Unsupervised Concept Representation Learning for Length-Varying Text Similarity

Xuchao Zhang, Bo Zong, Wei Cheng, Jingchao Ni, Yanchi Liu, Haifeng Chen
NEC Laboratories America, Princeton, NJ, USA
{xuczhang, bzong, weicheng, jni, yanchi, haifeng}@nec-labs.com

Abstract

Measuring document similarity plays an important role in natural language processing tasks. Most existing document similarity approaches suffer from the information gap caused by context and vocabulary mismatches when comparing varying-length texts. In this paper, we propose an unsupervised concept representation learning approach to address the above issues. Specifically, we propose a novel Concept Generation Network (CGNet) to learn concept representations from the perspective of the entire text corpus. Moreover, a concept-based document matching method is proposed to leverage advances in the recognition of local phrase features and corpus-level concept features. Extensive experiments on real-world data sets demonstrate that new method can achieve a considerable improvement in comparing length-varying texts. In particular, our model achieved 6.5% better F1 Score compared to the best of the baseline models for a concept-project benchmark dataset.

1 Introduction

Measuring the similarity between documents is a fundamental problem in several natural language tasks such as information retrieval (Manning et al., 2008), paraphrase identification (Yin and Schütze, 2015) and question routing (Zhang et al., 2020). A wide range of document similarity approaches (Kusner et al., 2015; Huang et al., 2016) have been proposed to handle the fundamental problem; however, most of them are based on the assumption that the documents being compared have similar document length. However, varying-length document matching tasks are ubiquitous in many real-world scenarios. For instance, in the news categorization task, the news articles may include both short reports for breaking news or narrative reports with cumbersome details.

The document matching in varying length may introduce the information gap between two documents in the following two aspects: (i) context mismatch, which is caused by the long-length documents usually provide more detailed context to support the key information while the short-length documents contain limited context information. The issue renders the existing pre-trained natural language representation models (Conneau et al., 2017a; Devlin et al., 2018) pay more attention to the long but less important contexts, which makes their document representations distinct from the short-length documents with little context information. (ii) vocabulary mismatch, which is usually caused by the different terms usage between short and long texts, which leads them do not share majority terms. Existing document distance such as word mover's distance (Kusner et al., 2015) focus on comparing the local features. Still, the vocabulary mismatch issue makes the local features hard to be matched while the majority of vocabulary is not shared.

To address the above challenges, our approach proposes a concept-based document matching method that incorporates both local phrase features and corpus-level concepts in an unsupervised setting, where concepts can be interpreted as a group of representative features that are interpretable for humans. The main contributions of this paper can be summarized as follows: (i) A novel unsupervised concept generation network is proposed to learn corpus-level concepts in the perspective of entire text corpus. Specifically, each concept and its phrase assignment is iteratively optimized by the reconstruction loss between local phrase features and global concept representations. (ii) A new concept-based document comparison method is proposed to measure the similarity between two text documents based on augmented concept representations, which leverages the advances of local phrases and corpus-level concepts. Moreover, an enhanced concept-weight constraint is proposed to

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5611–5620
June 6–11, 2021.

improve the performance in optimizing the concept-based document similarity. (iii) Extensive experiments on several length-varying text matching datasets demonstrate that the effectiveness of our proposed approach consistently outperforms existing state-of-the-art methods. In particular, our method improved 7.1% Accuracy and 6.5% F1-score in concept-project dataset compared to the best baseline method.

The rest of this paper is organized as follows. Section 2 reviews related work, and Section 3 provides a detailed description of our proposed model. The experiments on multiple real-world data sets are presented in Section 4. The paper concludes with a summary of the research in Section 5.

2 Related Work

In this section, we briefly describe recent advances in document similarity research. We start our discussion with recent progress in supervised methods, and then we shift our focus to unsupervised settings.

2.1 Supervised Methods

A large group of previous studies (Parikh et al., 2016; Liu et al., 2018; Zhang et al., 2019; Gupta et al., 2020; Zhang et al., 2020) learns document matching model between two text sequences in supervised settings. Tan et al. (2016) exploit attention mechanism to distil important words from sentences. Yang et al. (2019a) propose an inter-sequence alignment approach considering both previous aligned features and original point-wise features. Zhou et al. (2020) present a neural approach for general-purpose text matching with deep mutual information estimation. However, these semantic alignment approaches require massive human annotations in their training process, which are expensive and infeasible to obtain in many real-world scenarios.

2.2 Unsupervised Methods

Some approaches can be used to match document in unsupervised manners, including traditional statistical approaches (Metzler et al., 2007; Pincombe, 2004; Hua et al., 2016; Zhang et al., 2017). In past few years, neural-network-based methods have been used for document representation, which includes Doc2Vec (Conneau et al., 2017b), Skip-Thought vectors (Kiros et al., 2015). More recently, the state-of-the-art representation methods focus on the contextual representations to encode words in their context such as BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019b). A comparably long text may lose its local information after being encoded as a fix-length representation due to the informative contexts. Word Mover's Distance (WMD) approaches (Yokoi et al., 2020; Wang et al., 2019) can partially solve the problem since they focus on local feature matching. However, these methods still suffer from the vocabulary mismatch issue from length-varying texts, which makes the local features hard to be matched since texts share different majority of vocabulary terms.

Few approaches consider the length-varying texts in unsupervised settings. Hongyu Gong and Xiong (2018) proposed an unsupervised document matching approach by comparing documents in a common space of hidden topics (DSHT), which is optimized by Singular Value Decomposition (SVD). Compared to this approach, our method leverages both local features and global corpus-level concepts while DSHT only compares corpus-level topics. Moreover, the proposed CGNet can generate concepts in more scalable data set compared to the matrix decomposition solution in DSHT.

3 Model

We now describe our approach to calculate the document similarity for length-varying texts. We begin by introducing the overview of our model in Section 3.1. Then we provide details of the concept generation and document matching components in Section 3.2 and 3.3. Last, the implementation details are described in Section 3.4.

3.1 Model Overview

Given a corpus of documents $\mathcal{D} = \{d_1, d_2, ..., d_n\}$, we propose a concept-based document matching approach to compute the document distance $\text{dist}(d_i, d_j)$ between any two documents d_i and d_j in the corpus. The overall architecture is shown in Figure 1, which includes two main components: 1) **Concept Generation**, which is to generate the corpus-level concepts from the entire document corpus. Each concept c_i consists of a group of document phrases by minimizing the reconstruction loss between local phrase representation and global concept representation. Moreover, both cluster divergence and evidence regularization terms are proposed to regularize the generated concepts. 2) **Doc-**

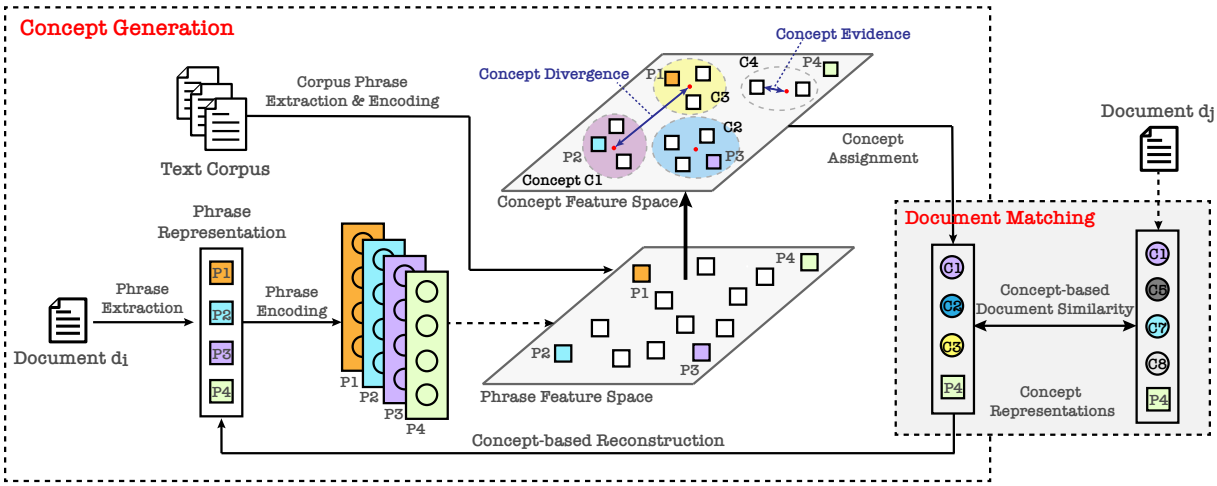

Figure 1: Overall Architecture

ument Matching. After the corpus-level concepts are learned from previous step, document matching is to calculate the document similarity based on concept-based document comparison method. Specifically, the concept-based similary adopt the Wasserstein distance (Fournier and Guillin, 2015) to compute similarity between two documents' concept representations in terms of enhanced concept-weight constraint.

3.2 Concept Generation

To generate concepts from a document corpus, we propose an unsupervised Concept Generation Network (CGNet). First, we extract a set of phrases $\mathcal{S}_p$ from the text corpus $\mathcal{D}$. The extracted phrases can be in different formats such as word tokens, noun phrases or n-grams according to the data corpus and language. Then, pre-trained language representation models such as Transformers (Devlin et al., 2018; Yang et al., 2019b) can be adopted to encode the extracted phrases into embeddings as their semantic representations. Specifically, we denote the embedding of the i-th phrase in document d_j as $\boldsymbol{p}_i^{(j)} \in \mathbb{R}^\theta$, where θ is the dimension of the phrase embedding. Suppose $\sigma(d_j)$ is the number of phrases in document d_j, we denote the phrase embedding set $\mathbb{P}^{(j)}$ for document d_j as $\mathbb{P}^{(j)} = \{\boldsymbol{p}_i^{(j)} \mid i \leq \sigma(d_j), i \in \mathbb{Z}^+\}$, where $\mathbb{Z}^+$ represents the set of positive integers. Specifically, we use $\mathbb{P} = \bigcup_{i=1}^{n} \mathbb{P}^{(i)}$ to represent the entire phrase set for all the documents.

We assume each document can not only be represented as a group of phrases but a set of corpus-level concepts, which are treated as good approximations of phrase representations. Especially for short-length texts, the limited phrases makes phrase representation hard to represent both text semantics and phrase importance. Instead, our concept representation can represent short-text semantics and weight document features in corpus perspective rather than individual document.

To learn the corpus-level concepts, we first randomly initialize κ concept centroid embeddings in the same feature space of phrases, where κ is the number of concepts. Specifically, we denote $\boldsymbol{c}_i \in \mathbb{R}^\theta$ as the embedding of the i-th concept centroid, where the concept dimension θ shares the same dimension as phrase representation. Noted that the concept centroid embeddings will be trained as model parameters in our CGNet model.

Then we assign each phrase to concepts based on its phrase embedding and concept centroids by student-t distribution as follows:

$$s_{ik}^{(j)} = \frac{\left(1 + \|\boldsymbol{p}_i^{(j)} - \boldsymbol{c}_k\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{k'=1}^{\kappa}\left(1 + \|\boldsymbol{p}_i^{(j)} - \boldsymbol{c}_{k'}\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}, \quad (1)$$

where $s_{ik}^{(j)}$ can be interpreted as the probability of the i-th phrase in document d_j assigned to the k-th concept. Since Student-t distribution has heavier tails, which makes it more prone to producing values that fall far from its mean. This characteristics can help to assign lower probability to phrases that do not belong to any concept. The parameter α can control the degrees of freedom of Student's t-distribution. Since our unsupervised setting, we let $\alpha = 1$ for all experiments.

Based on the phrase assignment on each concept, the concept representation for document d_j can be

represented as:

$$\mathbb{C}^{(j)} = \left\{ \boldsymbol{c}_k \middle| s_{ik}^{(j)} \geq \gamma, \forall i \leq \sigma(d_j), \forall k \leq \kappa \right\}, \quad (2)$$

where $\mathbb{C}^{(j)}$ is the set of concept centroid embeddings for document d_j and γ is a threshold to assign concepts for each document. When the probability $s_{ik}^{(j)}$ is greater than γ, the concept $\boldsymbol{c}_k$ is added into the concept embedding set $\mathbb{C}^{(j)}$; otherwise, the concept $\boldsymbol{c}_k$ is excluded.

To improve the concept assignment, we propose to optimize the concept centroids by minimizing the reconstruction loss between local phrases and corpus-level concepts for each document. The reconstruction loss is defined as follows:

$$\mathcal{L}_r = \frac{1}{n} \sum_{i}^{n} \text{sinkhorn}(\mathbb{P}^{(i)}, \mathbb{C}^{(i)}), \quad (3)$$

where $\mathbb{P}^{(i)}$ and $\mathbb{C}^{(i)}$ represents the embedding sets of phrases and concepts for the i-th document, respectively. Function $\text{sinkhorn}(\cdot)$ represents sinkhorn divergence (Cuturi, 2013), a sensible approximation of the Wasserstein distance (Fournier and Guillin, 2015) at a low computational cost. The experimental results in Section 4.2.4 show the sinkhorn divergence achieves empirical better performance than traditional mean squared error (MSE).

Only minimizing the reconstruction loss can easily get trivial local optima that assigns all the phrases to one concept. Thus, we propose two regularization terms, concept divergence loss and concept evidence loss, to regularize the concept centroid and avoid trivial solutions.

Concept Divergence. To prevent the similar or even duplicate concepts, we propose a divergence regularization term $\mathcal{L}_d$ that penalizes on concepts that are close to each other. The regularization term $\mathcal{L}_d$ is defined as follows:

$$\mathcal{L}_d = \sum_{i=1}^{\kappa} \sum_{j=i+1}^{\kappa} \max\left(0, \mu - \|\boldsymbol{c}_i - \boldsymbol{c}_j\|_2^2\right), \quad (4)$$

where μ is a threshold that justifies whether two concepts are similar or not. We set μ to 1.0 in our experiments. The divergence regularization exerts a large penalty when the L_2 norm distance between two concept embeddings are smaller than the threshold μ; otherwise, no penalty is produced.

Concept Evidence. To encourage each concept as close to encoded phrase instances, we propose a concept evidence regularization term $\mathcal{L}_e$, which penalizes the long distance between each concept embedding and its corresponding closest encoded phrases. The evidence regularization term $\mathcal{L}_e$ is defined as follows:

$$\mathcal{L}_e = \frac{1}{\kappa} \sum_{k=1}^{\kappa} \sum_{j=1}^{\tau} \min_j \left(\bigcup_{\boldsymbol{p}_i \in \mathbb{P}} \|\boldsymbol{c}_k - \boldsymbol{p}_i\|_2^2 \right), \quad (5)$$

where $\min_j(\cdot)$ represents the j-th minimum value in the given set and $\boldsymbol{p}_i \in \mathbb{P}$ is one of the phrase embedding from the entire phrase embedding set $\mathbb{P}$. We denote $\cup$ as the union operator to combine all the L_2 norm distance between concept centroids and phrase embeddings. We choose the sum of top-τ minimum distances as the concept evidence loss for each concept. The value of τ determines the minimum number of phrases we desired in each concept. By default, we set τ to five, which indicates a large penalty is produced while the k-th ($k \leq \tau$) closest phrases has a long distance to the concept centroid.

Finally, our loss function is the combination of reconstruction loss $\mathcal{L}_r$, concept divergence loss $\mathcal{L}_d$ and concept evidence loss $\mathcal{L}_e$ with their corresponding weights λ_r, λ_d and λ_e.

3.3 Concept-based Document Matching

Our concept-based document matching method is based on the concept generated in Section 3.2. According to the document concept assignment in Equation (2), some local phrases are excluded from any concept while its concept assignment probability $s_{ik}^{(j)} < \gamma$ for $\forall k < \kappa$. However, these local phrases may contain distinguished semantics that cannot be grouped with enough phrases as a concept, but play an important role in distinguishing the difference between documents. To involve the local phrases into our document matching task, we generate a local-feature augmented representation $\mathbb{C}_\dagger^{(j)}$ as follows:

$$\mathbb{C}_\dagger^{(j)} = \mathbb{C}^{(j)} \cup \left\{ \boldsymbol{p}_i^{(j)} \middle| \bigcup_{\substack{i \leq \sigma(d_j) \\ j \leq \kappa}} \left\{ \boldsymbol{c}_k \middle| s_{ik}^{(j)} \geq \gamma \right\} = \emptyset \right\}, \quad (6)$$

where all the local phrases that do not belong to any concept in $\mathbb{C}^{(j)}$ are added into the augmented concept embedding set $\mathbb{C}_\dagger^{(j)}$. The parameter γ is the same threshold as in Equation (2).

Based on the idea of the Wasserstein distance (Fournier and Guillin, 2015), we propose the

concept-based document similarity Ψ between augmented concept representations of documents d_i and d_j as follows:

$$\Psi(\mathbb{C}_\dagger^{(p)}, \mathbb{C}_\dagger^{(q)}) = \max \sum_{\boldsymbol{c}_i \in \mathbb{C}_\dagger^{(p)}} \sum_{\boldsymbol{c}_j \in \mathbb{C}_\dagger^{(q)}} f_{i,j} \frac{\boldsymbol{c}_i \cdot \boldsymbol{c}_j}{\|\boldsymbol{c}_i\| \|\boldsymbol{c}_j\|}$$
$$\text{s.t.} \quad \sum_{i \in \mathbb{Z}^+} f_{i,j} \leq w_{p,i}, \quad \forall i \leq \left|\mathbb{C}_\dagger^{(p)}\right|$$
$$\sum_{j \in \mathbb{Z}^+} f_{i,j} \leq w_{q,j}, \quad \forall j \leq \left|\mathbb{C}_\dagger^{(q)}\right|, \tag{7}$$

where the $f_{i,j}$ is a flow from concept representation $\boldsymbol{c}_i$ in $\mathbb{C}_\dagger^{(p)}$ to $\boldsymbol{c}_j$ in $\mathbb{C}_\dagger^{(q)}$. Parameters $w_{p,i}$ and $w_{q,j}$ represent the weight of concept i and j in document p and q, respectively. We choose the concept weight as the averaged TF-IDF weight of phrases that are assigned to the concept, which is used as upper bound constraint of the flow parameters. Overall, the concept-based document similarity is to find a flow between concept representations of two documents that maximize the similarity score.

3.4 Implementation Details

The proposed CGNet model described in this section is implemented using the Pytorch[1] framework and trained on a single Nvidia Quadro RTX 6000 GPU with 24GB memory. For phrase extraction, we set the minimum phrase frequency to 10 and maximum document frequency to 0.5. The phrases embeddings are initialized with pre-trained fastText model (Bojanowski et al., 2016) using the default dimensionality of 300. We set the number of training epochs of CGNet to 100 and batch size to 8. For the sinkhorn divergence used in Equation (3), we apply an approximate Wasserstein distance implementation[2]. For the settings of concepts, we set number of concept κ to 100 and concept threshold γ to 0.8. It should be noted that while we train our CGNet with the text corpus, the model – once trained – can be applied to new document in the same domain that is not included in the text corpus.

4 Experiment

In this section, we evaluate the performance of the model described in Section 3 on document matching task for length-varying texts.

[1]https://pytorch.org/

[2]https://github.com/dfdazac/wassdistance

4.1 Experimental Setup

We begin by introducing the evaluation settings, with details on the datasets, metrics and baselines that we use in our experiments.

4.1.1 Datasets and Labels

We conducted experiments on three publicly available datasets in different tasks: (i) **Concept-Project** (Hongyu Gong and Xiong, 2018). The dataset is to match science projects and concepts when people intend to search related projects that match a given concept. It includes 537 pairs of projects and concepts involving 53 unique concepts from the Next Generation Science Standards[3] (NGSS) and 230 unique projects from Science Buddies[4]. Each pair is labeled by human beings with the decision wther it is a good match or not. (ii) **CL-SciSumm 2017** (Prasad, 2017). The dataset consists of 494 ACL Computational Linguistics research papers covering 30 categories in total. Each category contains a reference paper and its corresponding human-annotated summary. We compare the reference summary with its corresponding reference paper and use all the citing papers as negative cases. The matching task is formulated as a ranking problem and use the reference paper as the top-1 ground-truth. (iii) **CL-SciSumm 2018** (Jaidka et al., 2019). The dataset consists of 605 research papers with reference papers including summaries and citing papers, which covers 40 categories. Different from dataset CL-SciSumm 2017, we randomly select 5 corresponding citing papers as the true candidate for each reference summary and choose the other 15 citing papers from all the citing papers as distractors.

4.1.2 Evaluation Metrics

For Concept-Project dataset, we use **Accuracy**, **Precision**, **Recall** and **F1-score** as evaluation metrics based on the binary classification predictions. The metrics including Precision, Recall and F1-score are based on positive predictions.

For the CL-SciSumm 2017 dataset, we use popular ranking evaluation metrics from the literature, which includes: (i) **Precision@1**: The proportion of predicted instances where the true reference paper appears in the ranked top-1 result. (ii) **Mean Reciprocal Rank (MRR)**: the average multiplicative inverse of the rank of the correct answer, represented mathematically as $\text{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i}$,

[3]https://www.nextgenscience.org/

[4]https://www.sciencebuddies.org/

where N is the number of samples and rank_i is the rank assigned to the true comment by a model. (iii) **Normalized Discounted Cumulative Gain (NDCG)**: the normalized gain of each reference paper based on its ranking position in the results. We set the relevance score of the true comment to one and those of the distractors (citing papers) to zero.

For the CL-SciSumm 2018 dataset, the results are evaluated by **Precision@K**: The proportion of predicted instances where the true citing papers appear in the ranked top-K result. For example, P@3 or "Precision at 3" corresponds to the percentage of cases where the true citing appears in the top 3 ranked results. We vary the value of K from 1 to 5 in our experiments.

4.1.3 Competing Methods

The following methods are included in the performance comparison: (i) **TF-IDF**, which uses the cosine similarity between the TfIdf-weighted vectors of the document as a measure of document similarity. (ii) **Infersent**, which finds the cosine similarity between document embeddings generated by the state-of-the-art sentence embedding method InferSent (Conneau et al., 2017a). (iii) **BERT**, which uses inner product between the document representation generated by the pre-trained deep bidirectional transformer (Devlin et al., 2018). (iv) **WMD** (Kusner et al., 2015), which uses word mover's distance metric based on the embeddings of document words generated by fastText[5]. (v) **WRD** (Yokoi et al., 2020), which is a variant of traditional WMD method. WRD separates word importance and word meaning by decomposing word vectors into their norm and direction. The alignment-based similarity is computed by earth mover's distance. (vi) **DSHT** (Hongyu Gong and Xiong, 2018), which matches documents by comparing them in a common space of hidden topics.

4.2 Experimental Results

We now present and discuss the empirical results of our evaluation for the three document matching tasks.

4.2.1 Concept-Project Matching

Table 1 summarizes results of the concept-project document matching task. Our model significantly outperforms all the baselines in accuracy, precision and F1-score. In particular, our model achieves 87.2% accuracy and 88.4% F1 score, which is 7.1% and 6.5% better than the best baseline method (DSHT). The improvements over all the baselines are statistically significant at a p-value of 0.01. The baseline methods including InferSent, BERT have high recalls, but low F1 scores and precision. This is because these approach cannot distinguish unmatched documents but predict most of documents are matched.

4.2.2 Summary-Reference Matching

Table 2 shows the result of Summary-Reference Matching task in CL-SciSumm 2017 dataset. From the results, we conclude that our approach outperforms all the baselines on all metrics. The results are statistically significant at $p < 0.01$ using the Wilcoxon signed rank test (Smucker et al., 2007). Since the summary-reference task only has one true reference, the other citing papers being distractors, the P@1 result becomes especially important for this task. Our approach achieves 90% precision, which is 3.3% better than the precision of the best baseline method (WMD). We also find the global representation methods such as BERT and InferSent performs worse than approaches using local features such as WMD and TF-IDF, which is different from the results in concept-project dataset. But our concept-based approach that utilizes both local and global features has consistently outperforms these baseline methods.

4.2.3 Summary-Citance Matching

Figure 2 shows the precision at K result of summary-citance matching task in CL-SciSumm 2018 dataset when k is set from one to five. Both mean and variance are presented by 10 experimental runs. From the result, we conclude that our method can significantly outperform the other baselines for all the settings of K. Specifically, our model performs around 5% better than the best baseline method, WMD. Moreover, the variance of our model is also much smaller than the other baselines, which indicates that our model is less impacted by random selected distractors. The result of WRD is not available to compute due to its out-of-memory issue. In addition, we also find the similar results that local feature-based approaches performs better than global representation methods.

4.2.4 Ablation Study

To verify the effectiveness of designed components in our approach, we make an ablation study in the

[5]https://fasttext.cc/

	Acc	Prec	Recall	F1
TF-IDF	0.538	0.540	0.993	0.700
InferSent	0.540	0.541	0.997	0.701
BERT	0.548	0.545	**1.000**	0.706
WMD	0.685	0.656	0.880	0.752
WRD	0.704	0.678	0.863	0.759
DSHT	0.801	0.807	0.832	0.819
CGNet	**0.872**	**0.865**	0.904	**0.884**

Table 1: Performance result of Concept-Project

	MRR	P@1	NDCG
TF-IDF	0.918	0.867	0.938
InferSent	0.457	0.267	0.581
BERT	0.181	0.033	0.357
WMD	0.933	0.867	0.951
WRD	0.701	0.533	0.773
DSHT	0.555	0.367	0.661
CGNet	**0.944**	**0.900**	**0.959**

Table 2: Result of Summary-Reference Matching

following settings: (i) *w/o Sinkhorn*: To demonstrate the effectiveness of the sinkhorn-based reconstruction loss, we remove the sinkhorn loss and instead simply use mean-square-error between the embeddings of phrases and concepts. (ii) *w/o Cluster Divergence Loss (CDL)*: We remove the cluster divergence loss in Equation (4) in our training process. (iii) *w/o Cluster Evidence Loss (CEL)*: To show the effectiveness of the concept evidence loss, we remove the CEL in the concept learning process in Equation (5). (iv) *w/o Enhanced Concept-weight Constraint (ECC)*: We replace the concept-weight constraint to one in our concept mover's distance to demonstrate the performance of the module.

Table 4 shows the results of the ablation study, which demonstrates that each component improves the overall performance in concept-project matching task, across our evaluation metrics. This indicates that our modeling choices are suited to tackle the inherent challenges involved in matching the length-varying documents. In particular, the cluster divergence loss has great impact on the performance since the loss can avoid assigning all the cluster centroids to the same value.

4.2.5 Parameter Analysis

We conduct several experiments to investigate the impact of the following two hyper-parameters: concept number and phrase length. (i) **Concept Number**. Figure 3 shows the results in concept-project dataset using different concept numbers from 1 to 30. We conclude that both the F1 score and Accuracy can continuously be improved when the number of concepts are increased to 10. After the concept number reaches to 10, the performance starts to degrade but still keep in a high level, which indicates that our model is not sensitive to the setting of concept number. (ii) **Phrase Length**. Figure 4 shows the performance results in concept-project dataset using different settings of phrase length. From the results, we conclude that the token-level phrases have the best performance compared to other settings even including the combination of length 1 and 2. The main reason is that the 2-gram features contain a large portion of noisy phrases that make the extracted concepts less effective for document matching.

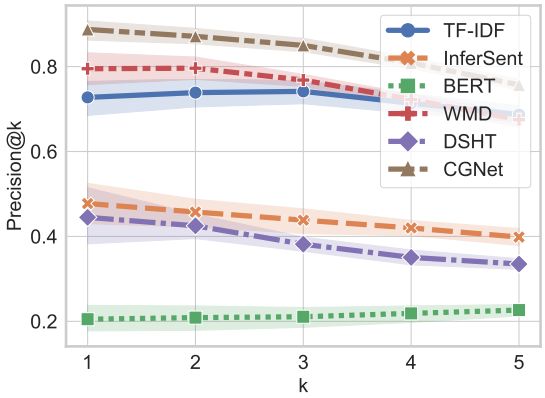

Figure 2: Result of Summary-Citance Matching

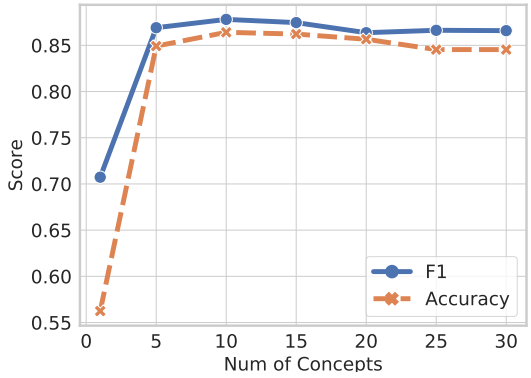

Figure 3: Parameter Analysis of Concepts Number

4.3 Efficiency Analysis

The running time of training are shown in Table 5. We can see the training time is increased linearly when the data size is increased. Since our model can be converged in a few epochs (usually

	Phrase assignments for each concept
Concept-1	fat, fur, fats, blubber, adipose, whale, beluga, oils, mammal, blubber adipose, carbohydrates, warm blooded, colligative, " or, fats, animal, calories, adipose, protein, mitochondria
Concept-2	sea, isle, tide, seas, tides, compass, vessel, boat, islands, ocean, waters, pirate, currents, winds, oceans, atlantic, ship, coast, oceanic, waves
Concept-3	bug, bugs, ants, bee, bees, insect, insects, katydid, spiders, grasshoppers, sowbugs, weevil, pillbugs, crickets, snails, peanuts, flies, do more, lions, peanut
Concept-4	odor, smell, scent, smells, rancid, taste, rancidity, emit, fishy, gone, emitting, tastes, auditory, sounds, emits, emitted, stimuli, unpleasant, bloom, sensation
Concept-5	cow, cows, age, milk, rex, ages, formula, rennet, lactase, horses, tablet, formulas, gestation, pasteurizing, calcium, matrix, ratio, breast, milkshake, cream

Table 3: Case Study of Concepts

	Acc	Prec	Recall	F1
w/o Sinkhorn	0.866	0.890	0.859	0.874
w/o CDL	0.562	0.555	0.976	0.707
w/o CEL	0.859	0.848	0.900	0.873
w/o ECC	0.805	0.780	0.890	0.832
CGNet	**0.872**	**0.865**	**0.904**	**0.884**

Table 4: Result of Ablation Study

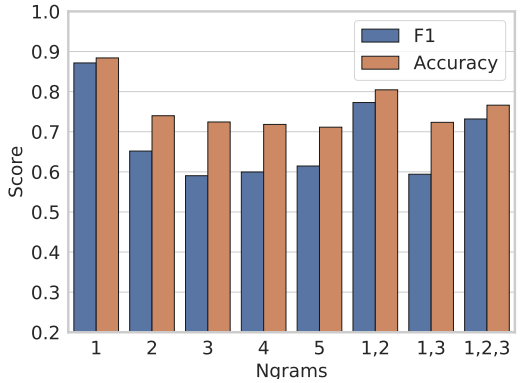

Figure 4: Parameter Analysis of Ngrams

	Concept-Project	Summary-Reference	Summary-Citance
Training Time (sec/epoch)	20.76	6.35	10.95
Eval Time (sec/pair)	0.412	0.143	0.281

Table 5: Efficiency Result for Training (second/epoch) and Testing (second/pair).

less than 100 epochs), our model can be trained in a reasonable duration. Moreover, we find the evaluation time of each dataset has less difference compared to the training time since the evaluation time is related to the size of phrases and concepts.

4.4 Interpretation of Concepts

Table 3 give some interpretation of concepts, showing top-20 phrases ranked by the phrase assignment probability in Equation (2) of five concepts generated by our CGNet model. From the results, we conclude that: (i) The generated concept is capable of representing high-level topics. For instance, the *Concept-1* relates to fat and energy of sea mammals when phrases such as fat, blubber adipose and carbohydrates appears; the *Concept-2* relates to the sea sailing when phrases such as tide, isle, compass are assigned to the concept. (ii) phrases in concepts are not only grouped by the similar semantics but the inherent co-occurrence in the text corpus. For example, the calories and mammal shares very few semantic similarity but these two terms can be connected by documents that introduce the energy storage system of sea mammals. (iii) The 2-gram phrases can introduce useful phrases such as "blubber adipose" in *Concept-1*. However, sometimes it produces some noises such as "do more" in *Concept-3*.

5 Conclusion

In this paper, an unsupervised concept representation learning method is proposed to address the length-varying text comparison problem. To achieve this, we propose a deep neural network based model to generate corpus-level concept representation and design a concept-based document matching method based on augmented concept representation that leverages the advances of both local phrase features and global concept features. Extensive experiments on real-world datasets demonstrated that our proposed method dramatically outperforms competing methods, exhibiting a significant improvement in all the metrics in different length-vary text comparison tasks.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017b. 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.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nicolas Fournier and Arnaud Guillin. 2015. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738.

Vivek Gupta, Ankit Saw, Pegah Nokhiz, Praneeth Netrapalli, Piyush Rai, and Partha Talukdar. 2020. P-sif: Document embeddings using partition averaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7863–7870.

Suma Bhat Hongyu Gong, Tarek Sakakini and Jinjun Xiong. 2018. Document similarity for texts of varying lengths via hidden topics. In *the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Ting Hua, Xuchao Zhang, Wei Wang, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Automatical storyline generation with help from twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2383–2388.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. In *Advances in Neural Information Processing Systems*, pages 4862–4870.

Kokil Jaidka, Michihiro Yasunaga, Muthu Kumar Chandrasekaran, Dragomir Radev, and Min-Yen Kan. 2019. The cl-scisumm shared task 2018: Results and key insights. *arXiv preprint arXiv:1909.00764*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966.

Xiaodong Liu, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for natural language inference. *arXiv preprint arXiv:1804.07888*.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Donald Metzler, Susan Dumais, and Christopher Meek. 2007. Similarity measures for short segments of text. In *European conference on information retrieval*, pages 16–27. Springer.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Brandon Pincombe. 2004. Comparison of human and latent semantic analysis (lsa) judgements of pairwise document similarities for a news corpus. Technical report, DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION SALISBURY (AUSTRALIA) INFO ....

Animesh Prasad. 2017. Wing-nus at cl-scisumm 2017: Learning from syntactic and semantic similarity for citation contextualization. In *BIRNDL@ SIGIR (2)*, pages 26–32.

Mark D Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 623–632.

Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473.

Zihao Wang, Datong Zhou, Yong Zhang, Hao Wu, and Chenglong Bao. 2019. Wasserstein-fisher-rao document distance. *arXiv preprint arXiv:1904.10294*.

Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019a. Simple and effective text matching with richer alignment features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4699–4709, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. Word rotator's distance: Decomposing vectors gives better representations. *arXiv preprint arXiv:2004.15003*.

Xuchao Zhang, Wei Cheng, Bo Zong, Yuncong Chen, Jianwu Xu, Ding Li, and Haifeng Chen. 2020. Temporal context-aware representation learning for question routing. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 753–761.

Xuchao Zhang, Dheeraj Rajagopal, Michael Gamon, Sujay Kumar Jauhar, and ChangTien Lu. 2019. Modeling the relationship between user comments and edits in document revision. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5003–5012.

Xuchao Zhang, Liang Zhao, Zhiqian Chen, Arnold P Boedihardjo, Jing Dai, and Chang-Tien Lu. 2017. Trendi: Tracking stories in news and microblogs via emerging, evolving and fading topics. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1590–1599. IEEE.

Xixi Zhou, Chengxi Li, Jiajun Bu, Chengwei Yao, Keyue Shi, Zhi Yu, and Zhou Yu. 2020. Matching text with deep mutual information estimation. *arXiv preprint arXiv:2003.11521*.