SLOLQ at the NTCIR-13 OpenLiveQ Task

Ryo Kashimura
Waseda University, Japan
dqlover2352@akane.waseda.jp

Tetsuya Sakai
Waseda University, Japan
tetsuyasakai@acm.org

ABSTRACT

The SLOLQ (Sakai Laboratory OpenLiveQ) team submitted six runs to the Offline Test of the NTCIR-13 OpenLiveQ Task, including a similarity ranking run and a diversity ranking run. Subsequently, our similarity ranking run was evaluated in the Online Test. Unfortunately, our offine results show that our Similarity Ranking and Diversity Ranking runs are statistically indistinguishable from those that rank questions at random. Our online results show that our Similarity Ranking run failed to outperform a baseline that simply ranks questions by the number of answers they received.

Team Name

SLOLQ

Keywords

community question answering; Doc2Vec; question answering

1. INTRODUCTION

The SLOLQ (Sakai Laboratory OpenLiveQ) team submitted six runs to the Offline Test of the NTCIR-13 OpenLiveQ Task, including a similarity ranking run and a diversity ranking run. Our runs utilise Doc2Vec [3] to generate a question vector that represents a given question. Subsequently, our similarity ranking run was evaluated in the Online Test. Unfortunately, our offine results show that our Similarity Ranking and Diversity Ranking runs are statistically indistinguishable from those that rank questions at random. Our online results show that our Similarity Ranking run failed to outperform a baseline that simply ranks questions by the number of answers they received.

2. RELATED WORK

Our runs utilise Doc2Vec proposed by Mikolov et al. [3], which generates a vector representation of a given document based on word vectors of the words contained in the document. While Doc2Vec accommodates two different network models, namely, PV-DM (Distributed Memory Paragraph Vectors) and PV-DBOW (Distributed Bag of Word Paragraph Vectors), we chose to use the PV-DM model as it can perform better than the PV-DBOW model [3]. Below, a brief description of PV-DM is provided.

Suppose that we have a document D_i and a context C that consists of a sequence of $2k+1$ words in the document. Hence the context can be denoted as $C = [w_{t-k}, \cdots, w_{t+k}]$. The PV-DM algorithm determines each word vector so that the probability that the next word w_{t+k+1} cooccurs with the context C is maximized. The IDs of the documents as well as those of words in the document are inputs to the network, and the IDs of the documents are also learned. Figure 1 shows the network of the PV-DM model.

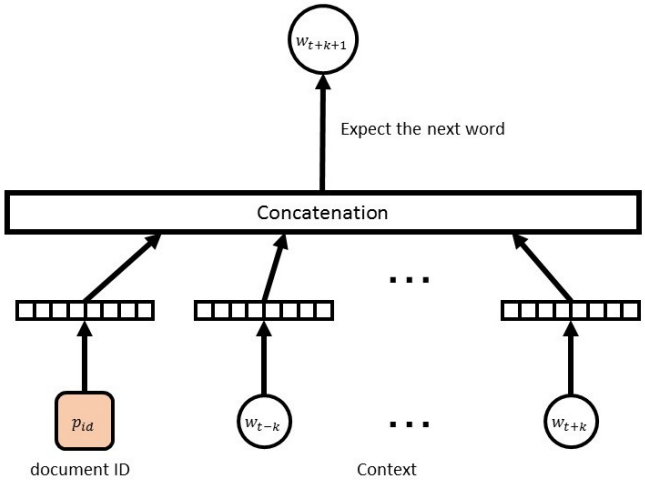

Figure 1: Doc2Vec: PV-DM model

3. PROPOSED METHODS

A concise description of the OpenLiveQ task would be: given a search query q, rerank the questions in the set D_q in descending order of estimated click counts. Our approach is conducted by two steps. First, we generate question vectors for each question in D_q using Doc2Vec. Next, we compute the simlarities among question vectors and use them for ranking the questions.

3.1 Question Vector

A Question Vector is a vector expression generated by Doc2Vec. The inputs to the Doc2Vec network are a question ID as a document ID, and the words from the snippet of that question as its context. Note that the users in the OpenLiveQ Online Test are also shown not only the question titles but also the snippets; our assumption was that if the training data resembles what is shown to the users during the Online Test, then it would be easy for the system to predict whether each question will receive many clicks.

3.2 Question Ranking Algorithm

Having generated the question vectors, we next rank the questions based on the similarities of the question vectors. First, we select the most viewed question from D_q and put it at the top of our ranked list: the assumtion is that frequently viewed items are likely to be frequently clicked. Then the remaining questions are ranked by either Similarity Ranking or Diversity Ranking, as described below.

3.2.1 Similarity Ranking

Our first strategy is to rank the questions based on the similarity with the most viewed question, where the similarity is computed based on the Doc2Vec question vectors. Algorithm 1 provides the pseudocode for this approach. Here, "$[\boldsymbol{v}]$" denotes a list containing one element, namely, the Doc2Vec vector $\boldsymbol{v}$, and the oparator "+" between two lists denotes a concatenation of the lists.

Algorithm 1 Similarity Ranking

Input: V_q(a set of question vectors of questions with respect to a query q),
${\boldsymbol{v}_q}^{(1)} \in V_q$(the question vector of the most viewed question)
Output: R_q
$R_q \leftarrow \left[{\boldsymbol{v}_q}^{(1)}\right]$
$V_q \leftarrow V_q \setminus \left\{{\boldsymbol{v}_q}^{(1)}\right\}$
while $V_q \neq \emptyset$ **do**
  $\boldsymbol{v} \leftarrow \underset{\boldsymbol{v}_q \in V_q}{\operatorname{argmax}} \operatorname{cosine}({\boldsymbol{v}_q}^{(1)}, \boldsymbol{v}_q)$
  $R_q \leftarrow R_q + [\boldsymbol{v}]$
  $V_q \leftarrow V_q \setminus \{\boldsymbol{v}\}$
end while
return R_q

3.2.2 Diversity Ranking

Our second strategy aims to diversify the question ranking, based on the observation that Similarity Ranking may put many similar questions close to one another and may not accurately reflect the ranking based on click counts. Hence, this strategy chooses a question that is most different from the most recently chosen one according to the Doc2Vec question vectors. Algorithm 2 provides the pseudocode for this approach.

Note that Diversity Ranking considers the most recently chosen question when choosing the next one, rather than the entire set of already chosen questions. We did not have time to submit a run based on a Maximal Marginal Relevance approach [1].

Algorithm 2 Diversity Ranking

Input: V_q(a set of question vectors of questions with respect to a query q),
${\boldsymbol{v}_q}^{(1)} \in V_q$(the question vector of the most viewed question)
Output: R_q
$R_q \leftarrow \left[{\boldsymbol{v}_q}^{(1)}\right]$
$V_q \leftarrow V_q \setminus \left\{{\boldsymbol{v}_q}^{(1)}\right\}$
$\boldsymbol{v}_{prev} \leftarrow {\boldsymbol{v}_q}^{(1)}$
while $V_q \neq \emptyset$ **do**
  $\boldsymbol{v} \leftarrow \underset{\boldsymbol{v}_q \in V_q}{\operatorname{argmin}} \operatorname{cosine}({\boldsymbol{v}_q}^{(1)}, \boldsymbol{v}_q)$
  $R_q \leftarrow R_q + [\boldsymbol{v}]$
  $V_q \leftarrow V_q \setminus \{\boldsymbol{v}\}$
  $\boldsymbol{v}_{prev} \leftarrow \boldsymbol{v}$
end while
return R_q

4. RESULTS

4.1 Offline Test

We submitted six runs to the Offline Test. In addition to the Similarity Ranking and Diversity Ranking runs, we submitted four baselines, three of which simply ranked the questions at random (Random1, Random2, Random 3). The fourth baseline simply ranked the questions in descending order of page view (PageView). Tables 1, 2, and 3 show the mean, maximum, and minium scores of each run in terms of nDCG@10, ERR@10, and Q-measure [4], respectively. The scores of the three random baselines are averaged here.

To compare the means of our six runs from a statistical viewpoint, we conducted randomised Tukey HSD tests with $B = 10,000$ trials using the `Discpower` toolkit[1]. For each run pair, we also computed the sample effect size (i.e., standardized mean difference) [5]. Tables 4, 5 and 6 show the p-value and the effect size for each run pair and for each evaluation measure. It can be observed that the p-values are either 1 or very close to one, and therefore that our runs are statistically indistinguishable from one another.

4.2 Online Test

As our Similarity Ranking run achieved the highest score in terms of mean nDCG@10 in the Offline Test, only this run was evaluated in the Online Test. In the Online Test, each run receives "credits" based on actual clicks by the users [2]. Table 7 compares our credit statistics with the Baselines provided by the organizers. The sum and the mean credits as well as per-query maximum and minimum credits are shown. It can be observed that our run failed to outperform the baseline that simply ranks questions by the number of answers they received.

5. CONCLUSIONS

Unfortunately, our offine results showed that our Similarity Ranking and Diversity Ranking runs are statistically indistinguishable from those that rank questions at random. Our online results show that our Similarity Ranking run failed to outperform a baseline that simply ranks questions by the number of answers they received.

One possible reason for our lack of success is that we only fed questions IDs, question titles and snippets to our Doc2Vec model. If additional data such as the body of each question, its best answer, and clickthrough data are utilized, this may help us build better models.

6. REFERENCES

[1] J. Goldstein and J. Carbonell. Summarization: (1) using MMR for diversity - based reranking and (2)

[1] http://research.nii.ac.jp/ntcir/tools/discpower-en.html

Table 1: nDCG@10 of each run in offline tests

Run		Random	Page View	Similarity Ranking	Diversity Ranking
nDCG@10	Mean	0.31010	0.31384	0.31908	0.31329
	Max	0.61447	0.59845	0.68306	0.65769
	Min	0.04029	0.04310	0.01946	0.04620

Table 2: ERR@10 of each run in offline tests

Run		Random	Page View	Similarity Ranking	Diversity Ranking
ERR@10	Mean	0.17987	0.18779	0.19760	0.20352
	Max	0.84895	0.97687	0.97715	0.97482
	Min	0.01567	0.01378	0.00781	0.00928

Table 3: Q-measure of each run in offline tests

Run		Random	Page View	Similarity Ranking	Diversity Ranking
Q-measure	Mean	0.65215	0.65497	0.65563	0.64613
	Max	0.84287	0.84088	0.84026	0.83140
	Min	0.39900	0.39308	0.32029	0.37863

Table 4: p-value/effect size with nDCG@10

	Random2	Random3	Page View	Similarity Ranking	Diversity Ranking
Random1	0.9342 / 0.1347	1 / 0.0194	1 / 0.0044	0.9963 / 0.0642	1 / 0.0020
Random2		0.8851 / 0.1541	0.9259 / 0.1391	0.7160 / 0.1989	0.9376 / 0.1328
Random3			1 / 0.0150	0.9994 / 0.0448	1 / 0.0214
Page View				0.9975 / 0.0598	1 / 0.0063
Similarity Ranking					0.9959 / 0.0662

Table 5: p-value/effect size with ERR@10

	Random2	Random3	Page View	Similarity Ranking	Diversity Ranking
Random1	1 / 0.0492	1 / 0.0233	1 / 0.0402	0.9912 / 0.1199	0.9391 / 0.1679
Random2		1 / 0.0258	0.9987 / 0.0894	0.9373 / 0.1691	0.8825 / 0.2171
Random3			1 / 0.0635	0.9740 / 0.1432	0.8944 / 0.1912
Page View				0.9994 / 0.0797	0.9853 / 0.1277
Similarity Ranking				/	1 / 0.0481

Table 6: p-value/effect size with Q-measure

	Random2	Random3	Page View	Similarity Ranking	Diversity Ranking
Random1	1 / 0.1210	1 / 0.0451	1 / 0.0782	1 / 0.1209	1 / 0.4948
Random2		1 / 0.1662	1 / 0.1993	1 / 0.2420	1 / 0.3737
Random3			1 / 0.0330	1 / 0.0758	1 / 0.5399
Page View				1 / 0.0427	1 / 0.5730
Similarity Ranking					1 / 0.6157

Table 7: Credits of Similarity Ranking in the online test

Run		Baseline (# of Answers)	Baseline (as is)	Similarity Ranking
Sum		18917.55	14037.08	14892.03
per query	Mean	19.50	14.47	15.35
	Max	302.39	679.48	748.59
	Min	0.00	0.00	0.00

evaluating summaries. In *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, pages 181–195, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[2] M. P. Kato, T. Yamamoto, T. Manabe, A. Nishida, and S. Fujita. Overview of the NTCIR-13 Open Live Q Task. In *Proceedings of NTCIR-13*, 2017.

[3] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

[4] T. Sakai. Metrics, statistics, tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*, pages 116–163, 2014.

[5] T. Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, June 2014.