SLWWW at the NTCIR-13 WWW Task

Peng Xiao
Waseda University, Japan
xp1994@fuji.waseda.jp

Lingtao Li
Waseda University, Japan
lilingtao@fuji.waseda.jp

Yimeng Fan
Waseda University, Japan
723603536@fuji.waseda.jp

Tetsuya Sakai
Waseda University, Japan
tetsuyasakai@acm.org

ABSTRACT

SLWWW participated in the Chinese Subtask of the NTCIR-13 WWW Task. We applied the query expansion methods based on word embeddings proposed by Kuzi, Shtok, and Kurland. However, according to our comparison with the baseline run, our runs were not successful. As the baseline run provided by the organisers was not included in the pools for constructing relevance assessments, we discuss condensed-list versions of the official evaluation measures in addition to the regular measures.

Team Name

SLWWW

Subtasks

Chinese subtask

Keywords

short text conversation; weibo; word2vec

1. INTRODUCTION

SLWWW participated in the Chinese Subtask of the NTCIR-13 WWW Task [2]. We applied the query expansion methods proposed by Kuzi, Shtok, and Kurland [1]. These methods rely on *word embeddings* [3].

2. RUN DESCRIPTIONS

Table 1 summarises the approaches used in our four Chinese subtask runs. All of them are based on query expansion methods proposed by Kuzi, Shtok, and Kurland [1]. Below, we follow the notations from their paper.

2.1 Term Scoring

We tried two term scoring methods from Kuzi, Shtok, and Kurland [1] to select expansion terms from the corpus, namely, the *centroid method* and *CombMAX*.

For query q, let q_i denote its i-th query term. The centroid method represents q by:

$$\vec{q}_{Cent} \overset{\text{def}}{=} \sum_{q_i \in q} \vec{q}_i \ . \tag{1}$$

Let $\vec{t}$ denote the L_2-normalised Word2Vec vector [3] representing term t. The selection score for term t in the corpus is:

$$S_{Cent}(t;q) \overset{\text{def}}{=} \exp(\cos(\vec{t}, \vec{q}_{Cent})) \ . \tag{2}$$

In contrast to the centroid method which scores terms based on the similarity with the query as a whole, the CombMAX method selects n most similar terms for each q_i according to $\cos(\vec{q}_i, t)$. Let L_{q_i} be the list of terms for q_i. For each term in L_{q_i}, the n similarities are softmax-normalised:

$$p(t|q_i) \overset{\text{def}}{=} \frac{\exp(\cos(\vec{q}_i, \vec{t}))}{\sum_{t' \in L_{q_i}} \exp(\cos(\vec{q}_i, \vec{t'}))} \ . \tag{3}$$

We let $n = 5$ in our experiments. Finally, the lists for q_i are fused:

$$S_{CombMAX}(t;q) \overset{\text{def}}{=} \max_{q_i \in q} p(t|q_i) \ . \tag{4}$$

2.2 Query Expansion

We select $\nu = 3, 5$ terms according to the term selection scores $S_{\mathcal{M}}(t;q)$ ($\mathcal{M} \in \{Cent, CombMAX\}$); we then sum-normalise the scores to obtain $p(t|\mathcal{M})$, a probability distribution over the corpus vocabulary. Zero probabilites are assigned to terms other than the selected ones. The above unigram language model is then integrated with a language model induced from q. More specifically, the maximum likelihood estimate of term t with respect to q is obtained as:

$$p_{MLE}(t|q) \overset{\text{def}}{=} \frac{tf(t \in q)}{|q|} \ , \tag{5}$$

where $tf(t \in q)$ is the count of t in q. The integrated model is given by:

$$p(t|\mathcal{M}, q) \overset{\text{def}}{=} (1-\lambda)p(t|\mathcal{M}) + \lambda p_{MLE}(t|q) \ . \tag{6}$$

We let $\lambda = 0.5$ in our experiments.

3. RESULTS AND DISCUSSIONS

Table 2 shows our official results, together with results for the baseline run provided by the organisers. Note that the baseline run was not included in the relevance assessment pools and therefore the effectiveness of the baseline run is underestimated here. There are 4.39 unjudged documents on average in the top 10 baseline results for each topic; that is, a total of 439 unjudged documents across the 100 topics. In particular, the Mean Q score for `baseline` is lower than those for the other four runs, reflecting the fact that Q-measure is more recall-oriented than nDCG and nERR [5].

For Topic 0099, Table 3 illustrates how MSnDCG@10, Q@10, and nERR@10 are computed for `baseline` which was not involved in the pooling process. The left side computes the intermediate scores for the system's ranked list, while the right side computes those for the ideal list. On the left

Table 1: Run descriptions

Run name	Term scoring method	#Expansion terms
SLWWW-C-NU-Base-1	Centroid	3
SLWWW-C-NU-Base-2	CombMAX	3
SLWWW-C-NU-Base-3	CombMAX	5
SLWWW-C-NU-Base-4	Centroid	5

Table 2: Official results and the baseline results. The highest mean score in each column is indicated in bold. Note that the baseline run was not pooled and there are a 4.39 unjudged documents on average in the top 10 baseline results for each topic. Hence we are underestimating the baseline performance. Each $*x$ indicates that the run is statistically significantly better than SLWWW-C-NU-Base-x according to the randomised Tukey HSD test.

run	Mean MSnDCG@10	Mean Q@10	Mean nERR@10
SLWWW-C-NU-Base-1	0.3206	0.3094	0.4753
SLWWW-C-NU-Base-2	0.3225	**0.3099**	0.4723
SLWWW-C-NU-Base-3	0.2909	0.2838	0.4327
SLWWW-C-NU-Base-4	0.2991	0.2949	0.4406
baseline	**0.3235**	0.2522	**0.5341**^{*3,*4}

Table 3: How MSnDCG@10, Q@10 and nERR@10 are computed for baseline with Topic 0099.

	System's ranked list					Ideal ranked list					
r	relevance level	$cg(r)$	$dg(r)$	$dsat(r-1)$ $*Pr(r)$	$dsat(r-1)$ $*Pr(r)\frac{1}{r}$	relevance level*	$cg^*(r)$	$dg^*(r)$	$dsat^*(r-1)$ $*Pr^*(r)$	$dsat^*(r-1)$ $*Pr^*(r)\frac{1}{r}$	$I(r)$ $*BR(r)$
1	$L9$	9	9	0.9	0.9	$L9$	9	9.0000	0.9000	0.9000	1
2	$L0$	9	0	0	0	$L9$	18	5.6784	0.0900	0.0450	0
3	unjudged	9	0	0	0	$L9$	27	4.5000	0.0090	0.0030	0
4	$L0$	9	0	0	0	$L9$	36	3.8761	0.0009	0.0002	0
5	unjudged	9	0	0	0	$L9$	45	3.4817	0.0001	0.0000	0
6	unjudged	9	0	0	0	$L9$	54	3.2059	0.0000	0.0000	0
7	unjudged	9	0	0	0	$L9$	63	3.0000	0.0000	0.0000	0
8	$L0$	9	0	0	0	$L9$	72	2.8392	0.0000	0.0000	0
9	unjudged	9	0	0	0	$L9$	81	2.7093	0.0000	0.0000	0
10	$L0$	9	0	0	0	$L9$	90	2.6016	0.0000	0.0000	0
SUM			9		0.9			40.8920		0.9482	1

side, $cg(r)$ denotes the *cumulative gain* at rank r, used by Q@10; $dg(r)$ denotes the *discounted gain* at rank r, used by MSnDCG@10; $dsat(r-1)$ denotes the probability that the user was not satisfied with the top $(r-1)$ documents, and $Pr(r)$ denotes the probability that the user is satisfied with the document at rank r, used by nERR@10. On the right side, there are corresponding notations for the ideal ranked list. Finally, in the rightmost column, $I(r)$ is a flag where $I(r)=1$ if the document at r is relevant; $I(r)=0$ otherwise; $BR(r)$ is the *blended ratio* at rank r, given by:

$$BR(r) = \frac{c(r)+cg(r)}{r+cg^*(r)} \ , \qquad (7)$$

where $c(r)$ is the number of relevant documents in top r. See Sakai [5] for more details.

From Table 3, it can be observed that there are five unjudged documents in the top 10 documents returned by baseline for Topic 0099. The Q@10 for this topic is $BR(1)/10 = ((c(1)+cg(1))/(1+cg^*(1)))/10 = ((1+1)/(1+1))/10 = 0.1000$. Whereas, the MSnDCG@10 is $\sum_r dg(r)/\sum_r dg^*(r) = 9/40.8920 = 0.2201$; nERR@10 is $\frac{\sum_r dsat(r-1)Pr(r)/r}{\sum_r dsat^*(r-1)Pr^*(r)/r} = 0.9/0.9482 = 0.9491$.

For each evaluation measure, we conducted a randomised Tukey HSD test for each 100×5 topic-by-run matrix using Discpower with $B = 10,000$ trials [5][1]. In terms of Mean nERR@10, baseline statistically significantly outperforms SLWWW-C-NU-Base-3 ($p = 0.0022$) and SLWWW-C-NU-Base-4 ($p = 0.0067$) even though we are underestimating the effectiveness of baseline; all other pairwise differences are not statistically significant. Hence, our query expansion experiments are not successful.

Because the baseline run contains many unjudged documents, we tried evaluating the five runs using *condensed-list* measures [4]. A condensed-list measure removes all unjudged documents from the ranked list before evaluating it. The reader is referred to Sakai [5] for a summary of the advantages of condensed-list measures over other measures that were designed specifically for handling incomplete relevance assessments.

NTCIREVAL can compute condensed-list measures[2]. However, it requires at least one judged relevant ($L0$) document as this is required for computing *bpref* (but not for computing the other condensed-list measures). It turns out that Topic 0033 did not have any $L0$ documents; hence we removed this topic and computed the condensed-list measures for the remaining 99 topics.

Table 4 shows the regular evaluation measures averaged over the 99 topics; Table 5 shows the corresponding condensed-list measure scores, denoted by MSnDCG$'$@10, etc. Note that the scores are identical for the four submitted runs, since for these runs there are no unjudged documents in the top 10 documents for any of the topics. Only the effectiveness scores of the unpooled baseline run are boosted, as

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

[2] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

Table 4: Official results and the baseline results after removing Topic 0033 which lacks judged nonrelevant documents. The highest mean score in each column is indicated in bold. Each $*x$ indicates that the run is statistically significantly better than `SLWWW-C-NU-Base-`x according to the randomised Tukey HSD test.

run	Mean MSnDCG@10	Mean Q@10	Mean nERR@10
`SLWWW-C-NU-Base-1`	0.3177	0.3058	0.4715
`SLWWW-C-NU-Base-2`	0.3196	**0.3064**	0.4686
`SLWWW-C-NU-Base-3`	0.2878	0.2799	0.4285
`SLWWW-C-NU-Base-4`	0.2961	0.2912	0.4365
`baseline`	**0.3208**	0.2486	**0.5311**^{*3,*4}

Table 5: Condensed-list measure scores after removing Topic 0033 which lacks judged nonrelevant documents. The highest mean score in each column is indicated in bold. Each $*x$ indicates that the run is statistically significantly better than `SLWWW-C-NU-Base-`x according to the randomised Tukey HSD test.

run	Mean MSnDCG@10	Mean Q@10	Mean nERR@10
`SLWWW-C-NU-Base-1`	0.3177	0.3058	0.4715
`SLWWW-C-NU-Base-2`	0.3196	0.3064	0.4686
`SLWWW-C-NU-Base-3`	0.2878	0.2799	0.4285
`SLWWW-C-NU-Base-4`	0.2961	0.2912	0.4365
`baseline`	**0.5259**^{*1*2*3,*4}	**0.5068**^{*1*2*3,*4}	**0.6999**^{*1*2*3,*4}

Table 6: How the condensed-list measures MSnDCG′@10, Q′@10 and nERR′@10 are computed for `baseline` with Topic 0099. The relevance levels of documents that were promoted from beneath top 10 are indicated with ↑.

	System's ranked list					Ideal ranked list					
r	relevance level	$cg(r)$	$dg(r)$	$dsat(r-1)$ $*Pr(r)$	$dsat(r-1)$ $*Pr(r)\frac{1}{r}$	relevance level*	$cg^*(r)$	$dg^*(r)$	$dsat^*(r-1)$ $*Pr^*(r)$	$dsat^*(r-1)$ $*Pr^*(r)\frac{1}{r}$	$I(r)$ $*BR(r)$
1	$L9$	9	9	0.9	0.9	$L9$	9	9.0000	0.9000	0.9000	1
2	$L0$	9	0	0	0	$L9$	18	5.6784	0.0900	0.0450	0
3	$L0$	9	0	0	0	$L9$	27	4.5000	0.0090	0.0030	0
4	$L0$	9	0	0	0	$L9$	36	3.8761	0.0009	0.0002	0
5	$L0$	9	0	0	0	$L9$	45	3.4817	0.0001	0.0000	0
6	↑ $L0$	9	0	0	0	$L9$	54	3.2059	0.0000	0.0000	0
7	↑ $L0$	9	0	0	0	$L9$	63	3.0000	0.0000	0.0000	0
8	↑ $L0$	9	0	0	0	$L9$	72	2.8392	0.0000	0.0000	0
9	↑ $L3$	12	0.9031	0.03	0.0033	$L9$	81	2.7093	0.0000	0.0000	0.1556
10	↑ $L0$	12	0	0	0	$L9$	90	2.6016	0.0000	0.0000	0
SUM			9.9031		0.9033			40.8920		0.9482	1.1556

the removal of unjudged documents promotes the retrieved relevant documents that were beneath them. The true effectiveness scores for the baseline runs lie *somewhere between* those shown in Table 4 and those shown in Table 5.

We applied the randomised Tukey HSD test to each of the new 99×5 topic-by-run matrices. In Table 4, even after removing Topic 0099, `baseline` statistically significantly outperforms `SLWWW-C-NU-Base-3` ($p = 0.0022$) and `SLWWW-C-NU-Base-4` ($p = 0.0067$) in terms of Mean nERR@10. In Table 5, based on condensed lists, `baseline` statistically significantly outperforms all four submitted runs in terms of all three evaluation measures ($p \approx 0$).

Table 6 shows how the condensed-list measures are computed for `baseline` with Topic 0099, for comparison with Table 3. It can be observed that condensing the ranked list has promoted four new $L0$ documents from beneath the original top 10 documents, as well as one $L3$ document. The Q′@10 for this topic is $(BR(1) + BR(9))/10 = 0.1156$; The MSnDCG′@10 is $9.9031/40.8920 = 0.2422$; The nERR′@10 is $0.9033/0.9482 = 0.9526$.

4. CONCLUSIONS

SLWWW participated in the Chinese Subtask of the NTCIR-13 WWW Task. We applied the query expansion methods based on word embeddings proposed by Kuzi, Shtok, and Kurland. However, according to our comparison with the baseline run, our runs were not successful. More specifically, even though the official relevance assessments underestimate the baseline run without query expansion, `SLWWW-C-NU-Base-3` and `SLWWW-C-NU-Base-4` statistically significantly underforms the baseline in terms of Mean nERR@10. Moreover, when we evaluate all runs based on condensed-list measures, all four submitted runs statistically significantly underform the baseline, although condensed-list measures are known to overestimate the true effectiveness of unpooled runs such as the baseline.

5. REFERENCES

[1] S. Kuzi, A. Shtok, and O. Kurland. Query expansion using word embeddings. In *Proceedings of ACM SIGIR 2016*, pages 1929–1932, 2017.

[2] C. Luo, T. Sakai, Y. Liu, Z. Dou, C. Xiong, and J. Xu. Overview of the NTCIR-13 we want web task. In *Proceedings of NTCIR-13*, 2017.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *https://arxiv.org/abs/1301.3781*, 2013.

[4] T. Sakai. Alternatives to bpref. In *Proceedings of ACM SIGIR 2007*, pages 71–78, 2007.

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