Expanding Comments Using Previously Responded Posts for Short Text Conversation in Open Social Media

Jian-qiang WANG∗, Ying SUN and Dagobert SOERGEL
Department of Information Science, University at Buffalo, Buffalo, NY 14260, U.S.A.
∗Corresponding author

Keywords: Short text conversation, Information retrieval, Social media, WeiBo.

Abstract. We report a study on improving the retrieval of comments made to previous posts that can potentially be relevant to new posts in the open social media domain. This retrieval task, known as an important component of short text conversation, is challenging due to various reasons, among which is the fact that both posts and comments are usually very short, thus not providing enough context for typical information retrieval systems to be very effective. We designed a technique of expanding each comment by adding words to it that are selected from the post(s) that the comment previously responded to. Our experiment using a Sina WeiBo collection shows that by properly considering the number of posts that a comment has previously responded to, the number of responding comments of each of these posts, as well as the number of words to be selected for expansion, the proposed technique can result in slightly improved retrieval effectiveness. Furthermore, the technique can retrieve relevant comments that have not been returned by any other previous systems.

Introduction

Natural language conversational systems use computer agents to converse with humans in a coherent manner. Such systems can be useful in many application areas, including education, healthcare, government, business and entertainment where fast and accurate responses to a potentially large number of user inquiries are desired. The development of the systems often requires techniques of natural language processing and understanding, machine learning, reasoning, dialog modeling, information extraction, knowledge base development, and automatic speech recognition and text-to-speech synthesis in the cases of speech-based systems.

A research focus in this area in recent years has been on social media interaction, thanks to the rapid proliferation of social media websites like Twitter and the Chinese Sina WeiBo as major platforms for online users to connect to each other with shared interests. The research, known as Short Text Conversation (STC), aims at developing computer systems that retrieve previous messages (known as comments) or generate new messages in response to initial messages (known as posts). This task is on short text because it focuses on application areas like Twitter and WeiBo where posts and comments are usually very short. Ultimately, a STC system should support multiple rounds of conversation between the user and the computer. As an initial but important step, however, it is useful to build a one-round STC system in which the user submits a post and accordingly the computer either generates new comments (generation-based method) or retrieves comments from an existing collection (retrieval-based method), in response to the post.

In our study reported here, we focus on retrieval-based methods for STC. A key challenge of this task lies in the fact that comments are often very short, thus rendering traditional techniques of information retrieval (IR) less effective as they usually rely on matching words contained in the post against words contained in each comment. To compensate for this problem, we designed a technique that for each comment in the collection, selects words from the post(s) that the comment has previously responded to and then adds these words to it so that an expanded comment is created. Our hypothesis is that the technique can bring in useful words so that the expanded comment has a better chance to be retrieved, i.e., the retrieval effectiveness can be potentially improved.
The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 describes our technique of expanding each comment based on its previously responded posts. Section 4 introduces the test collection, IR system and effectiveness measures used in our experiment and reports the results. We conclude the paper with Section 5, where we also suggest the future research direction.

Related Work
Human languages are incredibly rich and complex in that often a meaning can be expressed with different words and at the same time a given word can have different meanings in different contexts. This often leads to discrepancies of vocabulary use between the information searcher and the information creator, which is a fundamental problem that IR systems like Web search engines have to deal with since these systems are usually built upon matching words contained in the information request (known as *query*) against word contained in what’s being sought (generally referred to as *documents*). The problem becomes more severe when the documents being retrieved are very short, which is the case with messages used in social media interaction.

Many techniques have been designed to tackle this problem of word mismatch in IR. Relevance feedback is a commonly used technique of expanding the initial query by adding useful words to it [1]. These words are selected from either initially retrieved documents that are deemed relevant by the searcher or simply top-ranked documents in the initial set. Other techniques seek to enhance documents, which can be particularly useful when the documents are noisy or short. Singhal & Pereira proposed an idea of expanding each automatically transcribed broadcast news article with an error-free text document in the same domain [2]. Alvarez-Melis and Saveski studied the usefulness of aggregating tweets by conversation for the purpose of topic modeling [3]. Shang et al are among the first to study the problem of short text conversation [4]. Their work resulted in the creation of a collection of Sina WeiBo posts and comments, which was later used in the NTCIR STC task [5, 6]. Much research on STC focused on using deep neural networks or word embedding (e.g., [7]). In our previous study, we expanded comments that each has responded previously to only one post [8]. In the study reported here, we extend our work by taking into consideration comments that have responded to different numbers of posts and the popularity of each responded post, performing word selection based on their importance as well as looking into whether our technique can retrieve relevant comments that have never been found by other IR systems.

Our Technique of Expanding Comments
As described briefly in Section 1, the key idea of our technique is to expand each comment in a social media collection by adding useful words that are selected from the post(s) that the comment responded to previously. Below are a few examples illustrating the usefulness of the technique:

- An existing post \( P \): *I lost big money this year in the stock market.*
- A comment \( C \) that responded to \( P \): *Maybe you should consider investing in the real estate.*
- An expanded comment \( C_e \): *I lost big money this year in the stock market maybe you should consider investing in the real estate.*
- A new post \( P_n \): *I'm not sure if it is a good time to invest in the stock market.*

Clearly the comment \( C \) is relevant to the new post \( P_n \). However, an IR system of matching words will likely fail to retrieve \( C \) for \( P_n \) because they do not share content-bearing words. On the other hand, since \( C \) has responded to a previous post \( P \) that shares the content-bearing words *stock* and *market* with \( C \), the expanded comment \( C_e \) will more likely be retrieved for the new post \( P_n \).

Social media interaction may contain information like who posted the message, in which topic areas that person posted frequently, and how many posters responded to that person. It is reasonable to argue that the information can and perhaps should be considered when expanding comments for IR purposes. However, in our study we assume this kind of information is unavailable (e.g., it is removed due to privacy concern when social media collections are used for research and product development).
Instead, we focus on social media collections that contain only the text of each message and information on which comments having responded previously to which posts. Accordingly, we aim at investigating how the following three parameters affect the effectiveness of the expansion technique: (1) the number of posts that a comment has previously responded to, (2) the number of previously responding comments for each of these posts, and (3) the number of words to be selected from these posts for expanding the comment. To our best knowledge, there has not been any published work looking into these factors together for STC.

While the first two numbers are readily available in a given collection, there exist different ways of selecting words from a post or posts. In this study, we chose to select words based on their document frequency (DF) values. In IR the DF of a word is defined as the proportion of documents in a collection that contain the word. For words in a given document, those with smaller DF values often carry more weight about the topic of the document than those with larger DF values. Therefore, when expanding a comment, we want to select \( n \) words with the smallest DF values from the post(s) that the comment responded to, with \( n \) being a parameter to tune.

Experiment

In this section, we describe the test collection used in our study, the retrieval system, the effectiveness measures, the experiment results and our analyses of them.

The Test Collection, Retrieval System and Effectiveness Measures

We used the Chinese collection of the NTCIR-13’s STC task to test the effectiveness of our technique [6]. Gathered from China’s Sina WeiBo, the collection contains 4,305,706 unique comments that responded to 219,174 unique posts. The collection retains only the text of each message, which was assigned with a unique ID, as well as information on which comments responded to which posts previously. A post can have one to many responding comments and likewise a comment may have responded to one or many posts. After performing Chinese word segmentation, we excluded comments of two words or fewer. On average each post contains 13 words while each comment contains 10 words.

The test collection also contains 100 evaluation posts, to mimic new posts in the future. The task for an IR system is for each evaluation post, to retrieve 10 relevant comments and rank them in decreasing order of their relevance to the post. 18 teams submitted a total of 64 retrieval-based runs for the NTCIR-13 Chinese STC task, each of which was generated by a different IR technique. Three independent assessors were then hired by the NTCIR organizers to manually judge the relevance of each retrieved comment for its post. Relevance takes one of the three levels: “2” denotes “relevant,” “1” denotes “partially relevant,” and “0” denotes “nonrelevant.” The relevance judgments made by the three assessors were then summed up, resulting in an overall relevance level ranging from 0 to 6 of each comment for its post. The final official set contains 27,944 entries, each specifying the official relevance level of a comment for a post. Comments in the collection that were not retrieved by any team were simply regarded as nonrelevant, which is a common practice in IR evaluation. The official relevance judgments can be used as ground-truth for evaluating future IR techniques.

Our retrieval system is a local implementation of the well-known Okapi BM25 weighting [9]. A baseline run was generated by retrieving 10 comments from the original (unexpanded) collection. With each parameter setting we created a collection of expanded comments and then retrieved 10 comments from it for each post. To measure the retrieval effectiveness of each run, we computed each individual post’s \( \text{nG@1} \) (normalized gain at cutoff 1) and \( \text{nERR@10} \) (normalized expected reciprocal rank at cutoff 10) and then took the average over 100 posts. These two measures were also used in the NTCIR STC task evaluation. The average \( \text{nG@1} \) (or \( \text{nERR@10} \)) value is viewed as the overall effectiveness of each run. Due to the space limit, interested readers can find more information of these measures as well as the evaluation method in [5].
Results and Analysis

Table 1 shows our experiment results in terms of average nG@1 and average nERR@10 of the baseline run and of different expansion runs. Each expansion run is named in the form of exp.x-y-z where x is the maximum number of posts that a comment responded to previously, y is the maximum number of previous responding comments for each of those posts, and z is the number of words selected from those posts and added to the comment for expansion. For example, exp.1-20-4 means we only expanded those comments that each responded previously to only one post, the post had no more than 20 responding comments, and we selected four words from it with the lowest DF values; exp.all-all-all means every comment is expanded by simply adding all word contained in its responded post(s).

<table>
<thead>
<tr>
<th></th>
<th>nG@1 (official)</th>
<th>nG@1 (local)</th>
<th>nERR@10 (official)</th>
<th>nERR@10 (local)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.407</td>
<td>0.407</td>
<td>0.542</td>
<td>0.542</td>
</tr>
<tr>
<td>exp.1-20-4</td>
<td>0.3887</td>
<td>0.392</td>
<td>0.5368</td>
<td>0.5439</td>
</tr>
<tr>
<td>exp.1-all-4</td>
<td>0.4027</td>
<td><strong>0.414</strong></td>
<td>0.5304</td>
<td><strong>0.5547</strong></td>
</tr>
<tr>
<td>exp.5-20-4</td>
<td>0.3937</td>
<td>0.397</td>
<td>0.5391</td>
<td>0.5464</td>
</tr>
<tr>
<td>exp.5-all-4</td>
<td><strong>0.411</strong></td>
<td><strong>0.4227</strong></td>
<td>0.5337</td>
<td><strong>0.5582</strong></td>
</tr>
<tr>
<td>exp.5-20-8</td>
<td>0.3962</td>
<td>0.3962</td>
<td>0.5458</td>
<td>0.5505</td>
</tr>
<tr>
<td>exp.5-all-8</td>
<td><strong>0.411</strong></td>
<td><strong>0.426</strong></td>
<td>0.5297</td>
<td><strong>0.5581</strong></td>
</tr>
<tr>
<td>exp.all-all-8</td>
<td>0.3868</td>
<td><strong>0.4168</strong></td>
<td>0.519</td>
<td><strong>0.5556</strong></td>
</tr>
<tr>
<td>exp.all-all-all</td>
<td>0.3643</td>
<td>0.3777</td>
<td>0.4928</td>
<td>0.5152</td>
</tr>
</tbody>
</table>

Notice for each effectiveness measure, there are two columns included in the table. The two official columns contain evaluation results using the NTCIR-13 STC official relevance judgments whereas the two local columns are evaluation results by also considering our local relevance judgments of retrieved comments that were not included in the official relevance judgments (more on this later).

When evaluating with only the official relevance judgments, the effectiveness of our expansion technique generally does not seem to be improved over the baseline – it is actually lower if no restriction is applied on the number of responded posts and the number of responding comments of each post, although the differences are statistically insignificant based on t-tests. The main cause appears to be from not limiting the number of posts that a comment responded to. We manually checked a small set of comments that each responded to more than 50 posts and confirmed that in most cases they are very generic comments (e.g., *I and my little buddies are all surprised!*). Expanding such comments can undesirably increase their chance of being retrieved. Therefore, it is important to expand only comments that have responded to a small number of posts; in the limited number of runs we tried, 5 seems to be a reasonable number. On the other hand, limiting the number of responding comments of a post does not seem to have an effect. A possible explanation is that if a post has many responding comments, it may have high quality, so we actually want to select some words from it for expansion.

We further checked whether our expansion technique retrieved additional comments that had not been retrieved by any NTCIR participating teams and if so, whether any of these newly found comments are relevant. It turned out that our technique retrieved 705 such comments (pooled from all runs). We asked three voluntary native Chinese speakers to independently judge the relevance of each of these comments for its corresponding post. All three assessors hold a USA master degree; they were briefly introduced to the task of relevance judgments and practiced with judging 20 comments selected from the official set before starting the task. To compare our local relevance judgements with the official NTCIR ones, we randomly selected 100 entries from the official set and put them into the set of the 705 comments. The final results showed that 119 of these 705 documents received an overall relevance level of 3 or higher. Comparison between our local judgments of the 100 comments with the official judgments indicated our assessors were significantly stricter, meaning they tended to assign lower relevance levels than the NTCIR-13 STC assessors. After adding these new relevance
judgments to the official set, we recomputed the nG@1 and nERR@10 of each run. As shown in Table 1, our technique can lead to improvement 5% on nG@1 and 3% on nERR@10 when word selection was applied and only comments responding to a small number of posts were expanded.

Conclusion and Future Work

We demonstrated by evaluating with the NTCIR-13 STC collection of Sina WeiBo that expanding each comment with words selected from the post(s) it had previously responded to could result in limited improvement of the effectiveness of retrieving comments for new posts. In some cases, the technique can retrieve relevant comments that have not been found by previous techniques. The expansion technique seems to be sensitive to the number of posts that a comment has previously responded but less so to the number of responding comments that a post had. Furthermore, word selection should be performed for expansion. In the future, we plan to incorporate the poster’s information into the technique as well as study it in other frameworks, such as deep neural networks.

References


