How to Combine Translation Probabilities and Question Expansion for Question Classification in cQA Services

Kyoungman BAE^{†a)}, Student Member and Youngjoong KO^{†b)}, Nonmember

SUMMARY This paper claims to use a new question expansion method for question classification in cQA services. The input questions consist of only a question whereas training data do a pair of question and answer. Thus they cannot provide enough information for good classification in many cases. Since the answer is strongly associated with the input questions, we try to create a pseudo answer to expand each input question. Translation probabilities between questions and answers and a pseudo relevant feedback technique are used to generate the pseudo answer. As a result, we obtain the significant improved performances when two approaches are effectively combined.

key words: question classification, cQA service, pseudo relevant feedback (PRF), question expansion, translation probability

1. Introduction

The community-based Question Answering Service (cQA) allows users to ask just about any question and get answers from other users. A question classification is essential task in the cQA. The question classification is the task of automatically assigning unlabeled questions into predefined categories (or topics). Since input questions in the cQA service consist of a small size of texts, they commonly suffer from insufficient information to classify their category well. In this paper, we propose a new question expansion method for the question classification in the cQA service.

The most important task of the question expansion is to extract expansion words strongly associated with an input question [1], [2]. In general, the training data consists of the pairs of question and answer in the cQA service but the input of the cQA service does only a question. If the input question could have its answer, it would improve the performance of the question classification. Therefore, we try to create a pseudo answer using translation probabilities from questions to answers in training data to expand input question. It is based on an assumption that the answer is strongly associated with the input question.

Some studies have used translation probabilities to find relevance between two words in questions and answers [3], [4]. If the translation probabilities are utilized for question classification, words with high translation probabilities can be regarded as expansion words and a pseudo an-

swer can be generated by them. However, since the expansion words generated by the translation probabilities generally contain some noise words that are not related with the input question, they should be removed from the expansion words. A pseudo relevant feedback technique [5], which is a traditional query expansion method in information retrieval, is employed to remove these noise words effectively. Top ranked question-answer pairs to the input question are selected as relevant question-answer pairs from the Indri search engine and then the words from the relevant question-answer pairs are used to filter the noise words. That is, only the words in the relevant question-answer pairs are considered as the candidates of expansion words and they are sorted by a combination of their translation probabilities and relevant scores from the pseudo relevant feedback. Eventually, top ranked words, which have a high combined score, are added to an original question as a pseudo answer (or expansion words). The pair of an original input question and its pseudo answer is used as an input to the cQA service.

As a result, the proposed method improved the performance of question classification significantly. In our experiment, the proposed model achieved 89.9%, which is 7.0% higher performance than baseline, which is tested using only original input question.

This paper is organized as follows. Section 2 describes related work. In Sect. 3, our question expanding method for question classification is described in detail. In Sect. 4, we discuss the analysis of experimental results and discussion. Finally, we draw some conclusions in Sect. 5.

2. Related Work

There are a variety of studies for question classification. One is to exploit the feature weighting scheme [6], [7]. They employed Jensen-Shannon (JS) divergence and the frequency related with the question, category and inverse category. The other way is to prune the unrelated categories using the relevant data [8], [9]. They employed the subset of categories related to similar questions for the question classification. In addition, [9] expanded short questions by leveraging external Wikipedia to tackle the data sparseness problem. Translation-based retrieval and pseudo relevant feedback are exploited in the paper. Translation-based retrieval is the idea of finding similar questions in cQA using the translation probabilities [10]. The translation probability is very efficient for finding the relevance between two words. A lot of studies used the translation-based retrieval

Manuscript received June 23, 2015.

Manuscript revised September 15, 2015.

Manuscript publicized January 14, 2016.

[†]The authors are with Dong-A University, 840, Hadan 2-dong, Saha-gu, Busan, 604–714, Republic of Korea.

a) E-mail: kyoungman.bae@gmail.com

b) E-mail: youngjoong.ko@gmail.com (Corresponding author)

DOI: 10.1587/transinf.2015DAL0005

for finding the similar data with the input data [11], [12]. The pseudo relevant feedback has been shown to be an effective way of improving retrieval accuracy by reformulating an original query using pseudo-relevant documents from the initial retrieval result [13], [14].

3. Proposed Method

This study explores a novel question expansion method using the translation probability and the pseudo relevant feedback. The most important challenge for question expansion is to effectively construct the expansion words and we propose to extract them by the following steps in Fig. 1.

3.1 Finding Expansion Candidate Words (ECWs)

The ECWs are a set of words that can be translated from words in the input question. The first step is to find the ECWs using the averaged translation probabilities (ATP) from the translation probabilities using the *Giza++*[†]. In this approach, we estimate word-to-word translation probabilities by regarding questions as a source and answers as a target; the opposite case is also considered. The ATP between two words is calculated by averaging the probabilities of these two cases as follows:

$$ATP(w_t, w_s) = \frac{P_{QtoA}(w_t|w_s) + P_{AtoQ}(w_t|w_s)}{2} \quad (1)$$

where $P_{QtoA}(w_t|w_s)$ is the translation probability between from w_s in questions as a source to w_t in answers as a target, and $P_{AtoQ}(w_t|w_s)$ is the translation probability in the opposite case. A word w whose $ATP(w, qw)$ score is greater than zero is chosen as an expansion candidate word (ECW). qw is a word in the input question.

3.2 Extracting the RECWs from the ECWs Using the Pseudo Relevant Feedback

The ECWs commonly contains a lot of noisy words because the translation probability is estimated from the whole corpus. Thus we try to remove the noisy words in the ECWs using the pseudo relevant feedback and its result is called the refined expansion candidate words (RECWs). By using the

Indri^{††} search engine, we can obtain a set of top 20 ranked relevant question-answer pairs strongly related to each input question. We assume that words occurred in the top-ranked question-answer pairs are strongly associated to the input question. Thus, if the words in the ECWs are not occurred in the relevant question-answer pairs, they are regarded as noisy words. Eventually, the RECWs are constructed by removing the noisy words in the ECWs.

3.3 Ranking the RECWs Using the Combination of the Translation Score (TS) and the Relevant Score (RS)

Now, the importance scores of the RECWs are calculated by the linear combination of the translation score (TS) and the relevant score (RS) and then the RECWs are ranked according to their importance scores. The top-ranked words are selected as the final expansion words (EWs). Since TS and RS scores have different ranges, they have to be normalized by dividing them by the maximum score, respectively. Since some words in the RECWs can have translation probabilities with two or more words in an input question, the averaged TS score is calculated as the final translation score of each word and it is normalized by Eq. (3).

$$TS(Q, w) = \frac{1}{|Q|} \sum_{qw \in Q} ATP(w, qw) \quad (2)$$

$$NTS(Q, w) = \frac{TS(Q, w)}{\max\{TS(Q, t_i), t_i \in RECW, i = 1, \dots, n\}} \quad (3)$$

where Q is an input question, $|Q|$ denotes the length of the input question, $NTS(Q, w)$ is the normalized TS score, t_i is an RECW and n is the total number of the RECWs related Q .

The RS scores of words in the relevant question-answer pairs are calculated according to estimated important scores by reflecting the ratio of word occurrences in relevant and non-relevant question-answer pairs. If the RS score of a word is high, the word is considered as a relevant word of the input question. It is calculated using Robertson and Walker's relevance weights in Eq. (4) [15].

$$RS(Q, w) = r_w \log \frac{N}{n_w} - \log \left(\frac{R}{r_w} \right) - \log V \quad (4)$$

$$NRS(Q, w) = \frac{RS(Q, w)}{\max\{RS(Q, t_i), t_i \in RECW, t = i, \dots, n\}} \quad (5)$$

where $RS(Q, w)$ is the relevance score of word w , R is the total number of top-ranked relevant question-answer pairs, r_w is the number of top-ranked relevant question-answer pairs in which the word w occurs, N is the size of the collection, n_w is the number of question-answer pairs that contain the word w and V is the size of the vocabulary. The argument of the second logarithm is the number of ways one can choose

Fig. 1 The overview of our model

[†]GIZA++ is part of the SMT toolkit EGYPT. (<http://www.statmt.org/moses/giza/GIZA++.html>)

^{††}Indri is a new search engine from the Lemur project. (<http://www.lemurproject.org/indri/>)

r_w from R . R is here set to 20. $NRS(Q, w)$ is the RS score normalized by the maximum score. Finally, two scores of the NTS and the NRS are merged by linear combination as follows:

$$CS(Q, w) = \alpha NTS(Q, w) + \beta NRS(Q, w) \quad (6)$$

where $CS(Q, w)$ is the combination score, α and β are parameters for linear combination. α is set to 0.2 and β to 0.8 in the experiment. The $CS(Q, w)$ is also normalized because the CS is used to calculate the weight of the EW in Eq. (9).

$$NCS(Q, w) = \frac{CS(Q, w)}{\max\{CS(Q, t_i), t_i \in REC W, t = i, \dots, n\}} \quad (7)$$

After ranking the $REC W$ s according to the NCS scores, the high-ranked $REC W$ s are selected as the expansion words (EW s). Top 10 words ranked by NCS are chosen as the final EW s.

3.4 Adding Expanded Words into an Original Input Question

In general, since an input question is represented by a vector, the original vector of the input question is expanded by adding the vector of the expand words (EW s) using the *Rocchio* algorithm for relevance feedback [16] as follows:

$$\vec{Q}_{expand} = \vec{Q}_{original} + \delta \vec{EW} \quad (8)$$

where $\vec{Q}_{original}$ is an original input question vector and $\vec{EW}$ is an expansion words vector. δ is also a parameter for the linear combination of an input question and the EW s. δ is set to 0.3 in the experiment. The *TFIDF* scheme is exploited as the weight of each word in the original input question vector. Since the EW s consist of unique words, they have only one as a *TF* score. Thus the *IDF* and NCS scores are used as the weight of each word in the expansion words vector by the Eq. (9). IDF_{EW_i} is the *IDF* score of i -th word in EW s and NCS_{EW_i} is the ranking score of i -th word in EW s by Eq. (7), $NCS(Q, EW_i)$.

$$weight_{EW_i} = IDF_{EW_i} * (1 + NCS_{EW_i}) \quad (9)$$

4. Experiments

4.1 Data Sets and Experimental Settings

A total of 14,702 question-answer pairs in *Naver KiN*, which is a community based question answering service widely used in Korea, were used for evaluating the proposed question expansion method. The corpus is collected from 10 categories (*Computer Communication, Game, Travel, Shopping, Sport, etc.*) on *Naver KiN*. Note that the size of question-answer pairs in *Naver KiN* remarkably varies with respect to the number of the content words. For evaluation, we used the five-fold cross-validation. To set up parameters, validation set was constructed by randomly selecting

Table 1 Training and Test settings in the baseline systems

	<i>Baseline</i>	<i>GST</i>
<i>Training</i>	Question & Answer	Question & Answer
<i>Test</i>	Question	Question & Answer

Table 2 Performance comparison of the baseline systems (% , *micro* denotes a micro-averaging *FI*-score and *macro* indicates a macro-average *FI*-score)

	<i>Baseline</i>	<i>GST</i>
<i>micro</i>	82.9	87.4
<i>macro</i>	82.6	87.1

Fig. 2 Performance changes according to the different values of parameter δ

20% from the train data. As an evaluation measure for the performance of question classification, *FI*-score is used in the experiments. The *SVM* classifier was selected in our experiments because it is the state-of-the-art classifier in text classification. The baseline systems are defined in Table 1.

In practice, an input test question does not have its answer. However, we can use answers associated with test questions because the test questions are randomly selected from the training corpus. It is called the *Gold Standard Test (GST)* and its performances are compared to those of the proposed system.

As shown in Table 2, *GST* achieved better performance than *Baseline* because it uses golden standard answers.

4.2 Question Expanding Methods

We here evaluate proposed question expanding methods using the *NTS*, the *NRS*, and the *NCS*. In Sect. 3.4, we use the parameter δ in the *Rocchio* algorithm by Eq. (8). Since the term frequency for a word of an input question is mostly 1 and the importance scores ($1 + NCS$) of EW s are bigger than 1, the weights of EW s in $\vec{EW}$ are greater than that of question words in $\vec{Q}_{original}$ in many cases. Therefore, we tried to find better parameter δ using the validation set.

As shown in Fig. 2, when the parameter 0.3 is used, the proposed method achieved the best performance. The *validation* indicates the performances based on the validation set and the *test* indicates the performances of the real test data. In Fig. 3, we observed the performance changes according to the different number of EW s from 5 to 100. As can be seen from Fig. 3, when top 10 ranked $REC W$ s are selected as the EW s, the highest performance is reported.

Table 3 shows a performance comparison when input

Fig. 3 Performance changes according to the different number of top ranked expansion words

Table 3 Performance comparison of three word-ranking approaches and two baselines (%)

	Baseline	Baseline(ATP)	NTS	NRS	NCS
<i>micro</i>	82.9	81.7(-1.2)	84.8(+1.9)	89.2(+6.3)	89.9(+7.0)
<i>macro</i>	82.6	81.4(-1.2)	84.3(+1.7)	88.9(+6.3)	89.6(+7.0)

Table 4 Results of the *t*-test

Group1	Group2	P
Baseline	NTS	0.0082
Baseline	NRS	0.0003
NTS	NCS	0.0001
NRS	NCS	0.0180

questions are expanded according to three different word-ranking methods based on *NTS*, *NRS* and *NCS*. In the *NTS* and *NRS* methods, Eq.(8) is reformed by using *NTS* or *NRS* instead of *NCS*. *Baseline(ATP)* is expanding the input question using *ECWs*, which are not processed by our noise elimination step based on pseudo relevant feedback.

As a result, the question expansion method based on *NCS* achieved the best *F1*-score of 89.9%. We believe that our method can generate a pseudo answer very well. Significant tests were conducted between baseline and the proposed schemes by *t*-test. As a result, all the improvements are statistically significant, $p < 0.01$ except one case between *NRS* and *NCS*. This *t*-test is based on one-tailed paired *t*-test. Table 4 shows results of the *t*-test.

5. Conclusions and Future Work

This paper has studied a new question expanding method for question classification in a cQA service by combining the translation probability from *GIZA++* and the pseudo relevant feedback based on the Indri search engine. As a result, when we effectively combine the relevant scores from the pseudo relevant feedback and the translation scores from the translation probability, we achieved the best *F1*-score of 89.9%. It is 7.0% higher performance than one of baseline. In future work, we plan to apply the proposed expanding method to question retrieval for improving the performance of the cQA service.

Acknowledgments

This research was supported by Basic Science Research Pro-

gram through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. NRF-2013R1A1A2009937)

References

- [1] W. Magdy and G.J.F. Jones, "A Study on Query Expansion Methods for Patent Retrieval," Proc. PaIR'11, pp.19–24, 2011. DOI: 10.1145/2064975.2064982
- [2] Q. Huang, D. Song, and S. Ruger, "Robust Query-Specific Pseudo Feedback Document Selection for Query Expansion," Proc. ECIR'08, pp.547–554, 2008. DOI:10.1007/978-3-540-78646-7_54
- [3] J. Jeon, W.B. Croft, and J.H. Lee, "Finding similar questions in large question and answer archives," Proc. CIKM'05, pp.84–90, 2005. DOI: 10.1145/1099554.1099572
- [4] G. Zhou, L. Cai, J. Zhao, and K. Liu, "Phrase-Based Translation Model for Question Retrieval in Community Question Answer Archives," Proc. HLT '11, pp.653–662, 2011.
- [5] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," Proc. The Knowledge Engineering Review, vol.18, no.2, pp.95–145, 2003. DOI: 10.1017/S0269888903000638
- [6] W. Zhang, Y. Zhang, S. Xiao, L. Wang, and T. Liu, "Exploring Domain Specific Feature Weights for Question Classification Enhancement in Community-based QA Services," Proc. Computational Information Systems, vol.9, no.6, pp.2373–2381, 2013.
- [7] X. Quan and W. Liu, "Term weighting schemes for question categorization," Proc. Pattern Analysis and Machine Intelligence, IEEE, vol.33, no.5, pp.1009–1021, 2011. DOI: 10.1109/TPAMI.2010.154
- [8] A. Singh and K. Visweswariah, "CQC: Classifying Questions in CQA Websites," Proc. CIKM'11, pp.2033–2036, 2011. DOI: 10.1145/2063576.2063883
- [9] L. Cai, G. Zhou, K. Liu, and J. Zhou, "Large-Scale Question Classification in cQA by Leveraging Wikipedia Semantic Knowledge," Proc. CIKM'11, pp.1321–1330, 2011. DOI: 10.1145/2063576.2063768
- [10] X. Xue, J. Jeon, and W.B. Croft, "Retrieval Models for Question and Answer Archives," SIGIR'08, pp.475–482, 2008.
- [11] D. Bernhard and I. Gurevych, "Combining lexical semantic resources with question & answer archives for translation-based answer finding," Proc. ACL-IJCNLP'09, pp.728–736, 2009. DOI: 10.1002/asi.23153
- [12] M. Karimzadehgan and C.X. Zhai, "Estimation of statistical translation models based on mutual information for ad hoc information retrieval," Proc.SIGIR'10, pp.323–330, 2010. DOI: 10.1145/1835449.1835505
- [13] K.S. Lee, W.B. Croft, and J. Allan, "A Cluster-Based Resampling Method for Pseudo-Relevance Feedback," Proc.SIGIR'08, pp.235–242, 2008. DOI: 10.1145/1390334.1390376
- [14] Y. Lv and C.X. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," Proc. CIKM'09, pp.1895–1898, 2009. DOI: 10.1145/1645953.1646259
- [15] S.E. Robertson and S. Walker, "Okapi/Keenbow at TREC-8," TREC-8, pp.151–161, 1999.
- [16] C.D. Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval, pp.173–190, Cambridge University Press, Cambridge, England, 2007.