On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs

Liang Zhou and Eduard Hovy

University of Southern California
Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{liangz, hovy} @ isi.edu

Abstract

In this paper we describe computational approaches to summarizing dynamically introduced information: online discussions and blogs, and their evaluations. Research in the past has been mainly focused on text-based summarization where the input data is predominantly newswire data. When branching into these newly emerged data types, we face number of difficulties that are discussed here.

Introduction

The Internet has grown beyond merely hosting and displaying information passively. It provides easy access for people to share, socialize, and interact with one another. Information displayed and exchanged between people are dynamic, in contrast to static information depicted in the older age of the Internet.

Text summarization has been an interesting and active research area since the 60's. The definition and assumption is that a small portion or several segments of the original long document can represent the whole informatively and/or indicatively. Reading or processing this shorter version of the document would save time and other resources. This property is especially true and urgently needed at present due to the vast availability of information. However, we are yet to see true Natural Language Processing (NLP) techniques being deployed on the web for summarization purposes. Many online news agencies use clustering methods to group news articles by their categories and provide pseudo-summaries. News articles are written so that the sentences (or paragraphs) in the beginning (Position Hypothesis) are the most informative and can be used as summaries. In addition, news articles are "static" information because there is no interaction and no means to facilitate such interaction between its authors and readers.

Online discussion forums and personal blogs are "dynamic", involving frequent exchanges between various participants. However, there are differences between discussions and blogs. As it suggests, a discussion involves multiple participants and there isn't a clear distinction among the participants and the information they provide. Without extensive monitoring, it is impossible to rate whether the information provided by one participant is more important than those provided by another participant. Thus, in designing a summarization system for this purpose, we must treat all information and all participants equally. This is where personal blogs and the summarization of blogs are different. A blog entry is equivalent to an entry to a person's diary, albeit published on the web. From the blogs we have studied, people, other than the author, often insert comments responding to the original blog message, but the level of interactions observed is less significant than those from online discussions. The dominance the original blog message imposes simplifies the design and training for a summarization system. We examine this assumption in our analysis, discussed in the following sections.

This paper is organized in the following way: first, we will describe the current status of online discussion summarization; then we discuss different ways to summarize blogs, in particular, political blogs; and lastly, we provide a discussion on potential evaluation methodologies for generated blog summaries.

Online Discussions

Online discussions and blogs are closely associated; it is difficult to discuss the summarization of one without describing the summarization of the other. In recently years, the emergence of online discussions as a major information source has prompted increased interest in thread summarization within the NLP community. One might assume a smooth transition from text-based summarization to discussion-based summarizations (Lam and Rohall, 2002; Newman and Blitzer, 2002; Rambow et al., 2004). However, discussions fall in the genre of correspondence, which requires dialogue and conversation analysis. This property makes summarization in this area

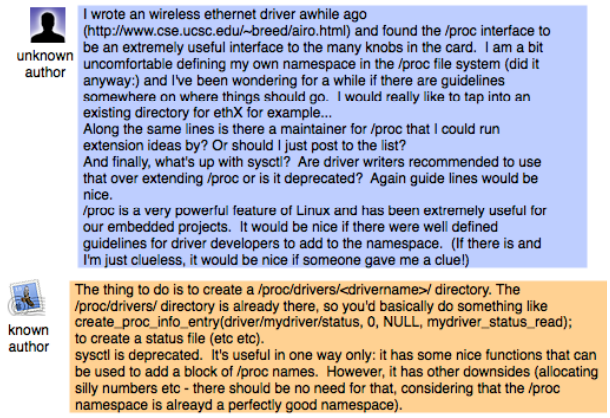

Figure 1. An excerpt from an OSS discussion.

even more difficult than traditional summarization. In particular, topic "drifts" occurs more radically than in written genres, and interpersonal and pragmatic content appears more frequently. Questions about the content and overall organization of the summary must be addressed in a more thorough way for discussion summarization systems.

Having recognized that email exchanges are more complex than written documents, Wan and McKeown (2004) introduce a system that creates overview summaries for ongoing decision-making email exchanges by first detecting the issue being discussed and then extracting the response to the issue. Galley et al., (2004) describe a system that identifies agreement and disagreement occurring in human-to-human multi-party conversations, which have a more complex task structure and dialogue structure.

Discussion Subtopic Structure

Online discussions can be short and light-hearted, or complex and attract heavy participation. For discussions with light participation, topic outlines are helpful. For more complex discussions, such as those that involve collective team efforts to accomplish a common goal, participants often volunteer to write summary digests that catalogue the past conversations on various issues. Human efforts are preferred if the summarization task is easily conducted and managed, and is not repeatedly performed. However, when resources (time, monetary compensation, and human) are limited, automatic summarization becomes more desirable.

In our research (Zhou and Hovy, 2005), we focus on the Open Source Software (OSS) development discussion forum which has matching participant-written summary digests. Keeping track of ongoing discussions is particularly important in large, complex technical discussions groups. In OSS, software designers and programmers have to be aware of all technical issues to ensure quality, compatibility, timeliness, etc., of their software. Figure 1 shows an excerpt of such a discussion. Discussions from the OSS forum are long and complex, making them more interesting to analyze than those from other domains. Each message submitted to a discussion addresses multiple subtopics relating to the main topic and invites more participation on related subjects. To produce summaries for these discussions, we must recognize and reorganize subtopics according to the level of relevance they are to one another.

Due to the complex structure of the dialogue, we observe similar *subtopic structure* in the participant-written summaries. As illustrated in Figure 2, the summary gives a chronology of what has been discussed in a particular discussion (Figure 2a) but does not distinctively group (technical) issues with their comments, necessitating the usage of reader guidance phrases such as "for the … question", "on the … subject", "regarding …", etc., to direct and refocus reader's attention. A better summary of the discussion would be the summary of Figure 2b in which we reorganized the original summary slightly by grouping issues and their corresponding responses.

In summary, the structure of a discussion involving

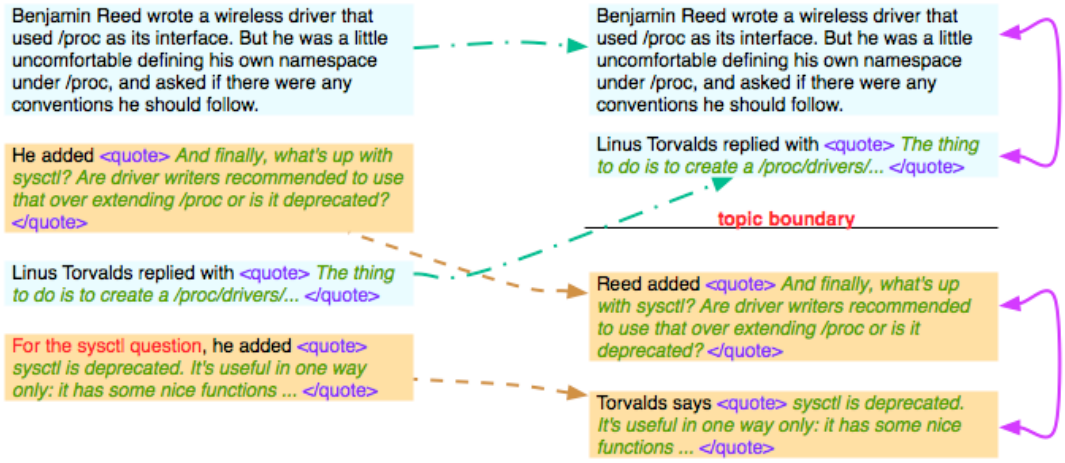

Figure 2. A participant-written summary and its modified version (ideal summary).

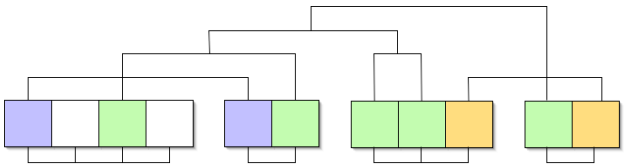

Figure 3. A two-level representation for technical online discussions.

multiple messages can be modeled with a two-level representation, shown in Figure 3. At the lower level, individual messages are partitioned into segments, assuming topic drifts occur linearly. At the higher level, we link segments that are related to each other across messages. Computationally, we use TextTiling (Hearst, 1994) for the intra-message segmentation (lower level in Figure 3) and hierarchical clustering on topic for inter-message-segment linking.

Modeling Interactions

Typical multi-document summarization (MDS) systems focus on content selection and synthesize redundancy and emphasize difference across multiple documents. The primary difference between an MDS system and an online discussion summarization system is that in a discussion multiple participants are involved and discussion topics are being passed back and forth by various participants. MDS systems are insufficient in representing this aspect of the interactions.

From our OSS discussions, we find that the interactions among participants can be modeled as a pair of actions:

1) Seek help or advice
2) Give advice or answer

This pairing behavior is also evident in the ideal summaries that we want to create (Figure 2b). These *problem (initiating, responding)* pairs can be modeled computationally by applying a concept borrowed from conversational analysis, called Adjacent Pairs (AP). Modeling APs has been shown to be effective in analyzing speech transcripts (Galley et al., 2004).

We take the segment that appeared first in time from a cluster as the problem-initiating segment, assuming that no question is answered without being asked first. To find the corresponding problem-responding segment, we use machine-learning methods, such as Maximum Entropy (ME) (Berger, Della Pietra, and Della Pietra; 1996) and Support Vector Machine (SVM) (Joachims, 1998).

The format for the final summary for a discussion is shown in Figure 4.

Blogs

Personal blogs are different from the online discussions that we have described so far. In discussions, it's not superficially clear which participant (and the information provided by him/her) is dominant over other participants

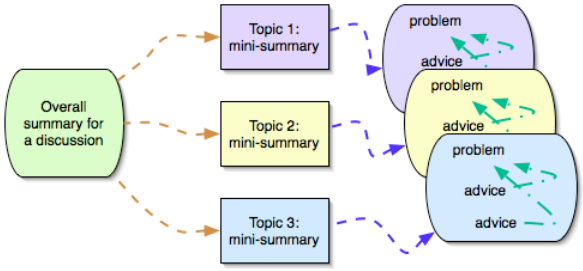

Figure 4. Final summary for a discussion.

(and the information provided by them). In blogs, the degree of interaction varies. In some blogs, we see a diary entry from one person and then comments from others. The original author of the diary entry rarely makes exchanges with his/her readers. In other blogs, authors do respond, but the exchanges are less frequent than those in online discussions.

One interesting feature blogs have is that in each original post, the author often includes URL links to the issues being discussed in the current entry. Is the blog entry a summary of the linked stories? If the blog entry were merely one's opinion, would it make sense to include a summary of those linked stories to provide some background information on the basis of why the author has come to his/her opinion? Is the background information (possibly facts) more important than the entry itself for the readers?

Political Blogs

We chose to investigate a popular blog called The DailyKos (http://www.dailykos.com). There, each log entry can be displayed either alone or together with all of its follow-up comments. Authors write about their opinions on specific topics related to news and politics, which are linked through URLs in the same entry. There are several different types of summaries we can provide to the readers.

If we assume that the blog entries are summaries, with personal opinion added, of the multiple texts (mostly news articles) that they are linked with, we can extract the relevant text segments from the entry as the summary. If this is indeed the case, we can also gain on collecting extraction-based summarization training data of aligned (*abstract, docs*) pairs. Marcu (1999) introduces an algorithm that produces corresponding extracts given (*abstract, text*) tuples with maximal semantic similarity. We modeled Marcu's approach in an opposite way. Since we are assuming that the blog entry contains parts that summarize the linked articles, it's also true that the linked articles, although not compressed at all, have all the important information that needs to be in the summary. So, starting with a complete blog entry, we keep deleting, one at a time, sentences that are not related to the linked articles, until any more deletion would result in a drop in similarity with the linked articles. This way we are assured a set of sentences (or text segments) that are a smaller size than the original, but still express the same amount of

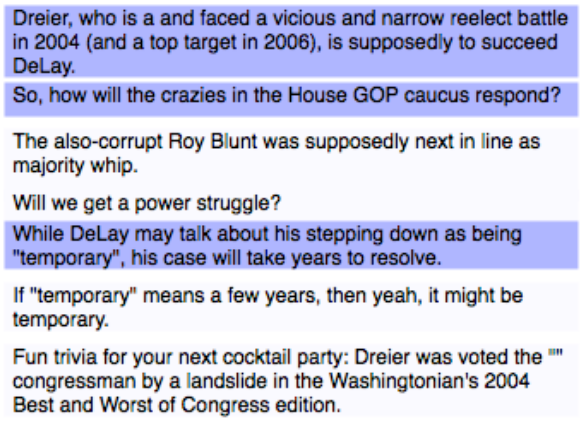

Figure 5. A blog entry and its summary (shaded area).

information as the original, with respect to the set of articles it has links for. Figure 5 shows an example blog entry. The shaded sentences are chosen to be in the summary for this entry.

Some readers are more interested in the "facts" that blog entries bring, instead of focusing on reading individuals' opinions on various topics. So we need to provide summaries of the linked news articles. This problem is simpler and has been the attention of text-based summarization. We applied our multi-document summarizer (Zhou, Lin, and Hovy; 2005), designed predominantly for newswire data, and provided a summary that companies each blog entry. Figure 6 shows the previously shown blog entry and a summary from its linked news articles.

Our analysis on the DailyKos data shows that the level of author-reader interaction is almost none. Sometimes the author of an original blog entry may come back and leave a follow-up comment in the comments section. But this is not common. Our hope is that maybe in other types of blogs people do need to interact and establish collaborative efforts, we would be able to apply the subtopic structure and dialogue modeling that was experimented with in analyzing online discussions.

Evaluation

In the previous sections, we have introduced various ways to create summaries automatically for online discussions and blogs. In this section, we provide a discussion on approaches to evaluate the quality of the machine-generated summaries.

Evaluating Online Discussions

To measure the goodness of system-produced summaries, gold standards are used as references. Human-written summaries usually make up the gold standards. The OSS discussion summary digests are written by Linux experts who actively contribute to the production and discussion of the open source projects. However, participant-produced digests cannot be used as reference summaries verbatim. Due to the complex structure of the dialogue, the summary itself exhibits some discourse structure, necessitating such reader guidance phrases such as "for the … question," "on the … subject," "regarding …," "later in the same thread," etc., to direct and refocus the reader's attention. Therefore, further manual editing and partitioning is needed to transform a multi-topic digest into several smaller subtopic-based gold-standard reference summaries (as previously shown in Figure 2).

We evaluated the online discussion summaries against two baseline systems, in addition to the rewritten reference summaries from human participants. A simpler baseline system takes the first sentence from each message in the sequence that they were posted, based on the assumption that people tend to put important information in the beginning of texts (*Position Hypothesis*). A second baseline system was built based on constructing and analyzing the dialogue structure of each discussion. Participants often quote portions of previously posted messages in their responses. These quotes link most of the messages from a discussion. The message segment that immediately follows the quote is automatically paired with the quote itself and added to the summary and sorted according to the timeline. Segments that are not quoted in later messages are labeled as less relevant and discarded. A resulting baseline summary is an interconnected structure of segments that quoted and responded to one another.

The participant-written summary digests consist of direct snippets from original messages, thus making the reference summaries extractive even after rewriting. This makes it possible to conduct an automatic evaluation. A computerized procedure calculates the overlap between reference and system-produced summary units. Our results show (Zhou and Hovy, 2005) that the summaries created to exploit the subtopic structure of discussions are better than (measured in precision and recall for content coverage) than the two baseline systems. It gains from a high precision because less relevant message segments are identified and excluded from identifying the *problem (initiating, responding)* pairs, leaving mostly topic-oriented segments in summaries. The simpler baseline system had a relatively good performance on recall, which reassured us that the Position Hypothesis still applies in conversational discussions. The second baseline performs extremely well on recall, which showed that quoted message segments, and thereby derived dialogue structure, are quite indicative of where the important information resides. Systems built on these properties are good summarization systems and hard-to-beat baselines.

Text-based Summarization Evaluation

In the previous section, we have discussed how to evaluate online discussion summaries created automatically from modeling discussion subtopic structure. The simple

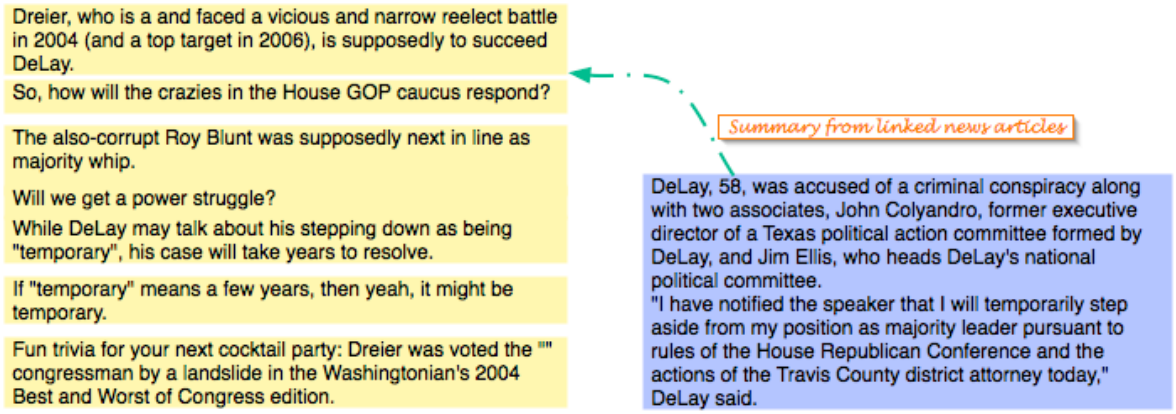

Figure 6. A blog entry and a summary of its linked news articles.

evaluation is reliable because both peer and reference summaries are extraction-based, which allows directed comparison at sentence level.

Creating a suitable summarization evaluation methodology has become an active research area. Many text-based summarization tasks and summaries are not extraction-based, rather abstraction-based. Several manual evaluation techniques have been introduced: SEE (Lin and Hovy, 2002), Factoid (Van Halteren and Teufel, 2003), and Pyramid (Nenkova and Passonneau, 2004). All three methods require human assessors to evaluate the quality of system-produced peer summary against one or more ideal reference summaries. Summaries are represented by a list of summary units (sentences, clauses, etc.).

Naturally, people trust manual evaluation methodologies since humans can infer, paraphrase, and use world knowledge to relate text units that are worded differently. However, there are two major drawbacks to a human evaluation: 1) determining the size of text units being compared, and 2) deciding how much of the *human-ness* to allow. Without a structured definition on the size of the summary units, humans cannot reliably perform this task consistently over multiple texts over a period of time. The second drawback refers to the level of inference that we should allow human assessors to incorporate in evaluation. If there is no constraint at all, then the judgment made on a summary may reflect primarily the assessor's knowledge on the subject being summarized. If we set the scope of inference (providing a list of paraphrases, synonyms, etc.), then the human assessors are behaving like machines making decisions according to rules. The point is that if we leave all the decisions to human, the problem becomes too hard and the results will be debatable.

This is where automated evaluation methodologies can help. We need an automated mechanism that performs both of the above tasks consistently and correctly, and yet correlates well with human judgments. ROUGE (Lin and Hovy, 2003) is an automatic evaluation package that measures the n-gram co-occurrences between peer and reference summary pairs. It was inspired by a similar idea of Bleu (Papineni, et al., 2001) adopted by the machine translation (MT) community for automatic MT evaluation. A problem with ROUGE is that the summary units used in automatic comparison are of fixed length: unigram, bigram, or n-gram. Non-content words, such as stopwords, could be treated equally as content words. A more desired design is to have summary units of variable size where several units may convey similar meaning but with various lengths. Basic Elements (BE) was designed with this idea in mind (Hovy, Lin, and Zhou, 2005). To define summary units, we automatically produce a series of increasingly larger units from reference summaries, starting at single-word level. The focus of BE is on minimal summary units where the unit size is small and paraphrase alternative are limited. Each BE is produced automatically from processing syntactic or dependency parse trees and is defined as:

head-word | head's modifier | head-modifier relation

Head words are the heads of major syntactic or dependent constituents (noun, verb, adjective, or adverbial phrases). BE has been shown to correlate with human summary judgments (on DUC 2002, 2003, 2004, and 2005 results) better than other automated methods.

The difficulties that we face and have been struggling with in text-based summarization evaluation are a forecast for evaluating summaries on dynamically introduced information, namely online discussions and blogs.

Evaluating Blogs: Questions Raised

In the previous two sections, we have discussed various text-based summarization evaluation methodologies and how to evaluate online discussions by factoring in the unique features associated with those discussions. Even though we have produced two different types of summaries on blogs, we have not yet developed a way to correctly quantify the quality of those summaries (intrinsic evaluation).

However, it maybe possible or more straightforward to carry out extrinsic evaluations where we can measure how

the summaries benefit people in performing other tasks. We have deployed a summarization system for online class discussions (Zhou and Hovy, 2005), we found that students and instructors are very open and willing to use NLP technologies. So ideally, we would like to ask the blog readers to rate and comment on how summaries help in achieving their tasks.

Conclusion

In this paper, we described the summarization experiments that we conducted on two types of dynamically introduced information: online discussions and blogs. Although online discussions are closely associated with blogs, evaluating summaries created for blogs are still largely unresolved. The problems that we face are not unique to summarizing blogs or discussions, but rather with summarization in general. But these two new data types allow us to explore further than previously limited newswire data. In analyzing the conversation and dialogue flow, we are provided with insights that may be applied to meeting summarization, which is even more difficult.

References

Berger, A., Della Pietra, S., and Della Pietra, V. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. 2004. Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of ACL 2004*.

Hearst, M. A. 1994. Multi-paragraph Segmentation of Expository Text. In *Proceedings of ACL 1994*.

Hovy, E., Lin, C. Y., and Zhou, L. 2005. Evaluating DUC 2005 using Basic Elements. In *Proceedings of DUC 2005*.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the ECML* 1998.

Lam, D., and Rohall, S. L. 2002. Exploiting E-mail Structure to Improve Summarization. *Technical Paper at IBM Watson Research Center* #20-02.

Lin, C.-Y. and E.H. Hovy. 2002. Manual and Automatic Evaluation of Summaries. In *Proceedings of DUC 2002*.

Lin, C.-Y. and E.H. Hovy. 2003. Automatic Evaluation of Summaries using n-gram Co-occurrence Statistics. In *Proceedings of the HLT-NAACL 2003*.

Marcu, D. 1999. The Automatic Construction of Large-scale Corpora for Summarization Research. In *Proceedings of SIGIR 1999*.

Nenkova, A., and Passonneau, R. 2004. Evaluating Content Selection in Summarization: the Pyramid Method. In *Proceedings of HLT-NAACL 2004*.

Newman, P., and Blitzer, J. 2002. Summarizing Archived Discussions: a Beginning. In *Proceedings of IUI 2002*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. 2001. IBM Research Report Bleu: a Method for Automatic Evaluation of Machine Translation. In *IBM Research Division Technical Report*, RC22176, 2001.

Rambow, O., Shrestha, L., Chen, J., and Laurdisen, C. 2004. Summarizing Email Threads. In *Proceedings of HLT-NAACL 2004: Short Papers*.

Van Halteren, H. and S. Teufel. 2003. Examining the Consensus between Human Summaries: Initial Experiments with Factoid Analysis. In *Proceedings of the HLT-NAACL 2003 Workshop on Automatic Summarization*.

Wan, S., and McKeown, K. 2004. Generating Overview Summaries of Ongoing Email Thread Discussions. In *Proceedings of COLING 2004*.

Zhou, L., and Hovy, E. 2003. A Web-trained Extraction Summarization System. In Proceedings of *HLT-NAACL 2003*.

Zhou, L., and Hovy, E. 2005. Digesting Virtual "Geek" Culture: The Summarization of Technical Internet Relay Chats. In *Proceedings of ACL 2005*.

Zhou, L, and Hovy, E. 2005. Classummary: Introducing Discussion Summarization to Online Classrooms. In *Proceedings of HLT/EMNLP 2005*.

Zhou, L., Lin, C.Y., and Hovy, E. 2005. A BE-based Multi-document Summarizer with Sentence Compression. In *Proceedings of ACL 2005 (MSE workshop)*.