, exploiting co-occurrence patterns. Doc2vec [14], an extension, introduces document identifiers to learn document-level embeddings within the same vector space, using inference tasks to reflect textual semantics. However, these classical models primarily focus on surface-level co-occurrence and often overlook external document properties, such as metadata (*e.g.*, genre, sentiment, or source), which convey crucial semantic distinctions. This limitation motivates the integration of additional semantic signals to enhance embedding interpretability and task relevance.

2.2 Disentangled Representation Learning

Disentangled Representation Learning (DRL) aims to separate the underlying generative factors of observed data into distinct, interpretable components [21], where each dimension captures a specific factor of variation [1].

Early work in this area focused on image processing using VAEs [12] with multivariate normal priors. Methods like β-VAE [11], DIP-VAE [13], and TC-VAE [4] encourage independence among latent dimensions by adjusting the KL divergence weight or adding correlation-penalizing regularization. Such disentangled representations enhance interpretability and support tasks like retrieval and recommendation [2, 7].

Beyond VAEs, InfoGAN [5] extends GANs by introducing a latent code and maximizing the mutual information between the code and generated outputs. Other approaches achieve disentanglement through orthogonality constraints [20, 22, 23] or vector space rotations [18] to reduce overlap among latent dimensions.

DRL has also been applied to natural language processing, with examples like DeBERTa [9] introducing disentangled attention and ABAE [10] encouraging topic-level separation. However, these prior studies primarily focus on large-scale or image-based models. In contrast, the present study investigates disentanglement within classical document embedding frameworks based on shallow neural networks, expecting each document vector dimension to represent an independent and semantically meaningful feature, enforced by regularization toward a multivariate normal distribution.

3 Shallow Neural Network-based Encoder for Document Disentangled Representation

This section presents the proposed encoder for learning disentangled representations of documents. The method extends a classical context-based embedding framework by incorporating a guidance task that leverages document metadata and by modifying the loss function to encourage proximity to a multivariate standard normal distribution. The following subsections outline the overall architecture, the base encoder utilizing context word prediction, the metadata-based guidance task, the independence mechanism grounded in multivariate Gaussian properties, and the final objective function employed during training.

3.1 Algorithm Overview

This subsection provides an overview of the algorithm used to learn disentangled representations of documents. Figure 1 shows the overall structure of the proposed method's training phase. The encoder's goal is to optimize a weight matrix whose rows correspond to document embeddings. The design follows principles from classical embedding models such as word2vec and doc2vec.

The proposed model extends this framework to document-level embeddings. During training, the input consists of a document ID and a one-hot vector representing a word in the document. The model outputs the probability of vocabulary words appearing in context and metadata labels associated with the

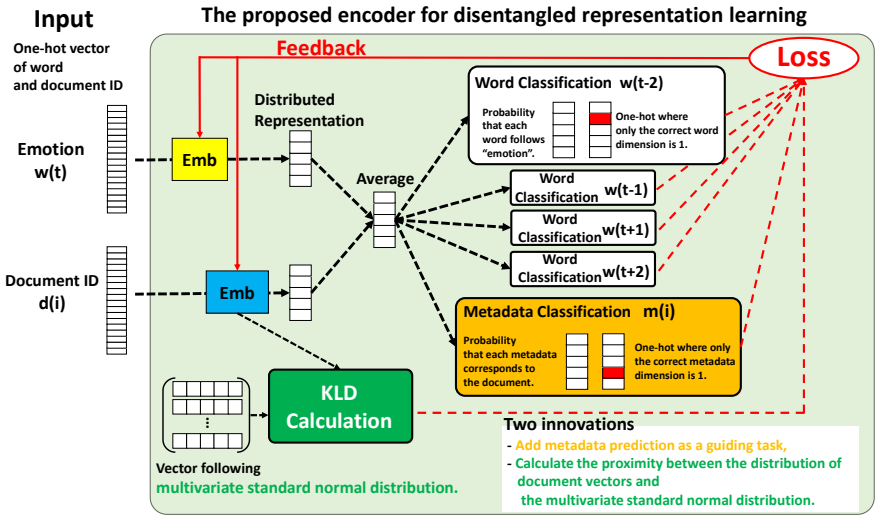

Fig. 1. The training flow of the proposed method. A batch of one-hot vectors representing document IDs and in-document words is input, from which the corresponding embeddings are extracted via a weight matrix. Using the mean of these embeddings, the model solves two tasks: a word classification task that predicts the surrounding words of each input word, and a guidance task that predicts metadata associated with each document. In parallel, the model computes KL divergence between the distribution of document embeddings and the multivariate standard normal distribution. Based on the results of both learning tasks and the computed KL divergence, the weight matrix is updated.

document ID. Training is performed over mini-batches of input pairs, where each document ID appears only once. Word vectors are embedded via the input-side weight matrix, averaged, and used to solve context word prediction and metadata prediction tasks jointly.

In parallel, the model calculates the KL divergence between the distribution of document embeddings in the batch and a multivariate standard normal distribution. The overall loss combines losses from both prediction tasks and the KL divergence regularization. Gradients from this total loss update the input-side weight matrix. Through this process, the encoder learns document embeddings in which each dimension is encouraged to be both statistically independent and semantically meaningful, resulting in disentangled representations.

3.2 Base Model

As the base model, we adopted an encoder that combines PV-DM and PV-DBOW from doc2vec to obtain effective document embeddings [14]. It predicts surrounding words given a document ID and a word from the document.

During training, it minimizes the loss using Negative Sampling [17]:

$$\mathcal{L}_{base}(i) = -\log p(W_{ci} \mid D_i) - \sum_{W_v \in V_{\text{neg}}} \log(1 - p(W_v \mid D_i)), \tag{1}$$

where D_i denotes the input (document ID and in-document word), W_{ci} is a positive context word, and W_v represents negative samples.

3.3 Multi-Task Learning with Metadata Prediction

As the first key design component, this study introduces a guidance task to promote semantic disentanglement in document embeddings. This aims to encourage each vector dimension to represent a distinct and interpretable feature. To this end, the model is trained not only on text-based context prediction but also on an auxiliary classification task predicting document metadata. Metadata, encoding explicit and semantically meaningful attributes such as genre or sentiment, directly provides interpretable axes for the embedding space. By aligning specific dimensions with these known semantic categories, this guidance task explicitly promotes the interpretability of individual dimensions and facilitates their disentanglement. This joint optimization enables the model to embed such information, potentially concentrating related information into specific dimensions and thereby promoting greater independence across the representation.

The encoder augmented with this guidance task, referred to as the guidance task model, maintains the base model's input format: a document ID and an in-document word. It outputs probability distributions over nearby context words and associated metadata labels. This joint objective encourages embeddings to reflect both the document's lexical content and its external semantic attributes.

The guidance loss function $\mathcal{L}_{guide}(i)$ is defined as:

$$\mathcal{L}_{guide}(i) = -\log p(W_{ci} \mid D_i) - \sum_{W_v \in V_{\text{neg}}} \log(1 - p(W_v \mid D_i)) - \log p(M_{gi} \mid D_i) - \sum_{M_s \in S_{\text{neg}}} \log(1 - p(M_s \mid D_i)), \tag{2}$$

where D_i represents the input document, W_{ci} is a positive context word, W_v represents negative samples, M_{gi} denotes a positive metadata label, and M_s denotes a negative metadata sample.

Minimizing $\mathcal{L}_{guide}(i)$ encourages high probabilities for relevant context words and metadata labels, while suppressing irrelevant ones. Posterior probabilities are computed using the sigmoid function, following the base model's formulation. Specifically, averaged document and word embeddings are dot-producted with target embeddings (word or metadata), then passed through a sigmoid for prediction probability. Given this formulation, the guidance loss $\mathcal{L}_{guide}(i)$ can be written as:

$$\begin{aligned} \mathcal{L}_{guide}(i) = &-\log(\sigma((\frac{V_{di} + V_{ti}}{2}) \cdot V_{ci})) \quad - \sum_{V_v \in V_{\text{neg}}} \log(1 - \sigma((\frac{V_{di} + V_{ti}}{2}) \cdot V_v)) \\ &-\log(\sigma((\frac{V_{di} + V_{ti}}{2}) \cdot V_{gi})) \quad - \sum_{V_s \in S_{\text{neg}}} \log(1 - \sigma((\frac{V_{di} + V_{ti}}{2}) \cdot V_s)), \end{aligned} \tag{3}$$

where V_{di} and V_{ti} are document and word embeddings, V_{gi} is a metadata label embedding, and V_s is a negative metadata sample.

3.4 Loss Function Considering Dimensional Independence

To encourage each dimension of the document embeddings to represent an independent semantic feature, we introduce a regularization term that aligns the embeddings' distribution with a multivariate standard normal distribution. This is achieved by computing the KL divergence between the empirical distribution of document vectors in each batch and the ideal distribution, thereby promoting statistical independence across dimensions.

A multivariate standard normal distribution has a zero mean vector and an identity covariance matrix, where all variables are uncorrelated and follow individual standard normal distributions. This property is leveraged in the β-VAE framework, weighting the KL divergence term to encourage the latent representation to approximate this distribution, thus promoting each dimension to capture a distinct and independent generative factor [15].

To ensure reliable covariance matrix estimation, each training batch is constructed with unique document IDs. Assuming each dimension of the document embeddings within a batch follows a multivariate normal distribution, the batch mean and covariance matrix estimate the empirical distribution $\mathcal{N}(\mu, \Sigma)$ of the document vectors. The target distribution is a multivariate standard normal distribution, $\mathcal{N}(0, \mathbf{I})$, with mutually independent and identically distributed dimensions. The KL divergence between the empirical and target distributions measures the deviation from this ideal structure, incorporated into the training loss as a regularization term to promote disentangled representations.

To compute the KL divergence between the empirical distribution of document embeddings and the multivariate standard normal distribution, we adopt the closed-form expression for the divergence between two multivariate Gaussian distributions:

$$D_{KL}\left(\mathcal{N}(\mu,\Sigma)\,\|\,\mathcal{N}(0,\mathbf{I})\right) = \frac{1}{2}\mathrm{tr}(\mathbf{I}^{-1}\cdot\Sigma) \quad + \frac{1}{2}(0-\mu)^{\top}\mathbf{I}^{-1}(0-\mu) \quad + \frac{1}{2}\log\left(\frac{\det(\mathbf{I})}{\det(\Sigma)}\right) - \frac{1}{2}d, \tag{4}$$

where μ is the d-dimensional mean vector, and Σ is the $d \times d$ covariance matrix of the batch.

In practice, the third term involving the log-determinant can introduce numerical instability, especially when Σ is close to singular. To mitigate this, we follow a common VAE approach, simplifying computation by assuming statistically independent embedding dimensions. Under this assumption, the KL divergence is calculated independently for each dimension by comparing its marginal distribution to the standard normal distribution. Letting the j-th dimension follow $\mathcal{N}(\mu_j, \sigma_j)$ and the target be $\mathcal{N}(0, 1)$, the KL divergence simplifies to:

$$D_{KL}\left(\mathcal{N}(\mu,\sigma)\|\mathcal{N}(0,1)\right) = -\frac{1}{2}\sum_{j=1}^{d}\left(1+\log(\sigma_j^2)-\mu_j^2-\sigma_j^2\right). \tag{5}$$

This dimension-wise computation provides a numerically stable and interpretable regularization term, encouraging each embedding dimension to follow a standard normal distribution and thereby promoting statistical independence.

3.5 Objective Function for Training

This subsection defines the overall training objective that integrates the components described above. The proposed encoder is trained using mini-batch processing, where each batch consists of document IDs and their associated one-hot word vectors.

Let B be the batch size, and let $b = 0, 1, 2, \ldots, \frac{K}{B}$ denote the batch index. For each batch, the training objective $\mathcal{L}_{dre}(b)$ combines the base loss, the guidance task loss, and the KL divergence-based regularization, to form the complete objective function for learning disentangled document embeddings. The final training objective is defined as:

$$\begin{aligned} \mathcal{L}_{dre}(b) = &- \log p(W_{cb} \mid D_b) - \sum_{W_v \in V_{\text{neg}}} \log(1 - p(W_v \mid D_b)) \\ &- \log p(M_{gb} \mid D_b) - \sum_{M_s \in S_{\text{neg}}} \log(1 - p(M_s \mid D_b)) \\ &+ \alpha \sum_{i=0}^{d} D_{KL}\left(\mathcal{N}(\mu_{bi}, \sigma_{bi}) \| \mathcal{N}(0, 1)\right), \end{aligned} \tag{6}$$

where α is a hyperparameter that balances the influence of the KL divergence term. This term promotes statistical independence across embedding dimensions by encouraging their distribution to approximate the standard normal distribution.

4 Evaluation

This section evaluates the effectiveness of the proposed method using two datasets. To assess the representational quality of the learned document embeddings, automatic evaluation was first conducted on synthetic data. In addition, both automatic and human evaluations were conducted on real-world data to assess the method's applicability in practical settings.

4.1 Datasets

Two datasets were used for evaluation: a synthetic dataset and a real-world movie review dataset. We selected movie reviews as real-world data because individuals often describe a single item (e.g., a film) from diverse perspectives, and the domain provides rich metadata such as genres and ratings.

Synthetic Dataset: The synthetic dataset consisted of 10,000 documents, each containing 20 word tokens, assigned to one of 20 distinct topics (A – T). Tokens 'a' – 't' were topic-related; 'u' – 'z' and '1' – '4' were topic-irrelevant. Each document contained one-third topic-specific tokens, one-third randomly sampled topic-related tokens, and one-third topic-irrelevant tokens. This structure simulates real-world variability, ensuring documents focus on a specific topic amid noise.

Movie Review Dataset: For real-world evaluation, the IMDb Review Dataset, ebD[3], was used. We selected 50,000 reviews from 1,000 movies, each with at

[3] IMDb Review Dataset - ebD: `https://www.kaggle.com/datasets/ebiswas/imdb-review-dataset`

least 50 reviews of 120 words or more, to minimize document length bias. Genre metadata was manually collected as it was not included in the original dataset. After filtering words occurring five times or fewer, the training corpus contained 7,581,077 tokens and a vocabulary of 29,543 unique words.

4.2 Implementation Details

The proposed model was implemented in Python, utilizing CuPy[4] for GPU acceleration. Training involved mini-batch processing with a batch size of 800 and a mini-batch size of 700. The embedding dimension was set to 50, and the number of negative samples was 5. The Sigmoid Annealing Scheduler (SAS) [3] was employed to control the influence of the KL divergence term, gradually increasing its weight after initial semantic learning, with a maximum KL weight of 1. Training was performed over 10 epochs with a window size of 5.

4.3 Baseline Methods

An ablation study was conducted to evaluate the contributions of the guidance task and KL divergence regularization. The document ID embeddings obtained under different configurations were compared on both the synthetic dataset and the movie review dataset.

For **Synthetic Dataset**, the following three types of document vectors were evaluated:

- **Random**: embeddings immediately after initialization, without training,
- **Baseline**: embeddings learned by the base model using only the context word prediction task, and
- **+KLD**: embeddings learned by augmenting the baseline model with a KL divergence regularization term, encouraging alignment with a multivariate standard normal distribution.

For **Movie Review Dataset**, we prepared four configurations as:

- **Proposed**: the proposed method learned with both the guidance task and KL divergence regularization,
- **W/O KLD**: guidance task included, but KL divergence regularization excluded,
- **W/O Guidance**: KL divergence included, but the guidance task excluded, and
- **Baseline**: context word prediction only (base model).

Each configuration isolates the effect of one or both proposed components, enabling the analysis of their individual and combined impact on the quality of the learned representations.

[4] CuPy: `https://cupy.dev/`

Table 1. KL divergence between the distribution of document ID vectors and the multivariate standard normal distribution, along with average variance and total entropy across dimensions

Method	KLD	Avg. Variance	Entropy
Random	36.15	0.01	33.35
Baseline	10.81	0.16	60.43
+KLD	0.19	0.41	54.16

Table 2. Average variance in the number of documents per topic for each dimension, and average number of zero-topic dimensions

Method	Avg. variance	Avg. # Zero-topic
Random	5.16	0.15
Baseline	44.62	7.80
+KLD	104.40	10.30

4.4 Experiment 1: Evaluation on Synthetic Data

To assess the representational properties of the proposed encoder, we conducted a series of evaluations on the synthetic dataset. Three types of document ID embeddings were compared: **Random**, **Baseline**, and **+KLD**. The first evaluation measured the KL divergence between the distribution of learned embeddings and the multivariate standard normal distribution using Equation (4). Second, the independence and interpretability of embedding dimensions were assessed using average variance across dimensions and total information entropy. A higher variance suggests distinct information per dimension [18], while higher entropy indicates greater overall expressiveness. Third, we examined topic-specific information capture by analyzing topic distributions among the top 100 documents with the highest values for each dimension, calculating variance in document counts across topics and the number of zero-topic dimensions. These metrics were averaged across all dimensions to assess disentanglement.

Table 1 summarizes the KL divergence, average variance, and total entropy. Incorporating the KL divergence term significantly reduced the divergence to the target distribution and increased the average variance per dimension, indicating greater independence and expressiveness. While entropy decreased slightly compared to the baseline, it remained high, suggesting that overall information content was preserved.

Table 2 reports topic-level focus per dimension. The model with KL divergence regularization showed substantially higher variance in topic distribution and more dimensions with zero-topic assignments, indicating stronger disentanglement with individual dimensions specializing in distinct topics.

4.5 Experiment 2: Applicability on Real-World Data

The applicability of the learned document ID embeddings was evaluated using real-world movie review data, treating movie reviews as documents and

movie IDs as document IDs for training. Four configurations were compared: **Proposed**, **W/O KLD**, **W/O Guidance**, and **Baseline** (as defined in Section 4.3).

We first examined whether the learned vectors exhibited similar statistical properties to those in the synthetic dataset. Following the procedure in Section 4.4, we computed KL-divergence between movie ID vector distributions and the multivariate standard normal distribution (Equation 4), along with average variance and total information entropy per dimension. Figure 2 shows the distribution of information entropy across dimensions. Compared to the baseline, the W/O KLD model exhibited greater variability, while the Proposed method and W/O Guidance model achieved more uniform entropy distributions, suggesting a better balance in information allocation.

The evaluation targeted three key criteria for disentangled representations:

- **Capturing real-world distance relationships:** Cosine similarities between query and training movies were computed. Tables 3 and 4 list the top three similar movies for *Iron Man II* and *Bohemian Rhapsody*. All methods captured basic semantic relationships. Notably, the Proposed method retrieved third-ranked movies more semantically aligned with the query compared to the baseline, especially for genre-specific or thematic attributes.
- **Independence across dimensions:** The degree of independence was evaluated by computing pairwise cosine similarities between the column vectors of the document embedding matrix. Figure 3 illustrates that both the Proposed method and the W/O Guidance model exhibited a higher concentration of near-orthogonal dimension pairs (i.e., cosine similarities between -0.1 and 0.1), indicating enhanced independence. In contrast, the Baseline and W/O KLD models showed more moderate inter-dimensional correlations.
- **Interpretability of dimensions (human evaluation):** A human subject experiment assessed whether individual dimensions encoded interpretable information. Movies were ranked by embedding values per dimension, and top-ranked/lowest-ranked movies were selected for semantic similarity ratings on a four-point scale. Results showed no substantial difference across methods in the number of dimensions where both participants consistently rated the top movie higher than the lowest-ranked one (Proposed: 22, Baseline: 24). This suggests comparable semantic coherence within individual dimensions. Quantitative analysis further indicated the Proposed method yielded the largest mean difference in cosine similarity between middle-ranked and top/bottom-ranked movies (Proposed: 0.63, Baseline: 0.45) , implying better semantic alignment of top-ranked movies with their group.Although the number of participants was limited to two, the inter-annotator agreement was moderately high, with a Krippendorff's α of 0.63.

5 Discussion

This section discusses evaluation results from both synthetic and movie review datasets. First, on synthetic data, adding the KL divergence term promoted di-

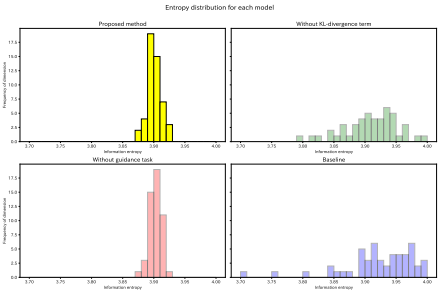

Fig. 2. Distribution of information entropy across dimensions of movie ID embeddings.

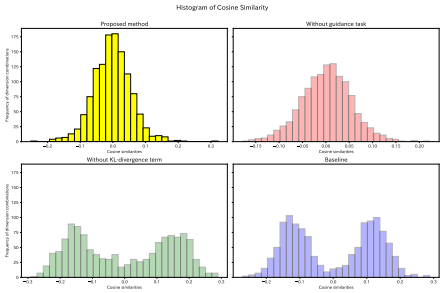

Fig. 3. Distribution of cosine similarities between dimensions in movie ID embeddings.

Table 3. Top three most similar movies to the query movie "Iron Man II"

Rank	Proposed	W/O KLD	W/O Guidance	Baseline
1st	Iron Man Three	Iron Man Three	Iron Man Three	Iron Man Three
2nd	Avengers: Age of Ultron	Spider-Man: Homecoming	Avengers: Age of Ultron	Avengers: Age of Ultron
3rd	Captain America: Civil War	Captain Marvel	Captain America: Civil War	Captain America: The Winter Soldier

mensional independence, comparable to β-VAE, even with a simple neural network. Analysis of the average variance across dimensions revealed that the KL divergence term enhances the interpretability of individual dimensions. Moreover, comparison of the total information entropy indicated that the KL divergence term constrains the amount of information carried by each dimension, suggesting a trade-off between independence and capacity. Evaluation of topic distributions across dimensions further showed that including the KL divergence term causes each dimension to focus on a smaller set of topics. These findings suggest that promoting dimensional independence via KL-regularization encourages each dimension to specialize in representing a single topic, resulting in a more compact and semantically focused embedding space.

Next, we discuss the results of the vector usability evaluation conducted on the movie review dataset. The distribution of information entropy across dimensions revealed that, compared to the baseline, the model without the KL divergence term exhibited several dimensions with notably high information content. This pattern suggests that the guidance task enabled certain dimensions to capture genre-related metadata. At the same time, dimensions with low entropy were also observed, possibly due to the concentration of metadata in only a few

Table 4. Top three most similar movies to the query movie "Bohemian Rhapsody"

Rank	Proposed	W/O KLD	W/O Guidance	Baseline
1st	Rocketman (I)	Rocketman (I)	Rocketman (I)	Rocketman (I)
2nd	Walk the Line	Walk the Line	Walk the Line	Walk the Line
3rd	The Greatest Showman	The Pianist	Yesterday (III)	First Man

dimensions, which reduced the informational content carried by others. These findings indicate that the guidance task facilitates the allocation of semantic attributes—such as genre—into specific dimensions of the embedding space.

In addition, both the proposed method and the model without the guidance task, each incorporating the KL divergence term, produced narrower entropy distributions than the baseline. This suggests that KL divergence-based regularization promotes more uniform information allocation across dimensions by encouraging statistical independence.

Case studies in Tables 3 and 4 indicate that all methods successfully captured real-world relationships among movies based on review data. Movies sharing representative features with the query consistently appeared among the top-ranked results. Notably, the inclusion of the KL divergence term did not compromise the expressive capacity of the embeddings, indicating that the proposed method meets the first requirement for a Disentangled Representation. Analysis of cosine similarities between embedding dimensions further showed that the KL divergence term promotes orthogonality, encouraging each dimension to encode a distinct semantic feature. This indicates that the second requirement—independence across dimensions—is also met.

Finally, the human subject experiment revealed no substantial difference across methods in the number of dimensions for which both participants rated the top movie higher than the lowest-ranked one. The consistency of responses between participants suggests that semantic coherence within individual dimensions was comparable across models, including the proposed method.

Collectively, the findings demonstrate that KL divergence-based disentanglement, initially developed for models such as β-VAE, is also effective in a simple neural network setting. The proposed method produces document embeddings that maintain representational performance while promoting independence across dimensions. However, the comparison between human and quantitative evaluations suggests a potential gap between machine-learned structure and human-perceived semantics. Although dimensions may exhibit statistical disentanglement, they do not always align with intuitive semantic categories.

6 Conclusion and Future Work

This study proposed an encoder for learning Disentangled Representations, where each dimension of a document embedding independently captures a distinct semantic aspect of the document. The method extends the doc2vec-based context prediction model by incorporating a metadata-guided auxiliary task and a KL divergence regularization term that encourages similarity to a multivariate standard normal distribution.

Experiments on a synthetic dataset showed that the KL divergence term promotes dimensional independence, allowing individual dimensions to encode information related to latent topics. The effect was particularly pronounced when input documents were of uniform length.

On real-world movie review data, the guidance task was shown to effectively embed metadata-related features into specific dimensions. The resulting repre-

sentations satisfied two of the three core criteria for Disentangled Representations: semantic meaningfulness and independence across dimensions.

In general, vector representations with independent dimensions are known to improve the performance of models such as Support Vector Machines and Naive Bayes. As future work, we plan to explore the broader applicability of disentangled embeddings in downstream tasks, while further improving their expressive capacity and interpretability.

7 Acknowledgments

This work was partially supported by JSPS KAKENHI under Grant Numbers 25K03229, 25K03228, and 24K03228.

References

1. Bengio, Y.: Deep learning of representations: Looking forward. In: Proceedings of International Conference on Statistical Language and Speech Processing (2013)
2. Bhattacharya, G., Kilari, N., Gubbi, J., V., B.L., Pal, A., P., B.: Datrnet: Disentangling fashion attribute embedding for substitute item retrieval. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2282–2286 (2022)
3. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. pp. 10–21 (2016)
4. Chen, R.T.Q., Li, X., Grosse, R., Duvenaud, D.: Isolating sources of disentanglement in vaes. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 2615–2625 (2018)
5. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 2180–2188 (2016)
6. Colombo, P., Staerman, G., Noiry, N., Piantanida, P.: Learning disentangled textual representations via statistical measures of similarity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2614–2630 (2022)
7. Faggioli, G., Ferro, N., Perego, R., Tonellotto, N.: Dimension importance estimation for dense information retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1318–1328 (2024)
8. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 855–864 (2016)
9. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. In: Proceedings of International Conference on Learning Representations (2021)
10. He, R., Lee, W.S., Ng, H.T., Dahlmeier, D.: An unsupervised neural attention model for aspect extraction. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 388–397 (2017)

11. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: Proceedings of International Conference on Learning Representations (2017)
12. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: Proceedings of 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014)
13. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. In: Proceedings of International Conference on Learning Representations (2018)
14. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. pp. 1188–1196 (2014)
15. Liu, X., Wang, J.: Latentvis: Investigating and comparing variational auto-encoders via their latent space. In: Proceedings of International Conference on Information and Knowledge Management (2020)
16. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of International Conference on Learning Representations (2013)
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119 (2013)
18. Park, S., Bak, J., Oh, A.: Rotated word vector representations and their interpretability. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 401–411 (2017)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning (2021)
20. Tokmakov, P., Wang, Y.X., Hebert, M.: Learning compositional representations for few-shot recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
21. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled representation learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9677–9696 (2024)
22. Wei, Y., Shi, Y., Liu, X., Ji, Z., Gao, Y., Wu, Z., Zuo, W.: Orthogonal jacobian regularization for unsupervised disentanglement in image generation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6701–6710 (2021)
23. Zhang, H., Yu, H., Yan, Y., Wang, R.: Gated domain-invariant feature disentanglement for domain generalizable object detection (2022)
24. Zhang, X., van de Meent, J.W., Wallace, B.: Disentangling representations of text by masking transformers. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 778–791 (2021)