Document Reranking: Information Retrieval Using Contextual Language Models for Document Representation Learning

2020-04-22

Celine Liu, Sheena Salwan, Kishan Maladkar, Kristian Berg
Mentor: Ganesh

Motivation & Task

Literature Review

BERT: Pre-training of Deep Bidirectional Transformers for Language (2016)

Deeper Text Understanding for IR with Contextual Neural Language Modeling (2019)

Document Ranking with a Pretrained Sequence-to-Sequence Model (2020)

Data

Dataset: MS MARCO Passage ranking dataset

Domain: Bing search engine query, passages (English), Relevance Labels

Size: 5057 queries, 102765 passages, 70% training, 20 dev, 10% test.

Data processing:

- Replace non-meaningful text such as "...." with empty string in passages
- Classification based - [CLS] + [<query_tokens>] + [SEP] + [<doc_tokens>] for query document pair learning with relevance scores.
- Distance based - [CLS] + [<tokens>] for representation learning with distance measure to between query & document.

Model Overview

Classification-Based

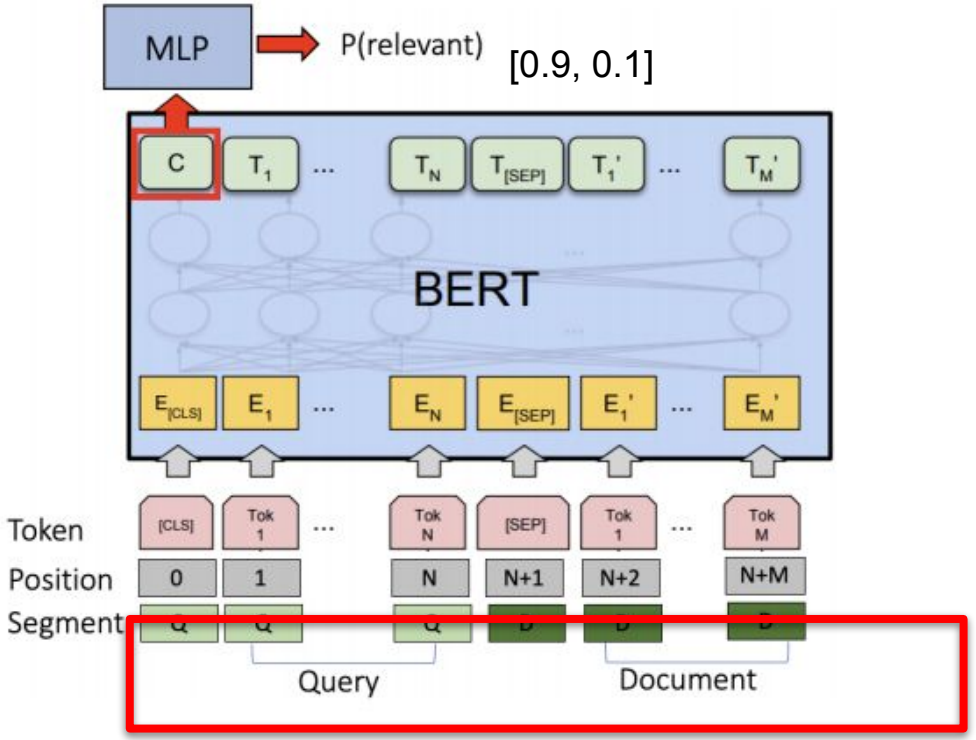
- **Model:** `DistilBertForSequenceClassification`
- First node of the final layer is used as a representation for the entire query-document pair
- MLP to predict the probability of relevance (binary classification)
- **Ranking:** based on classification probability

Distance-Based

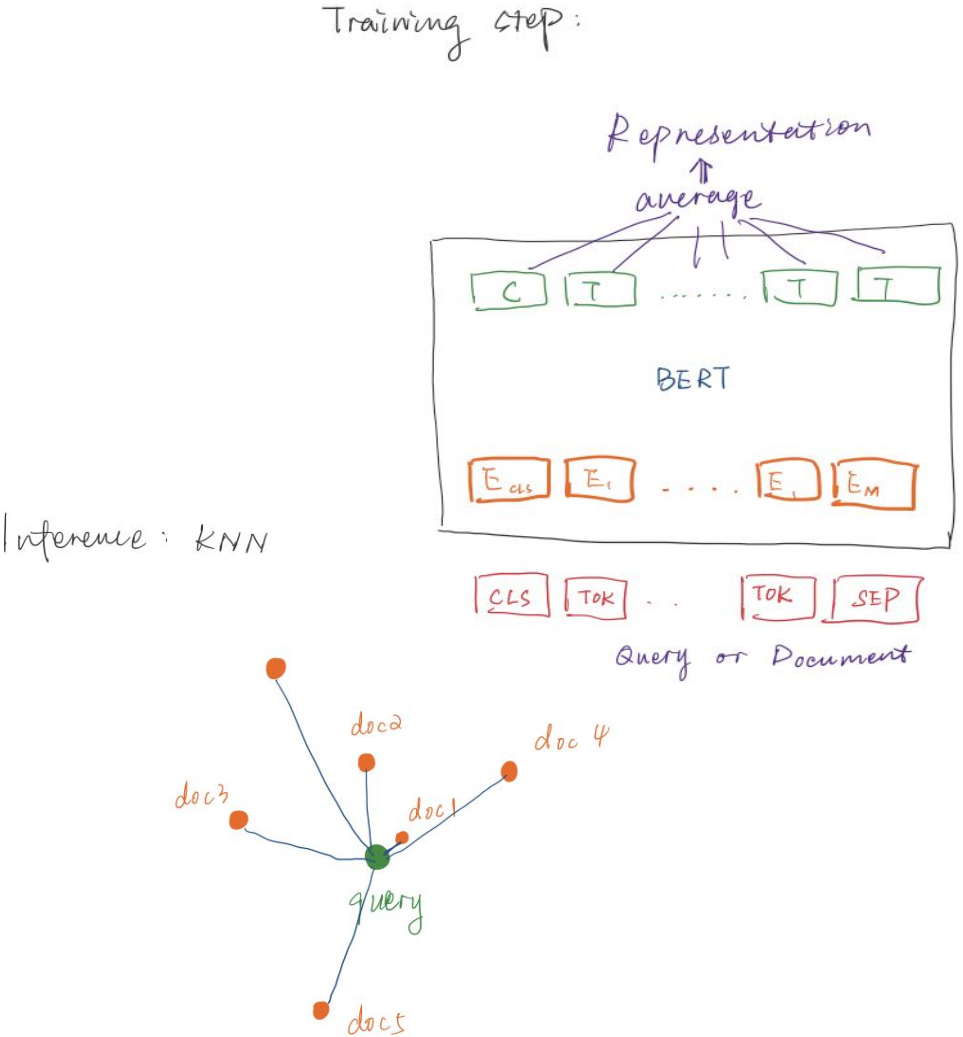
- **Model:** `DistilBertModel`
- First node of the final layer is used as the encoding of **each query and document**
- **Loss function:** `TripletMarginLoss`
- **Output:** representations for query and documents
- **Ranking:** based on different distance between query and documents (**nearest neighbour**)

Classification Based

- Only use first node of the last layer
- Two-layer MLP to turn representation to probability
- Input is query concatenated with document



Distance-Based



- Query and document are fed separately to the input. They are NOT concatenated.

- At inference time, documents can be **preprocessed** and stored as a vector representation, faster than classification based model

- TripletMarginloss is used here to make sure that query is closer to relevant document than non-relevant ones

  - loss (q, p, n) = max{dist(q, p} - dist{q, n} + margin, 0}
  - q = query, p = document1, n = document2

- To rank documents for a query, we can use the idea of k-NN in vector space

Demo

Project
Repository

Failure Cases

number of students at UBC

13.42: The University of British Columbia is a global centre for teaching, learning and research, consistently ranked among the top 20 public universities in the world and recently recognized as North America's most international university.

13.93: The University of British Columbia attracts, nurtures and proactively transform more than 58,000 students from Canada and 140 countries.

14.43: University of British Columbia Master of Data Science (MDS) is a 10-month, full-time, accelerated professional graduate program offered at both the University of British Columbia Vancouver and Okanagan campuses.

15.58: Vancouver is a major city in western Canada, located in the Lower Mainland region of British Columbia. As the most populous city in the province, the 2016 census recorded 631,486 people in the city, up from 603,502 in 2011. The Greater Vancouver area had a population of 2,463,431 in 2016, making it the third-largest metropolitan area in Canada.

Experiments & Results

Classification based

- **Experiment 1:** No training
- **Experiment 2:** Train end-to-end (prone to overfitting)
- **Experiment 3:** Freeze transformer, only train two-layer MLP

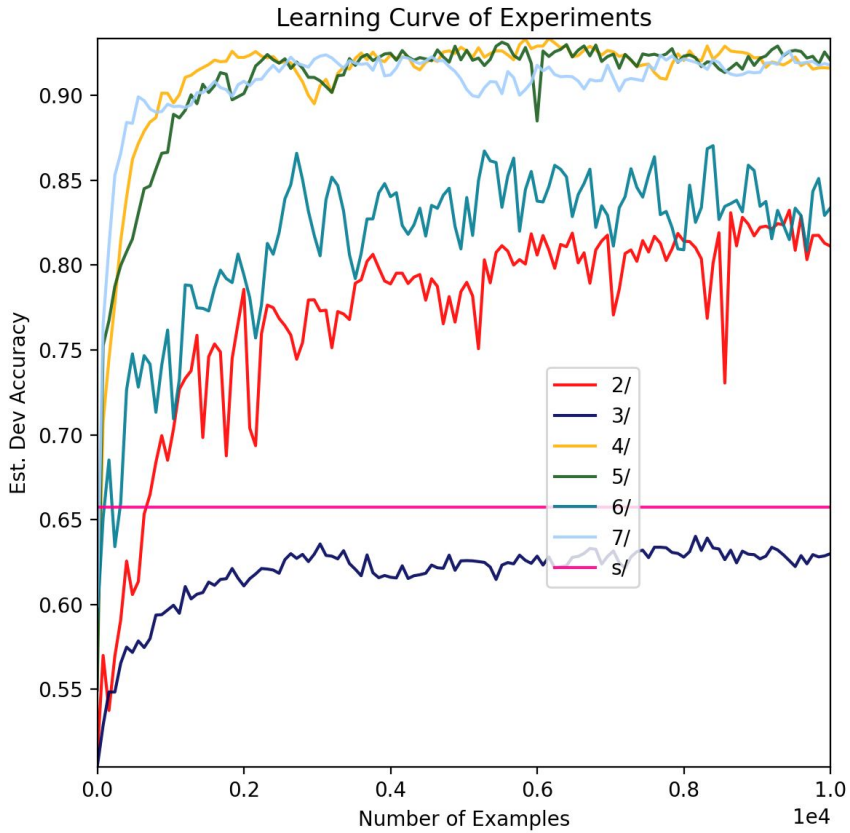
Distance based

Using **TripletMarginLoss function**

- **Experiment 4:** margin=1.0, distance=L2 (good)
- **Experiment 5:** margin=0.1, distance=L2
- **Experiment 6:** margin=1.0, distance=Cosine

Using **DistilBertForMaskedLM**

- **Experiment 7:** average of last layer (good)



Evaluation

Use Rank aware Evaluation Metrics like **MRR- Mean Reciprocal Rank**

MRR@10 - Mean Reciprocal Rank for 10 predictions

- high focus on the first relevant element of the list
- suited for targeted searches

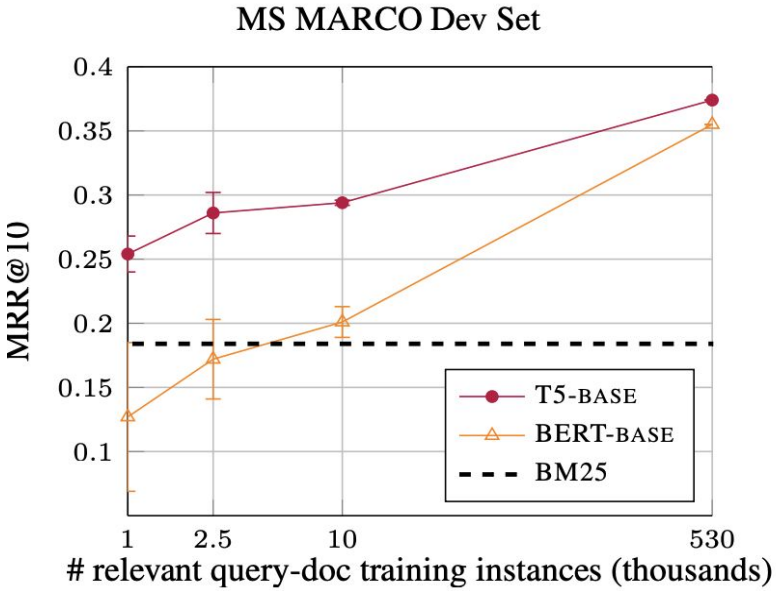


Image from: Document Ranking with a Pretrained Sequence-to-Sequence Model (2020)

Conclusion & Future Work

- Finish the T5 model
- Finish the MRR Evaluation of Models
- More hyperparameter tuning
- Train on full dataset leveraging TPU

References

Project Repo: https://github.ubc.ca/siliang6/585_teamTR/tree/master

A Deep Relevance Matching Model for Ad-hoc Retrieval. Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. *arXiv 2016*. https://arxiv.org/abs/1711.08611. (2016)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *arXiv 2018*. https://arxiv.org/abs/1810.04805. (2018)

Deeper Text Understanding for IR with Contextual Neural Language Modeling. Zhiyun Dai and Jamie Callan. *SIGIR 2019*. https://doi.org/10.1145/3331184.3331303. (2019)

Passage Re-ranking with BERT. Rodrigo Nogueira and Kyunghyun Cho. *arXiv 2019*. https://arxiv.org/abs/1901.04085. (2019)

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. *arXiv 2020*. https://arxiv.org/abs/1910.10683. (2020)

Document Ranking with a Pretrained Sequence-to-Sequence Model. Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. *arXiv 2020*. https://arxiv.org/abs/2003.06713. (2020)

What is UBC?

11.24: The University of British Columbia is a global centre for teaching, learning and research, consistently ranked among the top 20 public universities in the world and recently recognized as North America's most international university.

12.71: University of British Columbia Master of Data Science (MDS) is a 10-month, full-time, accelerated professional graduate program offered at both the University of British Columbia Vancouver and Okanagan campuses.

13.59: The University of British Columbia attracts, nurtures and proactively transform more than 58,000 students from Canada and 140 countries.

14.97: Vancouver is a major city in western Canada, located in the Lower Mainland region of British Columbia. As the most populous city in the province, the 2016 census recorded 631,486 people in the city, up from 603,502 in 2011. The Greater Vancouver area had a population of 2,463,431 in 2016, making it the third-largest metropolitan area in Canada.

16.76: R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

17.12: Python is a programming language that lets you work more quickly and integrate your systems more effectively. Python can be easy to pick up whether you're a first time programmer or you're experienced with other languages.

UBC MDS

10.27: University of British Columbia Master of Data Science (MDS) is a 10-month, full-time, accelerated professional graduate program offered at both the University of British Columbia Vancouver and Okanagan campuses.

11.85: The University of British Columbia is a global centre for teaching, learning and research, consistently ranked among the top 20 public universities in the world and recently recognized as North America's most international university.

13.93: The University of British Columbia attracts, nurtures and proactively transform more than 58,000 students from Canada and 140 countries.

15.84: Vancouver is a major city in western Canada, located in the Lower Mainland region of British Columbia. As the most populous city in the province, the 2016 census recorded 631,486 people in the city, up from 603,502 in 2011. The Greater Vancouver area had a population of 2,463,431 in 2016, making it the third-largest metropolitan area in Canada.

16.16: R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

17.07: Python is a programming language that lets you work more quickly and integrate your systems more effectively. Python can be easy to pick up whether you're a first time programmer or you're experienced with other languages.

best programming language

14.23: R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

14.43: Python is a programming language that lets you work more quickly and integrate your systems more effectively. Python can be easy to pick up whether you're a first time programmer or you're experienced with other languages.

16.95: University of British Columbia Master of Data Science (MDS) is a 10-month, full-time, accelerated professional graduate program offered at both the University of British Columbia Vancouver and Okanagan campuses.

17.38: The University of British Columbia is a global centre for teaching, learning and research, consistently ranked among the top 20 public universities in the world and recently recognized as North America's most international university.

17.56: The University of British Columbia attracts, nurtures and proactively transform more than 58,000 students from Canada and 140 countries.

17.77: Vancouver is a major city in western Canada, located in the Lower Mainland region of British Columbia. As the most populous city in the province, the 2016 census recorded 631,486 people in the city, up from 603,502 in 2011. The Greater Vancouver area had a population of 2,463,431 in 2016, making it the third-largest metropolitan area in Canada.

Experiments & Results

1. (Distance-based) No training

2. (Classification-based) Train end-to-end (prone to overfitting)

3. (Classification-based) Freeze transformer, only train two-layer MLP

4. (Distance-based) `TripletMarginLoss` (margin=1.0, distance=L2) (good)

5. (Distance-based) `TripletMarginLoss` (margin=0.1, distance=L2) (good)

6. (Distance-based) `TripletMarginLoss` (margin=1.0, distance=Cosine)

7. (Distance-based) `DistilBertForMaskedLM` (average of last layer) (good)