Topic-Focused Summarization of Chat Conversations

Arpit Sood, Thanvir P. Mohamed, and Vasudeva Varma

International Institute of Information Technology, Hyderabad
arpit.sood@research.iiit.ac.in, mohamed.thanvir@students.iiit.ac.in, vv@iiit.ac.in

Abstract. In this paper, we propose a novel approach to address the problem of chat summarization. We summarize real-time chat conversations which contain multiple users with frequent shifts in topic. Our approach consists of two phases. In the first phase, we leverage topic modeling using web documents to find the primary topic of discussion in the chat. Then, in the summary generation phase, we build a semantic word space to score sentences based on their association with the primary topic. Experimental results show that our method significantly outperforms the baseline systems on ROUGE F-scores.

1 Introduction

In the recent past, the communication of users through social media has seen an exponential increase. A substantial chunk of information exchange happens in the form of online chat conversations. These conversations attract a substantial number of participants and a single conversation tends to span a wide range of topics interspersed with irrelevant segments. In order to understand what transpired in such long and involved conversations, summarization becomes essential. The summaries provided can be of very good commercial and educational value, and can be used to analyze the impact of virtual social interactions and virtual organizational culture on product development [5]. To summarize a chat, first we identify the primary topic being discussed in the conversation. Then, we build a semantic space of the words using which a summary is produced based on the dependence of the primary topic on sentences. Specifically, the contributions of our work are as follows: (1) Applying topic modeling on web documents to find the topic distribution of sentences, thus dealing with data sparseness. (2) Incorporating topic information into HAL model [3] to generate better summaries.

2 Approach

2.1 Topic Modeling

We apply topic modeling to automatically discover the topics underlying the chat. We use Latent Dirichlet Allocation (LDA) [1] for topic modeling. Before

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 800–803, 2013.

discussing our main approach, we describe a simpler approach of discovering the topic structure of the chat. In this approach, we consider each sentence to be a document and apply LDA to the collection of sentences in the chat. For summary generation, we need the primary topic of discussion in the chat. We define the *primary topic* as the most prevalent topic in the longest sentences of the chat. The topic proportion of each topic in the longest sentences are added up and the topic having the highest sum is considered as the most prevalent.

As shown in the section 3.3, we obtain reasonably good results when summaries are generated by employing the above approach. But since the sentences are short in length, thus lacking in sufficient information, a better approach would be to use certain external information to supplement the information present in the chat. Therefore, we used documents from web. For every word, we query the web using Microsoft Bing API[1]. We fetch the first web document obtained as result and that document is considered to be a description of the word queried. Each sentence in the chat is thus associated with the set of web documents corresponding to the words in that sentence. For every sentence, its corresponding web documents are concatenated and considered as one single document. LDA is applied to this collection of expanded documents and the corresponding hidden topic structure is discovered. The topic distribution of every document is considered as the topic distribution of the corresponding sentence.

The primary topic is identified in the same manner as in the previous approach. In this case, the topics are distributions over the vocabulary of the web documents. However, for summary generation we require the primary topic to be a distribution over the vocabulary of the chat. The trained model, that has been obtained by applying LDA to the set of web documents, is applied to the chat considering each sentence as a document. On applying this, we obtain for each topic a distribution over the vocabulary of the chat.

2.2 Summary Generation

After the topic modeling phase, we use a co-occurrence HAL model [3] for sentence scoring. The HAL model constructs the dependencies of a word w on other words based on their occurrence in the context of w in a sufficiently large corpus. Such a semantic co-occurrence can be captured by building a $term \times term$ matrix. The weights assigned to each co-occurrence of terms are accumulated over the entire corpus. That is, if $matrix(w', k, w)$ denotes the number of times word w' occurs k distance away from w when considering a window of length K, and $W(k) = K - k + 1$ denotes the strength of this co-occurrence between the two words, then HAL co-occurrence score is given by:

$$HAL(w'|w) = \sum_{k=0}^{K} W(k) \times matrix(w, k, w') \tag{1}$$

The HAL space built this way can be used to create a probabilistic model to give scores by normalizing the count of the terms [4]. The probabilistic version of

[1] `http://msdn.microsoft.com/en-us/library/dd251072.aspx`

HAL, pHAL, gives the probability of associating a word w' with another word w in a window of size K. This can be expressed in terms of probability of observing w' at a distance $k < K$ from w, with $n(w)$ denoting the frequency of w, as

$$pHAL(w'|w) = \frac{HAL(w'|w)}{n(w) \times K} \tag{2}$$

After HAL Space is built, scoring of a sentence is done by accumulating scores over terms associated with the primary topic which was discovered in the second phase. Given topic terms, $t_1, t_2, \ldots, t_k$, score of a sentence S is given by

$$Score(S) = \prod_{w_i \in S} \left(P(w_i) \times \prod_{t_k} pHAL(t_k|w_i) \right) \tag{3}$$

3 Evaluation

3.1 Dataset

As guided by the work of Zhou and Hovy [5], we are aware of only one publicly available chat corpus with associated summaries, which is the GNUe Traffic archives[2]. It contains Internet Relay Chat (IRC) logs along with corresponding human created summaries. We have used a collection of 450 chat conversations with corresponding model summaries as the gold standards for evaluation.

3.2 Experimental Setup

In the topic modeling phase, the number of topics in LDA model has to be predetermined. To the best of our knowledge, there is no method to estimate the number of topics which best captures the topic structure. Therefore, we experimented with a range of values, between 5 to 100, for a set of 60 chats, and decided to use 20 as the number of topics. For evaluation, we used the remaining 390 chats. We use ROUGE[3] as the evaluation metric for summarization performance. ROUGE F-scores were computed for different matches: unigram (ROUGE-1), bigram (ROUGE-2) and longest subsequence (ROUGE-L). Length of the generated summaries were restricted to those of model summaries. We compare our methods with the following most widely used baseline systems. (1) MaxLength: Selects the longest sentences of every participant. This is a proven strong baseline for conversation summarization [2]. (2) DiaSumm: This system creates a summary by extracting an inter-connected structure of segments that quoted and responded to each other. (3) FirstSent: Selects the first sentence from each message in the chat sequence (*Position Hypothesis*). (4) LexCocc: the basic pHAL [3] system[4], which has been chosen to show the effect of topic modeling and web referencing on the model proposed. (2) and (3) are hard-to-beat baselines and were used by Zhou and Hovy [5] while summarizing IRC logs.

[2] `http://kt.earth.li/kernel-traffic/index.html`
[3] `http://www.berouge.com/`
[4] Secured 1st position in Document Summarization task in DUC 2007.

3.3 Results

Our methods outperform the baseline systems as shown in Table 1. Our simpler approach of topic-focused summarization (pHAL + LDA) achieving better performance than LexCocc indicates that incorporation of topic information is beneficial. Using Web Reference (WR) module with LDA further increases the performance by capturing the topic structure more accurately. LexCocc underperforms other baselines (MaxLength and DiaSumm) because there is less co-occurrence of related terms in spontaneous chats. The results presented are statistically significant at 99% significance level. We used paired t-test for testing statistical significance.

Table 1. ROUGE F-Scores for Chat Conversations

Configuration	ROUGE-1	ROUGE-2	ROUGE-L
MaxLength	0.42411	0.24927	0.41942
DiaSumm	0.40823	0.23726	0.39315
FirstSent	0.27913	0.15398	0.27016
LexCocc(pHAL)	0.39521	0.21591	0.38128
pHAL+LDA	0.46192	0.26654	0.43562
pHAL+LDA+WR	**0.59978**	**0.39284**	**0.58215**

4 Conclusions and Future Work

In this paper, we have explored a novel approach to chat summarization. We developed an approach to find the primary topic of discussion in the chat conversation, where the sentences in the chat are extended using web documents. This topic information is incorporated into HAL model to generate better summaries. Experiments showed that we have statistically significant performance gain over baselines with ROUGE as the evaluation metric. In future, we would like to extend this study by incorporating the dialogue structure of the chats.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Gillick, D., Riedhammer, K., Favre, B., Hakkani-Tur, D.Z.: A global optimization framework for meeting summarization. In: International Conference on Acoustics, Speech and Signal Processing, pp. 4769–4772 (2009)
3. Jagadeesh, J., Pingali, P., Varma, V.: A relevance-based language modeling approach to duc 2005. In: Document Understanding Conference (DUC 2005) (2005)
4. Lowe, W., McDonald, S.: The direct route: Mediated priming in semantic space. In: Annual Conference of the Cognitive Science Society, pp. 806–811 (2002)
5. Zhou, L., Hovy, E.H.: Digesting virtual geek culture: The summarization of technical internet relay chats. In: Meeting of the Association for Computational Linguistics, pp. 298–305 (2005)