Exploring Topic Landscape for Question-Answering Models in Hyperbolic Embedding Space

Sumanta Kashyapi
sumantakashyapi@gmail.com
University of New Hampshire
Durham, New Hampshire, USA

Laura Dietz
dietz@cs.unh.edu
University of New Hampshire
Durham, New Hampshire, USA

ABSTRACT

This notebook describes the submission from the TREMA-UNH team to the TREC 2023 deep learning track. Conventional DPR systems use dense vector representations from large language models such as BERT to measure how similar queries are to candidate passages. For effective open-domain question-answering, it's crucial for the embedding model to grasp both high-level topics and their detailed subtopics. While recent DPR systems implicitly learn topic similarities, explicitly integrating topic taxonomies would be beneficial. Vital article category scheme from Wikipedia is utilized to establish an overarching topic framework, and a hyperbolic embedding space is used to gain insights into topic hierarchies. When integrated into a DPR system, the entire topic landscape is considered while responding to a query. The resulting DPR system is utilized to produce runs for the reranking task of TREC 2023 deep learning track.

1 INTRODUCTION

Traditional Dense Passage Retrieval systems (DPR) leverage dense vector representations from large language models (e.g. BERT) to estimate the similarities between a query and a set of candidate passages. The underlying assumption behind this approach is that if a passage is relevant for a query then the embedding vectors of the pair should be in close proximity to each other in the latent embedding space learned by the language model. Therefore, the semantic knowledge captured in the embedding model used in the DPR system plays a pivotal role. However, for open domain question-answering tasks, the embedding model should also have intricate understanding of topics from various domains in fine-grained details in order to make the correct relevance judgements. Specifically, the embedding model should be able to distinguish between different coarse-grained topics (e.g. technology vs. politics) and be aware of the hierarchical relationships within each of the topic (e.g. technology, computer science, AI etc.). Recent DPR systems employ variants of contrastive learning approaches to implicitly learn about similarity between different topics by examples of relevant vs. non-relevant query-passage pairs. However, to answer queries from adhoc domains, it would be beneficial if we explicitly incorporate the taxonomic information about various topics. Hence, we propose a topic aware embedding model for DPR systems that has the following characteristics:

(1) *Property 1:* should contain information about coarse-grained and orthogonal topics such that it can distinguish between them.
(2) *Property 2:* should contain hierarchical information about fine-grained subtopics under each of those topics such that it can estimate the similarity within each topic.

To achieve *Property 1*, we take inspiration from the recent advancements in ideal prototype learning where a set of ideal points in the embedding space are determined apriori to guide the training process of the embedding model. For our task, we leverage the category scheme of Wikipedia vital articles as the ideal prototype of the topic landscape for our embedding model. For *Property 2*, we employ hyperbolic embedding space to efficiently learn the hierarchical relationships of the subtopic taxonomies within each topic. When such an embedding model is incorporated within a DPR system, it takes the complete landscape of topics into account while answering a query.

2 PROPOSED APPROACH

2.1 Ideal Topic Landscape

The set of topics covered by Wikipedia articles is large both in terms of breadth and depth. Therefore we derive our ideal topic landscape from Wikipedia categories. Specifically, we leverage 11 top-level categories of level 3 vital articles [1].

2.2 Capturing Subtopic Hierarchies

Topics related to a broad query are inherently hierarchical in nature where the scope of the topic narrows down with increasing depth of the hierarchy. Hence, hyperbolic embedding spaces seem to be the ideal choice in this case because it has been shown that hyperbolic spaces are more efficient than Euclidean spaces for embedding hierarchical data [1, 8]. For training data, we leverage the subtopic taxonomies from Wikipedia articles under the 11 top-level categories obtained from the previous step. Specifically, we utilize the Wikimarks dataset to obtain the topic hierarchies [3]. The following section presents our approach to train embedding models in hyperbolic space to capture these hierarchies.

2.3 Hyperbolic Embedding Space for Subtopics

Let A be the vocabulary of tokens present in a given corpus C. Also, let T be a directed acyclic graph, representing the taxonomy of all subtopics present in the corpus. Each node $t_k \in T$ is a particular subtopic. The directed edges form entailment relations [2, 11] of topics e.g. an edge $e(t_k, t_l)$ from topic t_k to t_l signifies that t_l is a subtopic of t_k.

Now, let us define a document a_i from our corpus C as a sequence of tokens taken from the vocabulary A:

$$a_i = (a_{i1}, ..., a_{in}) \text{where } a_{ij} \in A \tag{1}$$

$$\tag{2}$$

[1] https://en.wikipedia.org/wiki/Wikipedia:Vital_articles

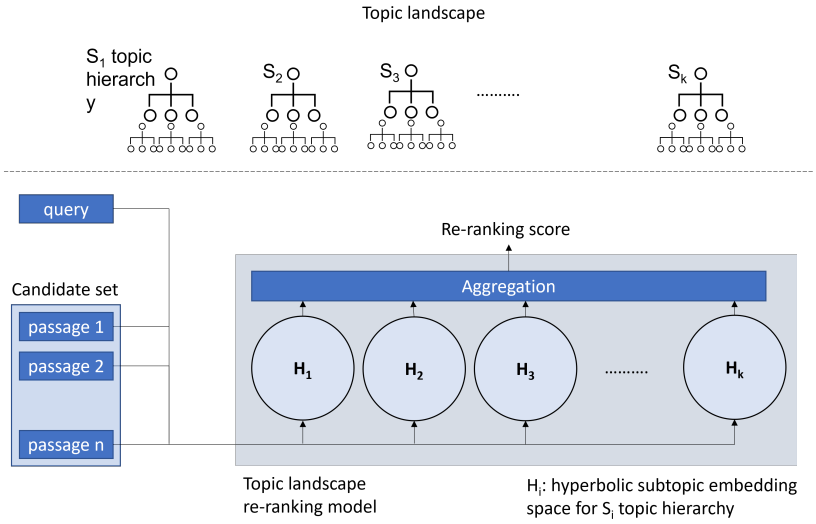

Figure 1: The diagram describes how the proposed approach is used to rerank passages in response to a query.

We also assume that a ground truth is provided to us that maps each document a_i to a set of nodes in the topic taxonomy:

$$\tau(a_i) = \{t_i \in T\} \tag{3}$$

Let $\mathbb{H}^n$ be a n-dimensional hyperbolic space. We use the hyperbolic embeddings of T in $\mathbb{H}^n$ as the ideal prototypes [6] for various topics present in the corpus. From here on, $\mathbb{H}^n$ is referred as the topic landscape.

$$\mathcal{H}(T) = \{t'_k \in \mathbb{H}^n\} \tag{4}$$

Here, $\mathcal{H}$ is a combinatorial approach to generate hyperbolic embeddings from a given hierarchy [9, 10] such that the relationships between all t'_k are preserved according to T.

Now, we learn a document embedding model $\mathcal{M}_\theta$ that produces embeddings of the texts of documents in the same topic landscape $\mathbb{H}^n$.

$$\mathcal{M}_\theta : A \longrightarrow \mathbb{H}^n \tag{5}$$

To learn $\mathcal{M}_\theta$, we establish a learning objective to minimize the entailment cone membership error [4, 12] of the embedded documents. Specifically, we minimize $\mathcal{L}(\mathcal{H}, \mathcal{M}_\theta)$ which is modified from equation 32 of [4] as follows:

$$\mathcal{L} = \sum_{(t,a)\in P} E(t, \mathcal{M}_\theta(a)) + \sum_{(t',a')\in N} max(0, \gamma - E(t', \mathcal{M}_\theta(a'))$$

γ = loss margin

P, N = set of positive and negative entailments

Here, based on how we define P and N in the above equation, we can have two variations of the trained model $\mathcal{M}_\theta$:

(1) $(t, a) \in P$ if there exists $t_k \in \tau(a)$ such that t entails t_k, meaning $e(t, t_k) \in T$. Otherwise, $(t, a) \in N$.

(2) $(t, a) \in P$ if and only if $e(t, t_k) \in T$ for all $t_k \in \tau(a)$. Otherwise, $(t, a) \in N$.

Once the embedding model $\mathcal{M}_\theta$ is trained, we can utilize it to answer queries regarding answer topics. We refer to the large body of work on dense retrieval [5, 7, 13] for implementation and design specifics.

3 APPLICATION IN RERANKING TASK

Our approach of utilizing the trained embedding model in a DPR system is depicted in Figure 1. Specifically, we apply the proposed system in the reranking task of the deep learning track. A set of 11 pretrained distilbert models are fine-tuned for each of the Wikipedia vital articles topics using poincare embeddings of the corresponding subtopic hierarchies as described in the previous section. These models are then used to embed candidate set passages and queries. Based on different aggregation methods and inference techniques, we obtain three variants of the proposed model. For each of these variants, the ranking score is calculated as the following:

(1) **hyperclose-min**: For each query passage pair, the minimum hyperbolic distance estimated across the 11 embedding model is considered as the reranking score.
(2) **hyperclose-mean**: For each query passage pair, the average hyperbolic distance estimated across the 11 embedding model is considered as the reranking score.
(3) **hyperclose-bestm**: For each query a model is selected from the set which has the lowest cumulative hyperbolic distance between the query embedding and the set of candidate passage embeddings. Then the distances are used to generate the final ranking

4 RESULTS ON DEEP LEARNING RERANKING TASK

We participate in the Deep Learning track of TREC2023 with runs corresponding to the reranking task. A trained reranking model

Table 1: Reranking performance of our runs in the reranking task of Deep Learning track TREC 2023.

Methods	MAP	Rprec
hypirclose-bestm	0.0447	0.1215
hypirclose-mean	0.0453	0.1175
hypirclose-min	0.0452	0.1144

obtained from the technique described in the previous section is utilized to rerank candidate passages corresponding to each queries for the task. We consider our approach unsupervised as our model is not trained directly on the MSMARCO data but a subset of Wikipedia articles. Table 1 demonstrates the reranking performance of our method in terms of MAP and Rprec.

We also analyze reranking performance of our model on a per-query basis. Following are some of the queries for which our model performs particularly well: *how to unlock the word document*, *koit number*, *what is myrrh essential*.

5 CONCLUSION

In our participation in TREC 2023 Deep Learning track, we focus on augmenting extrinsic topic information into dense passage retrieval systems in context of open domain question answering tasks. By integrating the vital article category hierarchies from Wikipedia into a DPR system operating in a hyperbolic embedding space, we achieve explicit incorporation of topic taxonomies into the dense vector representation. The model showed particular strength in specific queries, indicating its potential to improve precision in information retrieval across a broad range of topics.

REFERENCES

[1] Ivana Balazevic, Carl Allen, and Timothy Hospedales. 2019. Multi-relational poincaré graph embeddings. *Advances in Neural Information Processing Systems* 32 (2019).

[2] Bhavana Dalvi, Peter Alexander Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining Answers with Entailment Trees. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:233297051

[3] Laura Dietz, Shubham Chatterjee, Connor Lennox, Sumanta Kashyapi, Pooja Oza, and Ben Gamari. 2022. Wikimarks: Harvesting Relevance Benchmarks from Wikipedia. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3003–3012.

[4] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*. PMLR, 1646–1655.

[5] Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022. Tevatron: An Efficient and Flexible Toolkit for Dense Retrieval. *ArXiv* abs/2203.05765 (2022). https://api.semanticscholar.org/CorpusID:247411217

[6] Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. 2021. Hyperbolic busemann learning with ideal prototypes. *Advances in Neural Information Processing Systems* 34 (2021), 103–115.

[7] Jimmy J. Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, Rodrigo Nogueira, and David R. Cheriton. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). https://api.semanticscholar.org/CorpusID:235366815

[8] Maximilian Nickel and Douwe Kiela. 2017. Poincaré Embeddings for Learning Hierarchical Representations. *ArXiv* abs/1705.08039 (2017). https://api.semanticscholar.org/CorpusID:25418227

[9] Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. 2018. Representation Tradeoffs for Hyperbolic Embeddings. *Proceedings of machine learning research* 80 (2018), 4460–4469. https://api.semanticscholar.org/CorpusID:4753193

[10] Rik Sarkar. 2011. Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane. In *International Symposium Graph Drawing and Network Visualization*. https://api.semanticscholar.org/CorpusID:18268637

[11] Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 41–48.

[12] Tao Yu, Toni JB Liu, Albert Tseng, and Christopher De Sa. 2023. Shadow Cones: Unveiling Partial Orders in Hyperbolic Space. *arXiv preprint arXiv:2305.15215* (2023).

[13] Xinyu Crystina Zhang, Xueguang Ma, Peng Shi, and Jimmy J. Lin. 2021. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. *ArXiv* abs/2108.08787 (2021). https://api.semanticscholar.org/CorpusID:237213465