Matching Varying-Length Texts via Topic-Informed and Decoupled Sentence Embeddings

Xixi Zhou, Chunbin Gu, Xin Jie, Jiajun Bu*, Haishuai Wang*
Zhejiang Provincial Key Laboratory of Service Robot,
College of Computer Science, Zhejiang University
{xixi.zxx, guchunbin, jiexin, bjj, haishuai.wang}@zju.edu.cn

Abstract

Measuring semantic similarity between texts is a crucial task in natural language processing. While existing semantic text matching focuses on pairs of similar-length sequences, matching texts with non-comparable lengths has broader applications in specific domains, such as comparing professional document summaries and content. Current approaches struggle with text pairs of non-comparable lengths due to truncation issues. To address this, we split texts into natural sentences and decouple sentence representations using supervised contrastive learning (SCL). Meanwhile, we adopt the embedded topic model (ETM) for specific domain data. Our experiments demonstrate the effectiveness of our model, based on decoupled and topic-informed sentence embeddings, in matching texts of significantly different lengths across three well-studied datasets.

1 Introduction

Text matching is an important research area in natural language processing (NLP) applications such as information retrieval, natural language inference, and question answering. However, many text-matching approaches assume that the texts being compared have similar lengths (Gong et al., 2018; Zhou et al., 2020; Zhang et al., 2021a; Zou et al., 2022), and most pre-trained models, such as BERT (Devlin et al., 2019), focus on learning short sequences and are inadequate to represent complex domain-specific documents. Large language models (LLM) can directly process longer texts but require increasingly extensive training resources (Qin et al., 2023), which has evident limitations in some practical application scenarios.

A natural way for humans to deal with long texts is to break them down into smaller text segments before processing them (Nguyen et al., 2023). Inspired by this, for a varying-length text-matching task, we can split the two texts into natural sentences and put the sentences of the two texts into different sets. The motivation of our approach is: 1) If these two texts do not match, then their respective sets of sentences can be considered to belong to different classes. Therefore, the sentence points belonging to the same class should be close in their representation space, while the points from different classes should be further apart. 2) If two texts are semantically similar and consistent, the topics of the two texts should match. Accordingly, we present an effective method for varying-length text-matching with **T**opic-informed and **D**ecoupled sentence **E**mbeddings (TDE), as shown in Figure 1 and described in Section 2. First, we segment the text into sentences and utilize a state-of-the-art pre-trained model's transformer encoder to convert them into embedding points by extracting the last hidden state. We then employ SCL to optimize class-wise relations in the embedding space of sentence representations. Meanwhile, we discover the topic embeddings of the corpus by ETM (Dieng et al., 2020) and calculate the text embeddings with topic information. Finally, we concatenate the SCL-optimized and the topic-informed embeddings to predict scores for varying-length text matching.

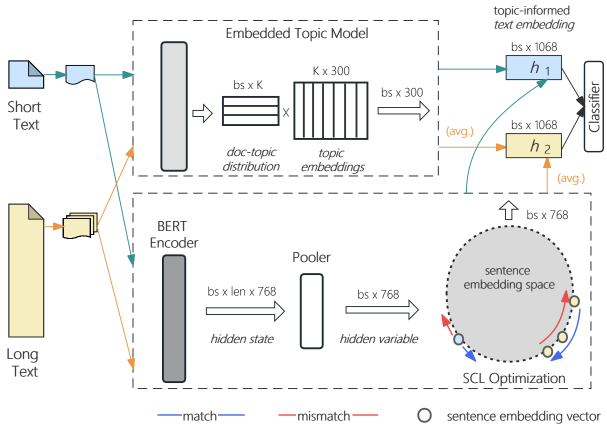


Figure 1: Architecture overview of TDE.

Two related works, the hidden topic comparison

*Corresponding authors

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1274–1280

method (HT) (Gong et al., 2018) and the unsupervised concept generation network (CGNet) (Zhang et al., 2021a), share similarities with our approach. HT extracts hidden topics and compares documents with varying lengths, while CGNet relies on local phrase features and corpus-level concept features for matching. In comparison, our method is simpler, more effective, and explainable. Moreover, our model can leverage the continuous evolution of LLMs to enhance performance.

2 Methodology

2.1 Topic-Informed Sentence Embedding

Truncating lengthy texts leads to information loss and diminishes the efficacy of text representations. To address the high space/time complexity in varying-length text matching, our proposed TDE, illustrated in Figure 1, first splits the two texts into sentences and organizes them into separate sets. Then, we employ a transformer encoder to produce fixed-sized sentence embedding vectors by extracting the last hidden state of the transformer encoder output. However, generating these embeddings can be challenging, as universal pre-trained models typically lack highly domain-specific information on real-world scenario data.

To address the problem, we leverage ETM because, first, as a neural topic model, ETM is useful for dealing with domain-specific data (Peinelt et al., 2020) and is also more suitable for joint training with other neural networks than other traditional models. Second, ETM treats both vocabulary words and topics as embeddings in a particular embedding space, which is useful when words of some specific domain are unknown. So, we follow the optimization objective ($\mathcal{L}_{etm}$) in ETM literature to calculate the topic-informed sentence embeddings by learning the document-topic distribution and topic embeddings. Then, we concatenate the topic-informed sentence embedding and the encoded output of BERT to get the full sentence representation.

At last, we calculate the average embedding of the vectors of each sentence set to obtain the text embedding (h_1 or h_2) for the original text corresponding to that set. A classifier layer takes the two average vectors as inputs and aggregates them into a single vector $m = [h_1; h_2; h_1 - h_2; h_1 \circ h_2]$ before it predicts a final matching probability with m (Mou et al., 2016). At last, a loss (e.g., cross-entropy) function is used to optimize the whole model with the predicted probability and the ground truth ($\circ$ denotes element-wise product and semicolons denote column vector concatenation).

2.2 Supervised Contrastive Optimization

After the classifier outputs the predicted probabilities of two classes through a Softmax operation, the usual cross-entropy (CE) loss ($\mathcal{L}_{ce}$) is adopted in a common Siamese text-matching network to optimize the whole model:

$$\mathcal{L}_{ce} = -\sum_{i} y_i log(\hat{y_i}), \tag{1}$$

where y_i is the ground-truth label of sample i and $\hat{y_i}$ is the sample's prediction label. However, the CE loss is unsuitable for optimizing varying-length text matching, as it is typically used for final representations, and the task-specific nature of varying-length texts can hinder its effectiveness for long texts. The CE loss exhibits shortcomings, including vulnerability to noisy labels (Zhang and Sabuncu, 2018) and potential poor margins (Elsayed et al., 2018), resulting in reduced generalization performance in varying-length text matching. We address this by dividing two input texts into multiple sentences and organizing them into sets, recognizing the matching process as a many-to-many relationship akin to a set matching problem.

Recently, contrastive learning has become the most popular self-supervised paradigm and has achieved remarkable success (Wang et al., 2024; Chen et al., 2020a; Henaff, 2020; Chen et al., 2020b; Wu et al., 2021). It is based on defining positive and negative pairs by which it aims to pull together the samples in the positive pair while pushing away the samples in the negative pair. With the development of contrastive learning, a few loss functions are proposed to improve the discrimination power, such as triplet loss (Weinberger et al., 2006) and N-pair loss (Sohn, 2016). In particular, supervised contrastive learning is proposed (Khosla et al., 2020) and used to effectively leverage label information in classifications (Nasiri and Hu, 2021) and pre-trained model fine-tuning (Zhang et al., 2021b), considering many positives and many negatives for each anchor. It is treated as a generalization of both the triplet and N-pair losses. The SCL method contrasts the set of all samples from the same class as positives against the negatives from the remainder of one batch.

Our TDE model is well-suited for optimization through SCL, effectively optimizing in-class and many-to-many sentence representations in embedding space. This approach complements the limitations of the CE optimization method, specifically addressing its local optimization deficiencies. Therefore, we introduce the supervised contrastive (SC) loss with labeled data for optimizing the Siamese text-matching network:

$$\mathcal{L}_{sc} = -\frac{1}{n|P_i|} \sum_{i=1}^{n} \sum_{z_j \in P_i} \log \frac{e^{(z_i^\top z_j/\tau)}}{\sum_{z_k \in A_i} e^{(z_i^\top z_k/\tau)}} , \quad (2)$$

where τ is the temperature, P_i denotes the positive pairs, and A_i denotes the full pair set of the anchor z_i. As shown in Equation 2, the contrastive loss with a small τ tends to make the embedding distribution more uniform (Wang and Isola, 2020). On the other hand, contrastive loss with a large temperature is less sensitive to the hard negative samples (Wang and Liu, 2021). As the temperature τ controls the strength of penalties on hard negative samples, we follow the previous literature (Gunel et al., 2021) and choose the τ values based on the hardness of negative samples to tune performance.

With supervised contrastive optimization, our TDE model has a hybrid loss function consisting of an original cross-entropy and a supervised contrastive loss: $\mathcal{L}_{OPT} = \mathcal{L}_{etm} + \lambda_1 \cdot \mathcal{L}_{sc} + \lambda_2 \cdot \mathcal{L}_{ce}$, where λ_i is a hyperparameter. According to our experiments, hybrid optimization gives a higher accuracy than training by these losses alone.

3 Experiments

3.1 Datasets

(1) Concept-Projects[1] (Gong et al., 2018) is designed for evaluating text-matching algorithms with varying-length inputs. A "Concept" represents a science curriculum summarizing a "Project," which is a science project document. To assess if a project aligns with a concept, individuals must determine the match. Concepts are typically brief, around 50 words, while project documents are lengthy, exceeding 1000 words. The dataset comprises 537 project-concept pairs annotated by contributors based on human judgment of matching.

(2) CL-SciSumm 2017[2]**.** This dataset (Prasad, 2017) contains 494 Computational Linguistics research papers in 30 categories. For each category, the dataset provides one reference paper and around 10 citing papers. At the same time, the dataset also provides a corresponding human-created summary for each reference paper. Following the literature (Zhang et al., 2021a), we label the reference paper as the positive candidate and all the citing papers as negative candidates when conducting paper-retrieval experiments by the summary. The matching task on this dataset is to retrieve and rank papers by a summary, which takes its corresponding reference paper as the top-1 ground truth.

(3) CL-SciSumm 2018[3]**.** This dataset (Jaidka et al., 2018) contains 605 research papers in 40 categories. For each category, the dataset provides one reference paper and at least 10 citing papers that cite this reference paper. At the same time, the dataset also provides a corresponding human-created summary for each reference paper. Following the literature (Zhang et al., 2021a), when conducting paper-retrieval experiments by a summary, we randomly take 5 citing papers from the same category as this summary to label them as positive candidates and randomly take 15 citing papers that do not belong to the same category as the summary to label them as negative candidates. The matching task on this dataset is to retrieve and rank papers by the summary, which takes the citing papers from the same category as this summary as the ground truth.

We follow the same splitting settings from the original literature of all three datasets for a fair comparison with all baseline models.

3.2 Evaluation Metrics

We evaluate three datasets with different metrics:

- For the Concept-Project, we follow the unified classification evaluation method (Gong et al., 2018) to evaluate the matching-prediction labels of texts by calculating a matching threshold using the similarity scores or relative distances of all text vectors in the model's pretrained representation space from the training data.

- For the CL-SciSumm 2017, we use popular ranking evaluation metrics from the literature, which include: (i) Precision@1: The proportion of predictions where the correct answer appears in the top-1 location of the retrieval result. (ii) Mean Reciprocal Rank (MRR). This ranking metric calculates the location of the

[1] https://github.com/HongyuGong/Document-Similarity-via-Hidden-Topics

[2] https://github.com/animeshprasad/clscisumm2017

[3] https://github.com/WING-NUS/scisumm-corpus

Method	Concept-Project Matching				Summary-Reference Retrieval		
	Acc.	Prec.	Recall	F1	Prec@1	NDCG	MRR
TF-IDF	53.8	54.0	**99.3**	70.0	86.7	93.8	91.8
Doc2Vec (Le and Mikolov, 2014)	90.5	86.2	96.1	90.9	65.2	77.2	71.5
WMD (Kusner et al., 2015)	68.5	65.6	88.0	75.2	86.7	95.1	93.3
HT (Gong et al., 2018)	80.1	80.7	83.2	81.9	36.7	66.1	55.5
SBERT (Reimers and Gurevych, 2019)	75.1	78.8	96.4	83.6	65.1	72.4	68.4
RE2 (Yang et al., 2019)	86.6	88.8	92.5	88.9	82.1	85.0	83.0
WRD (Yokoi et al., 2020)	84.9	83.3	89.2	86.2	53.3	77.3	70.1
CGNet (Zhang et al., 2021a)	87.2	86.5	90.4	88.4	90.0	95.9	94.4
TDE w/o ETM (ours)	94.1	97.8	95.2	94.4	88.7	94.7	94.0
TDE (ours)	**94.3**	**98.1**	95.8	**94.6**	**90.2**	**96.9**	**95.2**

Table 1: For Concept-Project Matching, TDE achieves the best the metric scores on accuracy, precision, and F1 score, which are 3.8%, 9.3%, and 3.7% better than the second-best models. For Summary-Reference Retrieval, TDE achieves the best metric scores on Prec@1, NDCG, and MRR, which are 0.2%, 1.0%, and 0.8% better than the second-best models.

first correct answer in the retrieval result. The higher the position of the first correct answer, the greater the MRR value, represented as MRR=$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$ for multiple queries N, where $\frac{1}{rank_i}$ is the reciprocal rank for a single query. (iii) Normalized Discounted Cumulative Gain[4] (NDCG): This ranking metric compares the relevance of the answers returned in the retrieval to the relevance of the answers in ideal order. NDCG is the quotient of the actual DCG and the ideal DCG.

- For the CL-Scisumm 2018, we use Precision@k as the retrieval metric to calculate the proportion of predictions where the correct answer appears in the top-k locations of the retrieval results. We tune k from 1 to 5 in our experiments.

3.3 Baseline Methods

We use eight strong baselines in experiments of matching varying-length texts: TF-IDF, Doc2Vec (Le and Mikolov, 2014), Word Movers' Distance (WMD) (Kusner et al., 2015), Hidden Topics (HT) (Gong et al., 2018), RE2 (Yang et al., 2019), Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), Word Rotator's Distance (WRD) (Yokoi et al., 2020), and CGNet (Zhang et al., 2021a). We follow the same data pre-processing rules and use the original implementation methods of baselines. We did twenty experiments on three datasets to observe those experiments' average results.

3.4 Experimental Results and Analysis

Table 1 and Figure 2 show that our proposed TDE performs better than the baseline models in match-

[4]https://scikit-learn.org/stable/modules/model_evaluation

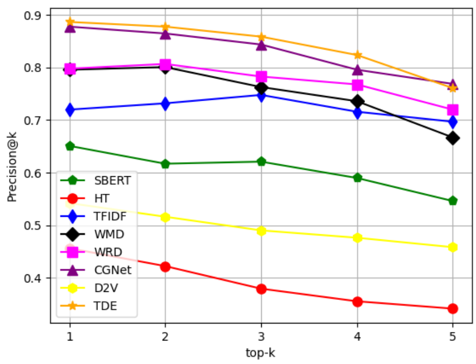
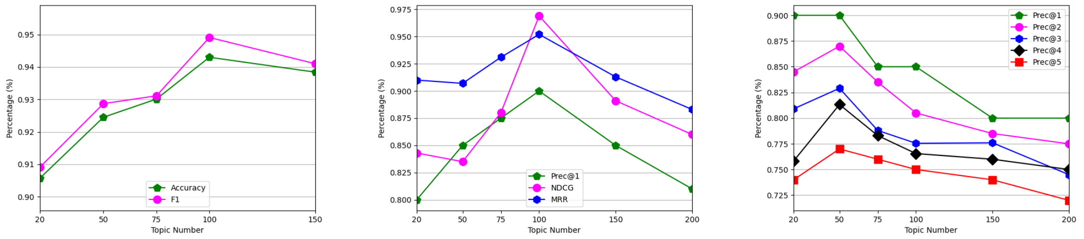


Figure 2: For Summary-Citance Retrieval, we adopt Precision@k as the experiment metric and tune k between 1 and 5. The experimental results show that our approach achieves the best performance in the baseline models when k is from 1 to 4.

	τ=0.07	τ=0.2	τ=0.7	τ=1.0
SC + CE (ours)	**94.6**	94.2	93.8	93.0
SC	93.7	93.5	93.2	92.5
CE (w/o τ)	92.3			

Table 2: Ablation study for SC loss and its temperature coefficients with the Concept-Project task (F1 score).

ing and retrieval performance on three well-studied datasets. As shown in Table 1, there is no obvious improvement with ETM enabled on the Concept-Project Matching dataset because it is from a more general domain and is not beneficial for ETM optimization with the universal pre-training model (BERT). For Summary-Reference Retrieval, the performance improvement is limited before ETM is enabled. That is because the universal pre-training model lacks domain-specific information necessary for text representation encoding of this dataset, a professional field dataset from linguistics papers. After ETM was enabled, there were noticeable performance enhancements, illustrating the impor-



(a) Concept-Project Matching (b) Summary-Reference Retrieval (c) Summary-Citance Retrieval

Figure 3: Parameter analysis of latent topic number.

tance of the ETM module in generic matching tasks with domain-specific cases.

We investigate the impact of the key hyper-parameter, latent topic number (K, as shown in Figure 1), in the three varying-length matching tasks. The three subfigures in Figure 3 show the experimental results using various topic number settings (from 20 to 200) on three datasets: (a) Concept-Project, (b) Summary-Reference, and (c) Summary-Citance, respectively. Figure 3 (a) shows that both Accuracy and F1 scores are improved until the number of topics increases to 100 in the Concept-Project matching task. After the topic number reaches 100, the performance starts to degrade, which indicates that the topic number has an optimal experimental value. For the other two tasks (Summary-Reference Retrieval and Summary-Citance Retrieval), the optimal values are 100 and 50 in our experiments, as illustrated in Figure 3 (b) and Figure 3 (c), respectively. Therefore, we conclude that setting the topic number too small or too large will lead to semantic confusion in each corresponding latent variable and performance degradation on the related dataset. A suitable setting of the latent topic number tends to make the representations better for the matching process. Therefore, in the performance comparison to the baseline models, we take the experimental results when K is set as 100, 100, and 50 for the three datasets, respectively, because of the optimal performance with these settings experimentally.

Our ablation experiments highlight the impact of varying temperature coefficients (τ) of the supervised contrastive loss on overall performance. The results in Table 2 indicate that adjusting some temperature values has negligible performance improvement, but notable gains are observed with specific τ values. This flexibility in tuning performance through supervised contrast learning allows for finding optimal temperature values tailored to different datasets. So, the ablation tests show the crucial role of supervised contrast learning in our task and its contribution to our method.

4 Conclusion

We present a novel and effective method for matching texts of varying lengths. We split long texts into sentences and encode them with the advanced transformer. Leveraging the in-class and many-to-many optimization characteristics of supervised contrastive learning and the domain-specific ability of embedded topic modeling, our model demonstrates superiority against the baseline models on three real-world datasets in both effectiveness and explainability of varying-length text matching.

Limitations

The efficacy of our method may be influenced by the quality of the natural sentences undergoing segmentation. In our experiments, we observed that some sentences generated through segmentation lack essential semantic information. Therefore, incorporating these sentences in contrastive learning training could potentially have a negative impact on the final quality of sentence representations and model performance. We anticipate that refining the pre-processed sentences will lead to further improvements in the existing framework in the future.

Acknowledgements

This work is supported by the National Key R&D Program of China (2022ZD0160703), the National Natural Science Foundation of China (Grant Nos. 62202422 and 62372408), and the National Key Research and Development Program of China (Grant Nos. 2022YFB4500300 and 2018YFB1403202).

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. 2018. Large margin deep networks for classification. In *Advances in NeurIPS*, volume 31.

Hongyu Gong, Tarek Sakakini, Suma Bhat, and Jinjun Xiong. 2018. Document similarity for texts of varying lengths via hidden topics. In *ACL*, pages 2341–2351.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR.

Kokil Jaidka, Michihiro Yasunaga, Muthu Kumar Chandrasekaran, Dragomir Radev, and Min-Yen Kan. 2018. The cl-scisumm shared task 2018: Results and key insights. In *In Proceedings of BIRNDL*, pages 74–83.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in NeurIPS*, volume 33, pages 18661–18673.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*, pages 957–966.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of ACL*, pages 130–136.

Alireza Nasiri and Jianjun Hu. 2021. SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification. *arXiv e-prints*, page arXiv:2103.01929.

Thong Nguyen, Sean MacAvaney, and Andrew Yates. 2023. Adapting learned sparse retrieval for long documents. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1781–1785, New York, NY, USA. Association for Computing Machinery.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Animesh Prasad. 2017. Wing-nus at cl-scisumm 2017: Learning from syntactic and semantic similarity for citation contextualization. In *Joint Workshop on BIRNDL*, page 26–32, Tokyo, Japan. CEUR.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP)*, pages 3982–3992.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in NeurIPS*, volume 29.

Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *2021 CVPR*, pages 2495–2504.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of ICML*, volume 119, pages 9929–9939. PMLR.

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. 2024. Contrast everything: A hierarchical contrastive framework for medical time-series. *Advances in Neural Information Processing Systems*, 36.

Kilian Q Weinberger, John Blitzer, and Lawrence Saul. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in NeurIPS*, volume 18. MIT Press.

Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, and Noah Goodman. 2021. Conditional negative sampling for contrastive learning of visual representations. In *International Conference on Learning Representations*.

Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019. Simple and effective text matching with richer alignment features. In *Proceedings of ACL*, pages 4699–4709.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. Word rotator's distance. In *Proceedings of EMNLP*, pages 2944–2960.

Xuchao Zhang, Bo Zong, Wei Cheng, Jingchao Ni, Yanchi Liu, and Haifeng Chen. 2021a. Unsupervised concept representation learning for length-varying text similarity. In *Proceedings of NAACL*, pages 5611–5620.

Yifan Zhang, Bryan Hooi, Dapeng Hu, Jian Liang, and Jiashi Feng. 2021b. Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. In *Advances in NeurIPS*, volume 34, pages 29848–29860.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in NeurIPS*, pages 8778–8788.

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

Yicheng Zou, Hongwei Liu, Tao Gui, Junzhe Wang, Qi Zhang, Meng Tang, Haixiang Li, and Daniell Wang. 2022. Divide and conquer: Text semantic matching with disentangled keywords and intents. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3622–3632, Dublin, Ireland. Association for Computational Linguistics.