IEICE TRANS. INF. & SYST., VOL.E92–D, NO.4 APRIL 2009

LETTER

Computing Word Semantic Relatedness for Question Retrieval in Community Question Answering

Jung-Tae LEE[†], **Young-In SONG**[††], ***Nonmembers***, ***and*** **Hae-Chang RIM**[†*a)], ***Member***

SUMMARY Previous approaches to question retrieval in community-based question answering rely on statistical translation techniques to match users' questions (queries) against collections of previously asked questions. This paper presents a simple but effective method for computing word relatedness to improve question retrieval based on word co-occurrence information directly extracted from question and answer archives. Experimental results show that the proposed approach significantly outperforms translation-based approaches.

key words: *question retrieval, community question answering, word semantic relatedness*

1. Introduction

Community question answering (CQA) services, e.g. Yahoo! Answers (http://answers.yahoo.com), where users ask or answer questions posed by other users in a collaborative fashion have been gaining popularity on the Web in recent years. Collections of questions and answers (Q&A) stored in these community-based services are valuable resources for many information seekers, because users can acquire hard-to-find information simply by searching through the collections for already answered questions similar to their information need.

As for any other IR task, a major challenge in designing a retrieval model for CQA is the handling of word mismatches between a user's question and answered questions stored in CQA. For example, "Where can I get cheap airplane tickets?" and "Any travel web-site for low airfares?" are two questions that are *semantically* similar to each other but share no words in common. Conventional word-based retrieval models would fail to match these two questions.

In the IR literature, various query expansion techniques, e.g. relevance feedback [1], thesaurus-based expansion [2], dimension reduction [3], and pseudo-relevance feedback [4], have been proposed to solve the word mismatch problem. In the case of question retrieval in CQA, approaches based on statistical translation techniques, namely the IBM Model 1 [5], have shown great promise [6], [7]. These translation-based approaches model the relatedness of words through word-to-word translation probabilities learned in advance from a parallel corpus derived from existing Q&A archives. However, as pointed out in [8], the incorporation of the IBM Model 1 based techniques into retrieval tasks is virtually a variant way of using co-occurrence information in the document collection, formulated in a different, statistical machine translation perspective.

Motivated by the studies on capturing relatedness of words by co-occurrences derived directly from the document collection to enhance retrieval performance [8]–[11], we propose an approach to using co-occurrence information directly obtained from Q&A archives in deriving word semantic relatedness[**] for question retrieval in CQA, *without* the use of any translation-based technique. Experimental results on real world Q&A data show that the proposed method significantly outperforms the previous translation-based approaches.

2. Previous Work

This work is most closely related to the work of Jeon et al. [6] and Xue et al. [7] on addressing the issue of word mismatches between user questions (queries) and questions stored in CQA. Motivated by Berger and Lafferty's approach [12] to viewing IR as statistical translation, [6] and [7] employ word translation probabilities for mapping a question word to a query word in their question retrieval models. Although the two differ in the way they derive their parallel corpora[***] from Q&A archives, they both use the classic IBM Model 1 [5] to learn translation probabilities between words.

The IBM Model 1 does not require any linguistic knowledge for neither the source nor the target language. Instead, the model learns the translation probability T from a source word s to a target word t as:

$$T(t|s) = \lambda_s^{-1} \sum_{i}^{N} c(t|s; J^i) \quad (1)$$

where λ_s is a normalization factor to make the sum of translation probabilities for the word s equal to 1, N is the number

Manuscript received September 19, 2008.
Manuscript revised November 28, 2008.

[†]The authors are with the Dept. of Computer and Radio Communications Engineering, Korea University, Seoul, South Korea.
[††]The author is with the Dept. of Computer Science and Engineering, Korea University, Seoul, South Korea.
[*]Corresponding author.
a) E-mail: rim@nlp.korea.ac.kr

DOI: 10.1587/transinf.E92.D.736

[**]We want to be very clear that, throughout this paper, we refer to the term "*semantic relatedness*" as the likelihood of a given word combination according to its frequency in a training corpus (in our case, the Q&A archives), not the similarity of word senses.

[***][6] artificially collects parallel pairs of questions that have similar answers, while [7] directly uses the existing question-answer pairs as parallel texts.

of training samples, and J^i is the ith pair in the training data. $c(t|s; J^i)$ is calculated as:

$$c(t|s; J^i) = \frac{P(t|s)}{P(t|s_1) + \cdots + P(t|s_n)} \#(t,J^i)\#(s, J^i) \quad (2)$$

where $\{s_1, \ldots, s_n\}$ are words in the source text in J^i; $\#(t, J^i)$ and $\#(s, J^i)$ are the number of times that t and s occur in J^i, respectively. Given the initial values of $T(t|s)$, Eq. (1) and (2) are used to update $T(t|s)$ repeatedly until the probabilities converge, with an EM-based algorithm.

Note that the IBM Model 1 estimates the translation probability from s to t, solely by considering how many times t *co-occurs* with s in parallel texts. As noted in [8], the approach of using such translation model is indeed a variant way of utilizing word co-occurrences observed from a corpus, formulated in a different, statistical translation setting. In other words, the majority of the translation probabilities learned by the IBM Model 1 virtually has no distinctive difference from co-occurrence probabilities.

In this paper, we propose a method that requires neither developing a parallel corpus nor training any translation models; we make use of word co-occurrences directly obtained from the Q&A corpus to capture word semantic relatedness for question retrieval in CQA.

3. The Question Retrieval Model

Given a user question q and a collection of answered questions $\{q_1, q_2, \ldots, q_n\}$, the task of question retrieval is to rank q_i according to the similarity score $sim(q, q_i)$.

The question retrieval model we use in this paper is based on the language modeling (LM) framework for IR [13]. Under the LM framework, $sim(q, q_i)$ can be modeled as the probability of the document language model D built from q_i generating q. If we assume that words occur independently, $sim(q, q_i)$ may be calculated as:

$$sim(q, q_i) \approx P(q|D) = \prod_{w \in q} P(w|D) \quad (3)$$

To avoid zero probabilities, a mixture between a document-specific multinomial distribution and a multinomial distribution estimated from the entire document collection is used in practice:

$$P(w|D) = (1 - \lambda)P(w|D) + \lambda P(w|C) \quad (4)$$

where $0 < \lambda < 1$, and C refers to a language model derived from the entire Q&A archive. $P(w|D)$ and $P(w|C)$ are calculated using maximum likelihood estimation (MLE).

To handle the word-mismatch problem, we assume that the generation of q follows a Markov process. First, a word t is chosen randomly from an original question q_i, according to the distribution estimated from q_i. Then, a word in q is generated based on the chosen word t, according to semantic relatedness among the words. Assuming that semantic relatedness can be represented by a conditional probability distribution $R(\cdot|\cdot)$ between words, the probability of choosing a word w as a representative of D is:

$$P(w|D) = (1 - \lambda) \sum_{t \in D} R(w|t)P(t|D) + \lambda P(w|C) \quad (5)$$

This question retrieval model is basically equivalent to the one used in [6], except that the translation model is replaced with $R(w|t)$.

4. Modeling Semantic Relatedness of Words from Q&A Archives

We compute semantic relatedness among words, i.e. $R(w|t)$ in Eq. (5), according to the word co-occurrence information directly obtained from Q&A archives. In general, word co-occurrences are observed within a certain context. In this paper, a window-oriented approach, i.e. counting co-occurrences within a window of fixed length, is taken. Then, using MLE, the semantic relatedness of words w_1 and w_2 conditioned on the appearance of w_1 is simply calculated as:

$$R(w_2|w_1) = f(w_1, w_2)/f(w_1) \quad (6)$$

where $f(w_1, w_2)$ is the frequency that w_1 and w_2 co-occur in the corpus, and $f(w_1)$ is the frequency of w_1 in the corpus. The frequencies in Eq. (6) may also be alternatively calculated using document frequencies.

Now, note that each Q&A data is actually a document with fields, e.g. a question field and an answer field. Depending on the CQA service, the question field (or the answer field) may be divided again into two sub-fields: a title and a body. Intuitively, the *purpose* of each field or subfield is unique; for example, a user would write his/her information need summarized in a few words in the question title field and then fill-in the question body field in order to elaborate in detail. Thus, we can assume that *each* field or sub-field would have interesting semantic relatedness information that may be either weakly captured or sometimes even neglected in other fields.

Therefore, we treat a given Q&A archive as a conjunction of unique sub-corpora distinguished by fields, each assigned with a relative weight of importance. We then compute the semantic relatedness of w_1 and w_2 by linearly combining evidences obtained from each field as follows:

$$R(w_2|w_1) = \sum_j \alpha_j R_j(w_2|w_1) \quad (7)$$

where α_j is the weight of the field j, and $\sum_j \alpha_j = 1$. $R_j(w_2|w_1)$ is the semantic relatedness of w_1 and w_2 obtained from j.

5. Experimental Settings

We have tested whether the word co-occurrences based semantic relatedness measure can improve retrieval performance. For the experiments, we have obtained a total of 43,001 questions[†] with a best answer (selected either by the

[†]The questions have two sub-fields: question titles and question bodies (optional).

questioner or by votes of other users) from the "*Science*" domain of Yahoo! Answers. Among them, 32 questions have been randomly selected as the test set (queries) and the remaining as the reference set to be retrieved. We have used a pooling technique [14] to find relevant questions for the queries. Three annotators judged a question as relevant if it addressed the same information need as the query. As a result, 177 relevant questions have been found in total.

The proposed model for question retrieval, namely the Semantic Relatedness based Language Model (SRLM), is compared to three baseline models:

- QLM: Query-likelihood Language Model widely used in the literature. This model estimates $P(w|D)$ in Eq. (3) by using MLE.
- TLM.Jeon: Translation-based Language Model proposed by Jeon et al. [6]. This model extends QLM by utilizing word translation probabilities in estimating $P(w|D)$. Translation probabilities are learned from a parallel corpus consisting of semantically similar question pairs. (Questions are considered similar if their answers are similar to each other above some threshold.)
- TLM.Xue[†]: Translation-based Language Model proposed by Xue et al. [7]. This model improves upon TLM.Jeon by linearly combining MLE-based estimation and translation-based estimation when computing $P(w|D)$. Translation probabilities are learned directly from question-answer pairs.

For both translation-based approaches, we have used the out-of-the-box GIZA++[††] to train the IBM Model 1, as in [6], [7].

6. Experimental Results and Discussions

All retrieval models tested here rank questions according to the similarity scores between queries and question titles, because the question title part is known to be most useful in finding relevant questions [6]. Table 1 summarizes the comparative results in Mean Average Precision (MAP) and R-Precision (R-Prec) [15]. For each method, the best performance after empirical parameter tuning according to MAP is presented.

Notice that QLM shows the lowest retrieval performance. This implies that word-based models often fail to retrieve relevant questions that have little word overlap with queries, as noted in [6]. Also, notice that SRLM achieves significantly better performance than all the baselines. This clearly indicates that co-occurrences obtained directly from the Q&A archive using a fixed-sized window is simple but effective in computing lexical semantic relatedness for improving the performance of question retrieval.

We suggest the performance loss of translation-based approaches compared to SRLM is due to the noise that may have been created while learning translation probabilities with SMT techniques. In the case of TLM.Jeon, noise may have been created while synthetically generating the parallel pairs of similar questions for training the IBM Model 1.

Table 1 Comparisons with three baselines. Percentage changes are with regard to QLM. The improvement of SRLM is tested to be statistically significant using paired t-test.

Model	MAP	R-Prec
QLM	0.1031	0.2396
TLM.Jeon	0.1131 (+9.7%)	0.2428 (+1.4%)
TLM.Xue	0.1417 (+37.4%)	0.2713 (+13.2%)
SRLM	**0.1480 (+43.6%)**	**0.2860 (+19.4%)**

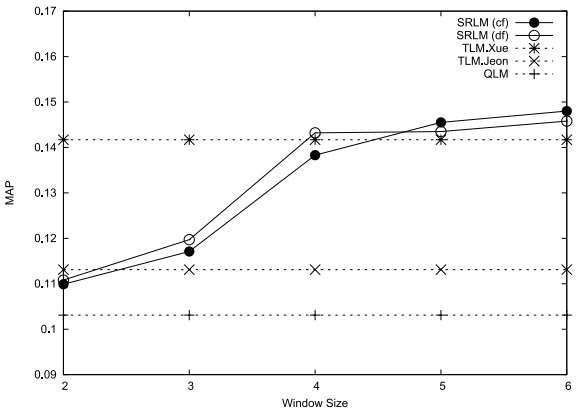

Fig. 1 Effect of window size for observing co-occurrences, based on MAP. "cf" and "df" denote the cases when collection frequencies and document frequencies are used, respectively.

Moreover, the probabilities are learned virtually from the question part only. Although TLM.Xue seems to resolve these defects by directly using the question-answer pairs instead to learn the translation probabilities, the relatively broad range of context in which word co-occurrences are observed within may have still caused noise to occur when estimating translation probabilities. The IBM Model 1 used in translation-based approaches obtains co-occurrences by considering *all* possible word alignments within given parallel texts, while SRLM captures co-occurrences only within a small fixed-sized context window. Figure 1 shows the influence of window size (in words) used for SRLM. Note that SRLM shows either comparable or better retrieval performance than TLM.Xue by using context of 4 to 6 words.

Figure 2 shows the effect of each weight parameter assigned to a field. Interestingly, SRLM has outperformed baseline methods noticeably when the weight of the question title field is assigned with a very low value. We hypothesize the reason for the ineffectiveness of question title field, which has been known to be the most effective search field, in computing semantic relatedness as follows. In many cases, users express the information need concisely in the question title and elaborate in the question body, causing the title to be relatively short. Thus, a very small amount of useful co-occurrence information would be extracted from

[†]The final version of Xue et al. [7]'s retrieval model uses both question and answer parts since their goal has been to rank Q&A pairs. We use a version of their model in which the answer part is *not* utilized, as a baseline, because the focus of our paper is to rank questions, *not* Q&A pairs.

[††]http://www.fjoch.com/GIZA++.html

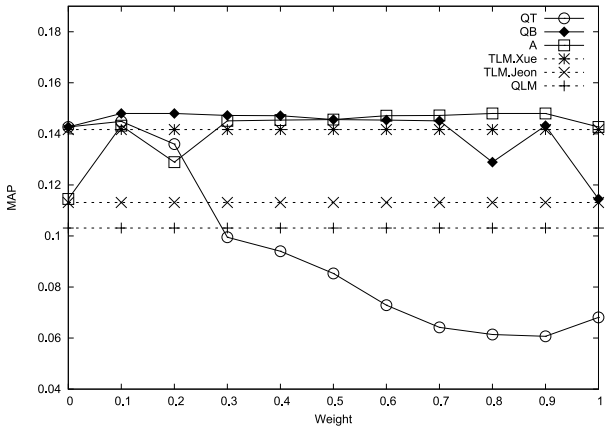

Fig. 2 Effect of each field parameter involved in estimation of semantic relatedness, according to MAP. "QT," "QB," and "A" denote weights assigned to question title, question body, and answer fields, respectively.

question titles compared to other fields that are relatively longer. For the weight of either the question body or the answer field, a relatively broad set of good parameter values is observed. This result supports our hypothesis that more useful co-occurrences are likely to be observed in longer texts.

7. Conclusions

Bridging the lexical gap between user questions and answered questions stored in Q&A archives in question retrieval has become an important issue due to the increasing popularity of online CQA services. In this paper, we have presented a simple but effective approach to computing word relatedness based on word co-occurrences obtained directly from Q&A collections for question retrieval in CQA. The proposed method has shown results comparable to statistical translation-based methods, which have been considered state-of-the-art for question retrieval in CQA, on real-world Q&A datasets.

Future work will focus on complementing the statistical semantic relatedness computed from co-occurrences observed in Q&A archives by additionally utilizing resources that contain manually recognized word relationships, e.g. WordNet. We also plan to investigate the effectiveness of computing one-to-many- or many-to-one-word relatedness for further enhancement of question retrieval in CQA.

Acknowledgments

This work was supported by Microsoft Research Asia and the Second Brain Korea 21 Project. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] J.J. Rocchio, "Relevance feedback in information retrieval," The SMART Retrieval System: Experiments in Automatic Indexing, pp.324–336, 1971.
[2] E.M. Voorhees, "Query expansion using lexical-semantic relations," Proc. 17th ACM SIGIR, pp.61–69, 1994.
[3] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman, "Indexing by latent semantic analysis," J. Am. Soc. Inf. Sci., vol.41, no.6, pp.391–407, 1990.
[4] V. Lavrenko and W.B. Croft, "Relevance based language models," Proc. 24th ACM SIGIR, pp.120–127, 2001.
[5] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," Computational Linguistics, vol.19, no.2, pp.263–311, 1993.
[6] J. Jeon, W.B. Croft, and J.H. Lee, "Finding similar questions in large question and answer archives," Proc. 14th ACM CIKM, pp.84–90, 2005.
[7] X. Xue, J. Jeon, and W.B. Croft, "Retrieval models for question and answer archives," Proc. 31st ACM SIGIR, pp.475–482, 2008.
[8] G. Cao, J.Y. Nie, and J. Bai, "Integrating word relationships into language models," Proc. 28th ACM SIGIR, pp.298–305, 2005.
[9] J. Gao, J.Y. Nie, G. Wu, and G. Cao, "Dependence language model for information retrieval," Proc. 27th ACM SIGIR, pp.170–177, 2004.
[10] R. Jin, A.G. Hauptmann, and C.X. Zhai, "Title language model for information retrieval," Proc. 25th ACM SIGIR, pp.42–48, 2002.
[11] R. Nallapati and J. Allan, "Capturing term dependencies using a language model based on sentence trees," Proc. 11th ACM CIKM, pp.383–390, 2002.
[12] A. Berger and J. Lafferty, "Information retrieval as statistical translation," Proc. 22nd ACM SIGIR, pp.222–229, 1999.
[13] J.M. Ponte and W.B. Croft, "A language modeling approach to information retrieval," Proc. 21st ACM SIGIR, pp.275–281, 1998.
[14] D. Harman, "Overview of the first trec conference," Proc. 16th ACM SIGIR, pp.36–47, 1993.
[15] J. Allan, "Hard track overview in trec 2005: High accuracy retrieval from documents," Proc. 14th TREC, 2005.