STIS at the NTCIR-16 Data Search 2 Task: Ad-hoc Data Retrieval Ranking with Pretrained Representative Words Prediction

Lya Hulliyyatus Suadaa
Department of Statistical Computing
Politeknik Statistika STIS
lya@stis.ac.id

Lutfi Rahmatuti Maghfiroh
Department of Statistical Computing
Politeknik Statistika STIS
lutfirm@stis.ac.id

Muhammad Luqman
Department of Statistical Computing
Politeknik Statistika STIS
luqman@stis.ac.id

Isfan Nur Fauzi
Department of Statistical Computing
Politeknik Statistika STIS
isfan@stis.ac.id

ABSTRACT

In this paper, we present the system and results of The STIS team for the Information Retrieval (English) subtasks of the NTCIR-16 Data Search 2 Task. The data collections in this task consist of a pair of metadata and a set of data files. We only used title, description, and tags of metadata as input documents of our proposed approach to retrieve a rank of query-related data files. We proposed using a pre-trained model to capture representative words prediction for each document then calculate the similarity between the query and the representative words as a rank score.

TEAM NAME

STIS

SUBTASKS

IR Subtask (English)

KEYWORDS

data retrieval, pre-trained model, representative word prediction

1 INTRODUCTION

The Open Government Data (OGD) policies encourage the distribution of government datasets to be explored by the citizen to solve any tasks. More than half of datasets in Google Dataset Search, one of the largest dataset search engines on the Web, arrive from the US Government Open Data portal (data.gov) [1]. In dataset search, the users submit their information needs using a query then the results are retrieved based on the similarity of the query to the metadata published about the datasets [2].

The STIS team participated in the Information Retrieval (English) subtasks of the NTCIR-16 Data Search 2 Task [4]. In this subtask, we have to generate a ranked list of statistical datasets for each query from data collections of the US government (data.gov). We explored metadata of the datasets consisting of title, description, and tags as input documents of our retrieval approach. Following Ma et al. [5], we proposed using a pre-trained Transformer model to capture representative words prediction for each document. The similarity between the query and the representative words was calculated as a rank score of the candidate retrieved dataset.

2 DATASETS

The datasets used in this paper consist of:

(1) Dataset collections
The collections were published by the US government[1] containing 46,615 datasets with 92,930 data files in various formats such as Excel (i.e., xls and xlsx), CSV, JSON, XML, RDF, PDF, and text files. Each dataset was supplemented by metadata that describe data information, including title, description, data format, data url, created date, and tags.
(2) Queries
Information needs were extracted from questions posted in community question answering such as "What are the largest causes of death in the United States in 1999-2016?". Queries for answering information needs were collected by asking ten crowdsourcing workers. The query example of previous information needs is "causes of death us 1999-2016".

3 PRETRAINED REPRESENTATIVE WORDS PREDICTION

The representative words idea is inspired by a query likelihood model that ranks the documents based on the relationship between a query and the document contents. A query consists of terms that are likely to appear in documents representing the representative words that discriminate the ideal documents from others.

Pre-trained models have been successfully applied in many downstream natural language processing tasks, including information retrieval. We implemented Pre-training with Representative wOrds Prediction (PROP) for ad-hoc retrieval [5] in constructing the representative words prediction for each document. The architecture of the proposed ranking approach with pretrained representative words prediction can be seen in Figure 1.

3.1 Representative Word Sets Sampling

We sample a pair of representative word sets from vocabulary $V = \{w_i\}_n^1$ based on document language model following dirichlet distribution with probability $P(w_i|D)$ for word w_i and document D. Query likelihood score function $QL(w_i, D)$ were calculated to each

[1] https://data.gov

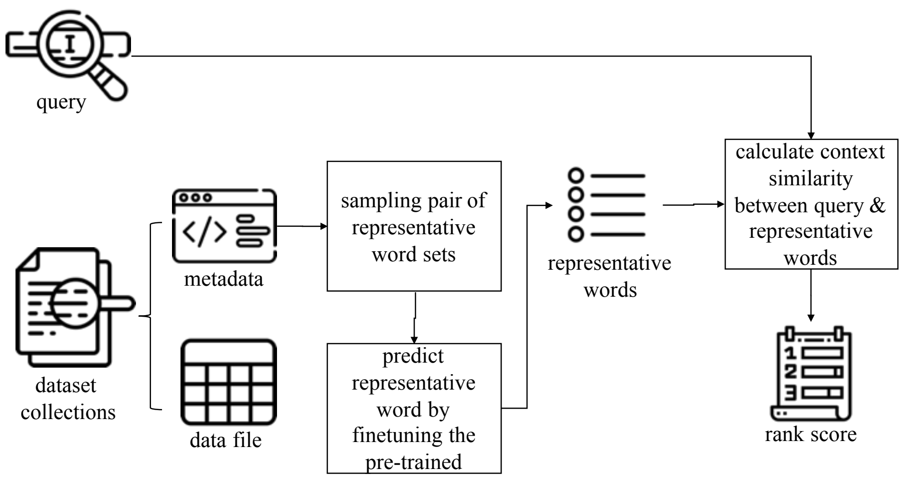

Figure 1: Proposed ranking architecture with pretrained representative words prediction.

word sets (S_u for the first word sample score and S_v for the second), resulting positive w_{pos} and negative words w_{neg} as follows:

$$\begin{aligned} \text{if } QL_u > QL_v \text{ then } w_{pos} = u \text{ and } w_{neg} = v \\ \text{else } w_{pos} = v \text{ and } w_{neg} = u \end{aligned} \tag{1}$$

The higher the query likelihood score, the more representative the words to the documents.

3.2 Representative Words Prediction

Using pairs of positive and negative words from the previous step, we finetune pre-trained Transformer BERT for representative words prediction task. Adopting the finetuned BERT approach in question answering task [3], we preprocess positive and negative word tokens and document tokens as input by inserting two special tokens, [CLS] and [SEP]. The [CLS] token is added to the beginning of input, and the [SEP] token is inserted after the query token to separate the representative words and document segments. The hidden state was obtained as follow,

$$h_{[CLS]} = \text{Transformer}([CLS] + w_{rep} + [SEP] + D + [SEP]), \tag{2}$$

where $w_{rep} = \{w_{pos}, w_{neg}\}$. Then, we compute the probability of word represent to the the document, as follows:

$$P(w_{rep}|D) = \text{MLP}(h_{[CLS]}). \tag{3}$$

We used hinge loss function for a pairwise loss of positive and negative word representation, as follows:

$$\mathcal{L} = \max(0, 1 - P(w_{pos}|D) + P(w_{neg}|D)). \tag{4}$$

3.3 Rank Score Calculation

The k list of representative word prediction ($\{w_{prop}\}_1^k$) from the previous finetuned model were used as representative words of documents. We obtain the rank score by calculating the context similarity between the query and $\{w_{prop}\}_1^k$. We used $BERT_{Score}$ [6] which takes advantage of the pre-trained contextual embeddings from BERT and computes the similarity of words in query and documents by cosine similarity.

$$S_{prop} = \text{avg}(BERT_{Score}(query, w_{prop_i})), i = 1...k \tag{5}$$

As the second model, we also try to combine the base score of traditional information retrieval model (e.g. BM25) and S_{prop} as follows:

$$S = (1 - \alpha)S_{base} + \alpha S_{prop}, \tag{6}$$

where α is the weight of the relevance score using representative words.

4 RESULTS

The overall performances of our proposed ranking approach using pre-trained representative words prediction (PROP) and $BERT_{Score}$ are shown in Table 1. We can see that our ranking mechanism using the Finetuned BERT of representative words predictor (PROP) were not outperformed the traditional information retrieval of BM25. However, a slightly better nDCG@10 score of the re-ranking mechanism could be a good sign of the representative words prediction effect.

The samples of representative words prediction using the finetuned model in NTCIR-16 dataset were shown in Table 2.

5 CONCLUSIONS

In this paper, we proposed an ad-hoc retrieval approach for governmental statistical data. We used metadata of data files as document features consisting of title, description and tags. We proposed using a pre-trained model to capture representative words prediction for each document then calculate the similarity between the query and the representative words as a rank score. We also combined the representative similarity score to re-rank candidate documents of BM25 model for each query.

REFERENCES

[1] Omar Benjelloun, Shiyu Chen, and Natasha Noy. 2020. Google Dataset Search by the Numbers. arXiv:2006.06894 [cs.IR]

Model	nDCG@3	nDCG@5	nDCG@10	nERR@3	nERR@5	nERR@10	Q-mea
BM25	**0.191**	**0.188**	**0.211**	**0.199**	**0.222**	**0.233**	**0.248**
Ranking by PROP and BertScore	0.172	0.175	0.201	0.191	0.188	0.201	0.218
BM25 + Reranking by PROP and BertScore	0.163	0.173	0.202	0.192	0.183	0.200	0.221

Table 1: Results from NTCIR-16 Data Search 2. The best score is in bold.

Query	Representative Words	$Rank_{BM25}$	$Rank_{PROP}$
causes of death us 1999-2016	x county cause numerous rate, new census, update scientifically deaths deaths, death estimating internationally, direction mortality death poisoning low ages rates, base longer meets deaths, drug death computer selected, estimated coded, affect states death poisoning, pending ratings x poisoned	6	1
	deaths finalization deaths, provisional death provisional comparisons, classifications updated, deaths reported death deaths causative categories, specifically cause, drug, counting drugs provisional, drug delay provisional, drug x updates vital, drug numbered pending drug	9	2
	rated rate x nchs differing, rated, rating census causes estimate number baseline, poisoned base acquisition, references death, poisoned drug includes, update cdcs adjusted, www published demographic, x rating ageing, death defined wonder	1	8
annual turnover care workers	disabled, incoming, benefits statistics, insurance social, socially calculations workers, securing three, three, receiving edition calculate, disability series people disability tables workers, peoples social statistical	56	1
	low researchers, teams v u shorebird shores forage, forest v tennessee foraging group, radio forest primary radio migrate, ponds estimating migratory shorebird, opening vulnerable estimated network rates, calidris, western flight shorebird reservation created migration, establishing rate, build regional valleys	1	35

Table 2: Samples of representative words prediction for each query and their ranks.

[2] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis-Daniel Ibáñez, Emilia Kacprzak, and Paul Groth. 2020. Dataset search: a survey. In *The VLDB Journal 29*. 251–272. https://doi.org/10.1007/s00778-019-00564-x

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[4] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2022. Overview of the NTCIR-16 Data Search 2 Task. In *NTCIR-16*.

[5] Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021. PROP: Pre-Training with Representative Words Prediction for Ad-Hoc Retrieval. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 283–291. https://doi.org/10.1145/3437963.3441777

[6] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr