Summarizing Blog Entries versus News Texts

Shamima Mithun and Leila Kosseim
Concordia University
Department of Computer Science and Software Engineering
Montreal, Quebec, Canada
{*s_mithun, kosseim*}@*encs.concordia.ca*

Abstract

As more and more people are expressing their opinions on the web in the form of weblogs (or blogs), research on the blogosphere is gaining popularity. As the outcome of this research, different natural language tools such as query-based opinion summarizers have been developed to mine and organize opinions on a particular event or entity in blog entries. However, the variety of blog posts and the informal style and structure of blog entries pose many difficulties for these natural language tools. In this paper, we identify and categorize errors which typically occur in opinion summarization from blog entries and compare blog entry summaries with traditional news text summaries based on these error types to quantify the differences between these two genres of texts for the purpose of summarization. For evaluation, we used summaries from participating systems of the TAC 2008 opinion summarization track and updated summarization track. Our results show that some errors are much more frequent to blog entries (e.g. topic irrelevant information) compared to news texts; while other error types, such as content overlap, seem to be comparable. These findings can be used to prioritize these error types and give clear indications as to where we should put effort to improve blog summarization.

Keywords

Opinion summarization, blog summarization, news text summarization.

1 Introduction

Everyday, people express their opinions on a variety of topics ranging from politics, movies, music to newly launched products on the web in weblogs (or blogs), wikis, online-forums, review sites, and social networking web sites. As more and more people are expressing their opinions on the web, the Internet is becoming a popular and dynamic source of opinions. Natural language tools for automatically mining and organizing these opinions on various events will be very useful for individuals, organizations, and governments.

Various natural language tools to process and utilize event-related information from texts have already been developed. Event-based question answering systems [21] and event-based summarization systems [12] are only a few examples. However, most of the event-based systems have been developed to process events from traditional news texts. Blog entries are different in style and structure compared to news texts. As a result, successful natural language approaches that deal with news texts might not be as successful for processing blog entries; thus adaptation of existing successful NLP approaches for news texts to process blog entries is an interesting and challenging task. The first step towards this adaptation is to identify the differences between these two textual genres in order to develop approaches to handle this new genre of texts (blogs) with greater accuracy. In this study, we compare automatically generated summaries of blog entries with summaries of news texts with the goal of improving opinion summarization from blog entries. In particular, we compared summaries for these two genres of texts on the basis of various errors which typically occur in summarization.

In this paper, we first investigate what kind of errors typically occur in query-based opinionated summary for blog entries. The errors that we have identified are categorized and then used to compare blog summaries with news texts summaries. For evaluation, we used summaries from participating systems at the TAC 2008 [1] opinion summarization track and updated summarization track. Summaries of the TAC 2008 opinion summarization track and updated summarization track were generated from blogs entries and traditional news texts, respectively. The systems participating in the TAC opinion summarization track and in the updated summarization track are quite different in several aspects, as they are targeted to resolve two different tasks. The systems participating in the updated summarization track were mainly required to find the answers to given queries and detect redundant information while the systems participating in the opinion summarization track were required to perform opinion mining and polarity classification in addition. Moreover, the systems participating in the opinion summarization track were provided optional snippets (described in section 3.1) and were restricted with a maximum summary length which were much higher compared to the updated summarization track. Despite these differences, these two datasets were used in our work because they are the most comparable datasets for our task.

2 Characteristics of Blogs

Blogs (or weblogs) are online diaries that appear in chronological order. Blogs reflect personal thinking and feelings on all kinds of topics including day to day activities of bloggers; hence an essential feature of blogs is their subjectivity. Some blogs focus on a specific topic while others cover several topics; some describe personal daily lives of bloggers while others describe common artifacts or news. Many different sub-genres of blogs exist. The two most common are personal journals and notebooks [5]. Personal journals discuss internal experiences and personal lives of bloggers and tend to be short [5]. They are usually informal in nature and written in casual and informal language. They may contain much and sometimes only unrelated information such as ads, photos, and other non-textual elements. They also contain spelling and grammatical errors, and punctuation and capitalization are often missing. On the other hand, notebooks contain comments on internal and external events. Similarly to newspaper articles, they are usually long and written in a more formal style [5]. Most NLP work on blogs has tended to study personal journals as opposed to notebooks. For example, the Blog-06 corpus [15], used at TREC and at TAC, contains mostly personal journals.

3 Blog Summarization

Opinion summarization, and in particular blog summarization, is a fairly recent field. Some systems (e.g. [9, 10]) have been developed for opinion summarization to generate a summary from a document. In 2008, the Text Analysis Conference (TAC) introduced a query-based opinion summarization track. They provided questions, a blog corpus and optional snippets which are found by QA systems. These query-based summarization systems are designed to retrieve specific answers on an event or entity instead of an overview of the whole document.

Opinion summarization uses opinionated documents such as blogs, reviews, newspaper editorials or letters to the editor to answer opinionated questions. On the other hand, summarization of traditional news texts uses fact-based information such as formal and non-opinionated texts. As we are interested in opinion summarization from blog entries, we will use the two terms *opinion summarization* and *blog summarization* interchangeably.

3.1 Current Approaches

A query-based opinion summarizer recapitulates what people think or feel on a particular topic (or an event or entity) by answering a specific query. For example, one such opinionated query could be *What has been Russia's reaction to U.S. bombing of Kosovo?*. A query-based opinion summarizer can answer opinion questions posed in natural language; thus it helps users to get specific answers to questions they are interested in, instead of retrieving an entire document.

At the TAC 2008 opinion summarization track, a set of target topics on various events or entities were given on which participating systems were evaluated. For each topic, a set of questions and a set of relevant blog entries (mostly personal journals) were provided. For example, for the topic "*UN Commission on Human Rights*", two questions were asked:

1. "*What reasons are given as examples of their ineffectiveness?*"
2. "*What steps are being suggested to correct this problem?*"

and a set of IDs of related blog entries were provided. Systems needed to extract answers to questions from these specified sets of blog entries. Additionally, some sample answer snippets were provided for every topic that summarization systems may use. These snippets were extracted by the participating QA systems at the TAC 2008 QA track. Here are two sample snippets for the topic *UN Commission on Human Rights*:

1. "*Issues regular resolutions condemning Israel while overlooking real offenders.*"
2. "*To ensure this new body would be no facsimile of its predecessor, the legislation prohibits membership to countries that violate human rights or are subject to specific human rights resolutions.*"

Two types of summarization approaches were used by TAC participants, namely: snippet-driven approaches and snippet-free approaches. Snippet-driven approaches use snippet information to extract sentences which contain these snippets from the input blog entries. They then generate a summary by incorporating these sentences. Snippet-free approaches do not use snippets. They mainly utilize query information and sentiment degree for sentence scoring. Participating systems first filter blog entries to identify the relevant content and remove irrelevant information such as ads, photos, music, videos, and other non-textual elements. The focus and polarity of the question are identified; then sentences are ranked according to their relevance with the query. The polarity of the sentences is also calculated and matched with the polarity of the query. To find the relevance with the query, overlap with the query terms is calculated using different techniques such as the cosine similarity, language models etc. Opinion dictionaries and different machine learning techniques are used to identify the polarity of the question and sentences. Finally, the summaries are generated using the ranked sentences.

3.2 Evaluation

Evaluation of blog summaries use the same criteria as for traditional news text summarization. The quality of a summary is assessed mostly on its content and linguistic quality [14]. Content evaluation of a query-based summary is performed based on the relevance assessment (with the topic and query) and inclusion of important contents from the input documents.

Currently, the automatic evaluation tool ROUGE [11] is the most popular evaluation approach for content evaluation. ROUGE automatically compares system generated summaries with a set of model summaries (human generated) by computing n-gram word overlaps between them. Conferences and workshops such as TAC and DUC (Document Understanding Conference) [2] use ROUGE. The pyramid method [18] is also used for content evaluation. In the pyramid method, multiple human generated summaries are analyzed manually to generated a gold standard. In this process, summary analysis is done semantically such that information with the same meaning (expressed using different wording) is marked as summary content unit (SCU). A weight is assigned for each SCU based on the number of human summarizers that express it in their summaries. In this method, the pyramid score for a system generated summary is calculated as follows [17]:

score = *(the sum of weights of SCUs expressed in a generated summary) / (the sum of weights of an ideally informative summary with the same number of SCUs)*

The linguistic quality of a summary is evaluated manually based on how it structures and presents the contents. Grammaticality, non-redundancy, referential clarity, focus, structure and coherence are the commonly used factors considered to evaluate the linguistic quality. Mainly, subjective evaluation is done to assess the linguistic quality of an automatically generated summary. In this process, human assessors directly assign scores on a scale based on agreement or disagreement with predefined set of questions such as "*Are they ungrammatical?*", "*Do they contain redundant information?*". The assessments are done without reference to any model summaries.

3.3 News Text Summarization versus Blog Summarization

As most work has been done on news text summarization, it is not surprising that the performance of such systems are generally higher than blog summarizers. For example, as shown in Table 1, at the TAC-2008 conference, the average scores for news text summaries (updated summarization track) are higher than for blog summaries (opinion summarization track) using all 3 evaluation criteria.

Table 1: *Average TAC-2008 Summarization Results - Blogs vs. News Texts*

Genre	**Pyramid Score**	**Linguistic Score**	**Resp. Score**
Blogs	0.21	2.13	1.61
News	0.27	2.33	2.32

Table 1 shows summary evaluation using the pyramid score, linguistic quality and responsiveness (Resp.). The last two criteria were evaluated by human assessors on a scale of 1 to 5 (1, being the worst). In this evaluation, the responsiveness of a summary was judged to measure the overall quality or usefulness of the summary, considering both the information content and readability.

This difference in performance between blogs and news texts can be attributed to the differences in the two textual genres. Indeed, one of the essential characteristics of blogs as opposed to news texts, is their subjectivity (or opinion). Unlike traditional news text summarization, sentiment (subjectivity) plays a key role for blog summarization. For blog summarization, sentiment degree is often used to rank sentences. In the case of query-based blog summarization, the sentiment polarity of the question needs to be matched with that of summary sentences.

In addition, as opposed to traditional news texts, blogs are usually written in casual language. For blogs, it is usually very difficult to identify which portions of blog entries are relevant to the topic. News texts are more uniform in style and structure. Blogs may contain many unrelated information such as ads, photos, music, videos. For blogs, it is often difficult to find sentence boundaries. In most cases punctuation and capitalization are unreliable. As a result, for blog summarization, systems need to put additional efforts to pre-process the text compared to news text summarization. Furthermore, because blogs do not exhibit a stereotypical structure, some features such as position of sentence, or similarity with the first sentence, which are shown to be useful for traditional news text summarization are not as useful for blog summarization [6].

4 Error Analysis

To identify the different challenges posed by blog summarization as opposed to traditional news texts summarization, we have studied 50 summaries from participating systems at the TAC 2008 opinion summarization track and compared these to 50 summaries from the TAC 2008 updated summarization tracks. The average summary length of the opinion summarization track was 1224 words, while that of the updated summarization track was 179 words. The average input documents length of the opinion summarization track was 1888 words, while that of the updated summarization track was 505 words. Summaries were randomly selected for the evaluation. However, we ensured that we selected summaries from all participating systems on all topics. The task of the updated summarization track was chosen for comparison because it is similar in nature to the blog summarization task in the sense that its goal is also to generate query focused (but non-opinionated) summaries (using news articles). Even though there are several differences between the summarization approaches in TAC opinion summarization track and updated summarization track, these two datasets are the most comparable datasets for our task.

In this study, we have analyzed the most common types of errors in our 100-summary corpus and have categorized them in 3 main categories:

1. *Summary-Level Error (SuLE)*

2. *Sentence-Level Error (SeLE)*

3. *Intra-Sentence-Level Error (ISLE)*

These are shown in Figure 1 and discussed in the following sub-sections.

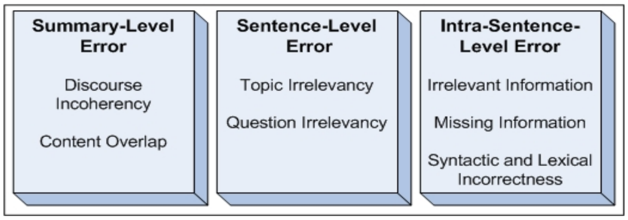

Fig. 1: *Types of Errors in Blog vs. News Summaries*

4.1 Summary-Level Errors

We define a Summary-Level Error (SuLE) as the textual contents which reduce the understandability and readability of the overall summary. There are two types of SuLE:

1. *Discourse Incoherency (DI)*
2. *Content Overlap (CO)*

Discourse Incoherency (DI)
A summary will exhibit a Discourse Incoherency (DI) if the reader cannot identify the communicative intentions of the writer from the propositions or if the propositions do not seem to be interrelated [8]. In the sample summaries that we have studied, Discourse Incoherency occurred both at the sentence level and at the proposition level.

Consider the following summary (ID:T1004.20 [1]) where a DI occurs at the sentence level:

Topic: *Starbucks coffee shops*
Question: *Why do people like Starbucks better than Dunkin Donuts?*
Summary: *I am firmly in the Dunkin' Donuts camp. It's a smooth, soothing cuppa, with no disastrous gastric side effects, very comforting indeed. I have a special relationship with the lovely people who work in the Dunkin' Donuts in the Harvard Square T Station in Cambridge. I was away yesterday and did not know. [...]*

In this summary, the underlined sentence is not coherent with the rest of the text because it is not interrelated with the rest of the text.

Consider the following summary (ID:T1001.3) where a DI occurs at the proposition level:

Topic: *Carmax*
Question: *What motivated positive opinions of Carmax from car buyers?*
Summary: *At Carmax, the price is the price and when you want a car you go get one. Tyler Sink Says: December 7th, 2005 at 9:22 am, with Carmax you will generally always pay more than from going to a good used car dealer. [...]*

In the underlined sentence, the textual fragments *Tyler Sink ..* and *with Carmax ..* do not seem to be related to one another.

Content Overlap (CO)
If semantically similar information exists in different units of a text, then we define it as Content Overlap (CO). Content overlap can range from a simple duplication of text fragments to a more complex textual entailment problem. For example, consider the summary below (ID:T1019.35):

Topic: *China one-child per family law*
Question: *What complaints are made about China's one-child per family law?*
Summary: *[...] If you have money to pay the fines, you can have 2 or 4 children. [...] $6400 - a typical fine for having more than one child- in China is about 2-3 years salary. [...] Imagine losing your job, being fined 2-3 years salary for having a second child. [...]*

In this summary, the underlined sentences carry similar contents. So it may seem redundant to include all these sentences in the final summary.

Table 2: *Summary-Level Errors - Blogs vs. News Texts*

Error Type		Blogs	News	Δ
DI	Discourse Incoherency	30.44%	10.66%	19.78%
CO	Content Overlap	19.14%	14.66%	4.48%

Table 2 compares Summary-Level errors in our 50 blog summaries corpus and our 50 news texts summaries corpus. Table 2 shows that opinionated blog summarization and non-opinionated news texts summarization both exhibit an important number of *Discourse Incoherency* and *Content Overlap* errors. However, blog summarization have around 20% more *Discourse Incoherency* and about 4.5% more *Content Overlap* errors, than those of news article summarization. We suspect that the reason behind this is that blogs are generally informal in nature. As a result, in blogs, propositions are often incoherent and contain redundant information. On the other hand, the formal nature of news articles reduces these errors for news texts summarization.

4.2 Sentence-Level Errors

If a summary sentence is irrelevant to the central topic of the input documents or to user query, then the summary contains a Sentence-Level Error (SeLE). Two types of SeLE were identified:

1. *Topic Irrelevancy (TI)*

[1] All summaries numbered ID:Txxxx.xx are taken from the TAC 2008 opinion summarization track.

2. *Question Irrelevancy (QI).*

Topic Irrelevancy (TI)
As mentioned in Sections 3.1 and 4, in both the TAC 2008 opinion summarization track (blogs) and the updated summarization track (news texts), participating systems needed to generate a summary answering a set of questions on a specific target (topic). However, in both tasks, many systems generated a summary containing sentences that are not related to the specified topic. Here is an example of a TI (ID:T1004.33):

Topic: *Starbucks coffee shops*
Question: *Why do people like Starbucks better than Dunkin Donuts?*
Summary: *Well ... I really only have two. [...] I didn't get a chance to go ice-skating at Frog Pond like I wanted but I did get a chance to go to the IMAX theatre again where I saw a movie about the Tour de France it wasn't that good. [...]*

Question Irrelevancy (QI)
Many of the system generated summary sentences are not relevant to the question even though they are related to the topic. An example of a QI is shown below (ID:T1004.3):

Topic: *Starbucks coffee shops*
Question: *Why do people like Starbucks better than Dunkin' Donuts?*
Summary: *Posted by: Ian Palmer — November 22, 2005 at 05:44 PM Strangely enough, I read a few months back of a coffee taste test where Dunkin' Donuts coffee tested better than Starbucks. [...] Not having a Dunkin' Donuts in Sinless City I am obviously missing out... but Starbucks are doing a Christmas Open House today where you can turn up for a free coffee. [...]*

The underlined sentence is relevant to the topic but not to the question.

Table 3: *Sentence-Level Errors - Blogs vs. News Texts*

Error Type		Blog	News	Δ
TI	Topic Irrelevancy	41.67%	5.86%	35.81%
QI	Question Irrelevancy	47.87%	16.67%	31.20%

Table 3 compares Sentence-Level errors for blog summaries and news text summaries. Note that in the table, *Topic Irrelevancy* is calculated based on the entire corpus. However, *Question Irrelevancy* is calculated based only on the sentences which are related to the topic. Table 3 shows that a large number of sentences from blog summaries have *Topic Irrelevancy* and *Question Irrelevancy* errors. In contrast, in news text summarization, *Topic Irrelevancy* error occurs only occasionally and *Question Irrelevancy* error is also not very frequent. Blogs summarization has around 30% more of these two errors than that of news text summarization. We suspect that the main reason behind such a difference is brought about by the summary evaluation scheme. Indeed, many systems use the optimal summary length (7000 characters per question) allowed in TAC which results in many out of context sentences to be used as filler. As a result, the average summary length of the opinion summarization track is much longer than that of the updated summarization track (1224 words versus 179 words). Another important reason for these errors is the informal style and structure of blog entries. Indeed, sentences in blog entries do not have a predictable rhetorical structure (e.g. in formal writing, the first and the last sentences of a paragraph usually contain important information) which can be used to rank sentence during summarization. As a result, it is much more difficult to rank blog sentences compared to news text sentences. Opinion (sentiment) information is typically used to rank blog sentences for summarization. We also believe that because opinion identification can be quite imprecise, it can possibly add more noise to the blog sentence ranking process. Moreover, unlike pre-focused news articles, blogs are quite unfocused. In blogs, bloggers express various opinions about the topic which are not relevant to the question. Together all these issues may lead to a high number of topic and question unrelated sentences in blog summarization.

4.3 Intra-Sentence-Level Errors

Intra-Sentence-Level (ISLE) errors occur within a sentence and involve irrelevant or missing information, grammatical errors, or lexical errors (e.g. spelling errors). Intra-Sentence-Level Errors include:

1. *Irrelevant Information (II)*
2. *Missing Information (MI)*
3. *Syntactic and Lexical Incorrectness (SLI)*

Each of these categories are described below with examples.

Irrelevant Information (II)
Under Irrelevant Information (II) errors, a significant portion of a sentence is irrelevant to the summary topic or question. For example, consider the summary below (ID:T1003.9):

Topic: *Jiffy Lube*
Question: *What reasons are given for liking the services provided by Jiffy Lube?*
Summary: *They know it's fine cause Jiffy Lube sent them a little card in the mail and they have about a month before they need an oil change. [...] Well, they suppose it is a little bit of a PITA to figure out what to do with the spent oil, but after some digging, they found out that every Jiffy Lube will take used oil for free! [...]*

The underlined snippet above is irrelevant to the question even though it holds a coherent discourse relation with the last proposition.

Missing Information (MI)
If a sentence does not contain all the necessary information to make it comprehensible for the reader and the required information to understand the sentence is also not available in the context then this error is defined as a Missing Information (MI) error.

Here is an example of MI. In the following summary (ID:T1021.17):

Topic: *Sheep and Wool Festival*
Question: *Why do people like to go to Sheep and Wool festivals?*
Summary: *[...] i hope to go again this year and possibly meet some other knit bloggers this time around since i missed tons of people last year. I love going because of the tons of wonderful people, yarn, Sheep, rabbits, alpacas, llamas, cheese, sheepdogs, fun stuff to buy, etc. , etc. [...]*

The underlined sentence contains incomplete information, which cannot be resolved from the context either making it incomprehensible.

Syntactic and Lexical Incorrectness (SLI)
Syntactical level errors such as grammatical incorrectness and incompleteness of a sentence or lexical level errors such as spelling errors, short forms, stylistic twists of informal writing ... in a sentence is defined as Syntactic and Lexical Incorrectness (SLI) error.

For example, consider the following summary (ID:T1009.32):

Topic: *Architecture of Frank Gehry*
Question: *What compliments are made concerning his structures?*
Summary: *Central to Millennium Park in Chicago is the Frank Gehry-designed Jay Pritzker Pavilion, described as the most sophisticated outdoor concert venue of its kind in the United States. [...] Designing a right-angles-be-damned concert hall for Springfield, hometown of Bart et al.. [...]*

In this summary, the underlined sentence is an example of a SLI.

Table 4: *Intra-Sentence-Level Errors - Blogs vs. News Texts*

Error Type		Blog	News	Δ
II	Irrelevant Information	30.91%	15.66%	15.25%
MI	Missing Information	9.33%	2.33%	7.00%
SLI	Syntactic and Lexical Incorrectness	18.79%	4.00%	14.79%

Table 4 compares Intra-Sentence-Level errors for blog summaries and news text summaries. From Table 4, we can see that *Irrelevant Information*, *Missing Information*, and *Syntactic and Lexical Incorrectness* errors appear about 15%, 7%, and 15% more respectively in blog summarization. Here again, we believe that the informal nature of blogs explains these difference.

5 Discussion

Compared to a manual linguistic evaluation of a summary, our work tries to identify and quantify the differences in error types between two textual genres: blogs and news.

Our error types incorporate both what the automatic and manual summary evaluation try to measure. Indeed, Sentence-Level Errors (Topic Irrelevancy and Question Irrelevancy) evaluate the content and relevance of the summaries similarly to what ROUGE tries to evaluate; whereas the remaining errors (Summary-Level Errors and Intra-Sentence Errors) evaluate more the linguistic quality of a summary.

It is not surprising to see that Topic Irrelevancy, Question Irrelevancy, Discourse Incoherency, Irrelevant Information and Syntactic and Lexical Incorrectness are much more frequent in blogs than in news texts (from 36% to 19% more frequent). Content Overlap and Missing Information, on the other hand, seem to be only slightly more frequent (5% and 7%) in blogs summaries than in news texts summaries. These results give a clear idea of how difficult it is to process blog entries for summarization compared to news texts and where efforts should be made to improve such summaries.

6 Related Work

6.1 NLP on blogs

Recently, the availability of opinions on current events on weblogs opened up new directions in natural language research. Even though natural language processing on blogs is a fairly new trend, its popularity is growing rapidly. Many conferences and workshops (e.g. [1, 3, 4, 15]) are taking place to address different aspects of the analysis of blog entries. Current NLP work on blog entries include: subjectivity and sentiment analysis; question answering; and opinion summarization.

Subjectivity and sentiment analysis include classifying sentiments of reviews [19] and analyzing blogger mood and sentiment on various events [16]. Sentiment classification of reviews on different events is often done on movie or product reviews. Rating indicators of reviews are used to identify the polarity of the blogs namely positive, negative or neutral. To analyze blogger mood and sentiment, systems make use of information regarding bloggers' mood varying over time. To record bloggers' varying mood, the polarity information of the blog post is often used. Some works (e.g. [16]) are done to measure how bloggers' varying mood affects different events. In addition, the TREC blog track [15] provides an opportunity to build new techniques of sentiment tagging on blog posts. The task is to identify and rank blog posts on a given topic from a corpus of blog entries.

Question answering (QA) on blog entries is a relatively new field. Most notable QA work on blog entries was conducted at TREC 2007 [15] and TAC 2008 [1]. To answer queries on an event or entity, TREC provided a blog corpus in addition to the AQUAINT newspaper corpus [15].

6.2 Analysis of blogs versus news

To the best of our knowledge there have been only a few work carried out to compare the difference between blog entries and news texts; however, none seems to have analyzed it at the linguistic level for a specific NLP application.

Ku et al. [10] developed a language independent opinion summarization approach. For summarization, they retrieved all sentences which are relevant to the main topic of the document set and determined the opinion polarity and degree of these relevant sentences. They also found that the identification of correlated events on a time interval is also important for opinion summarization. They tested their approach for blog entries and news texts for English and Chinese languages. From their evaluation, they found that blog entries contain more topic irrelevant information compared to news texts. Their results confirm our own results. Ku et al. also found that news texts use a larger vocabulary compared to blog entries which makes the filtering of non-relevant sentences task harder for news texts. On the other hand, this larger vocabulary helps to decide sentiment polarities. Due to the limited vocabulary the judgment of sentiment polarity of blog entries was difficult.

Somasundaran et al. [20] developed an opinion question answering approach for blogs and news texts. They exploited attitude information namely sentiment and argument types to answer opinion questions. They received comparable result with both text types.

Lloyd et al. [13] developed the Lydia system to analyze blog entries. They analyzed temporal relationship between blogs and news texts. In particular, they analyzed how often bloggers report a story before newspapers and how often bloggers react to news that has already been reported. To study this leads/lag relationship, they analyzed frequency time series of 197 most popular entries in news texts and blog corpora over six week period. Lydia first recognized name entities to extract information from both corpora. Then the system resolved noun phrase coreference because a single entity is often mentioned using multiple variations on their name. Then it performed a temporal analysis to identify which entities are referred more frequently over a certain period of time. In their analysis, they found that 30 entities exhibited no lead/lag relationship, 73 had news leading the blogs, and 94 had blogs leading the news.

Godbole et al. [7] developed a large-scale sentiment analysis system on top of the Lydia text analysis system [13] for news texts and blog entities. They determined the public sentiment on various entities and identified how this sentiment varies with time. They found that the same entities (person) except certain controversial political figures received comparable opinions (favorable or adverse) in blogs and news texts. Controversial political figures received different opinions in blogs compared to news texts because of the political biases among bloggers, and perhaps the mainstream press.

Though both the work Lloyd et al. and Godbole et al. handle news text and blog entries, their application domains (temporal relationship and sentiment analysis) are different from ours. Somasundaran et al. tested their question answering approach for news texts and blogs. They compared their approach for both genres of text mainly on the basis of subjectivity information. On the other hand, we compared summaries of both text types on the basis of errors which mainly occurred from informal style and structure of blog entries. Our work is most similar to Ku et al.'s work. However, we identified a larger number of errors of summarization and compared blog summaries with traditional news texts summaries on the basis of these errors. As a result, our work will better enable us to pinpoint the difference between these two genres of texts for summarization task.

7 Conclusion

As the performance of blog summarization is generally much lower than for news text summarization, we set out to compare automatically generated summaries for blogs entries with news texts based on the most common errors which occurred in summarization. The goal of our comparison was to assess whether these summary related errors affect traditional news texts based non-opinionated summaries differently than opinionated blog summaries.

We first analyzed and categorized errors that occur in opinion summarization on blogs using the summaries from participating systems at the TAC 2008 opinion summarization track. Then we compared these results with those of the TAC 2008 updated summarization track. Our results show that all types of summary related errors occur more often in blog summarization than news texts summarization. However, topic and question irrelevancy pose a much greater problem for blog summarization than for traditional news texts; while content overlap and missing information seem to be only slightly more frequent in blog than traditional news texts. These findings can be used to prioritize these error types and give clear indications as to where we should put effort to improve blog summarization.

8 Acknowledgements

The authors would like to thank the anonymous referees for their valuable comments on an earlier version of the paper.

This work was financially supported by NSERC.

References

[1] Text Analysis Conference (TAC): http://www.nist.gov/tac. (Last accessed 2009-05-20).

[2] Document Understanding Conferences (DUC): http://duc.nist.gov. (Last accessed 2009-05-20).

[3] Third International AAAI Conference on Weblogs and Social Media, San Jose, California, May 2009.

[4] Third Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics. In *Workshop Of WWW-2006*, Edinburgh, May 2006.

[5] A. Andreevskaia, S. Bergler, and M. Urseanu. All Blogs are Not Made Equal: Exploring Genre Differences in Sentiment Tagging of Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2007)*, Boulder, Colorado, March 2007.

[6] A. Bossard and M. Genereux. Description of the LIPN Systems at TAC 2008: Summarizing Information and Opinions. In *Notebook Papers and Results, Text Analysis Conference (TAC-2008)*, Gaithersburg, Maryland, USA, November 2008.

[7] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-Scale Sentiment Analysis for News and Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM'2007)*, pages 219–222, Boulder, Colorado, USA, March 2007.

[8] E. H. Hovy. Automated Discourse Generation using Discourse Structure Relations. *Artificial Intelligence*, 63(1-2):341–385, 1993.

[9] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *SIGKDD 2004*, pages 168–177, 2004.

[10] L. W. Ku, Y. T. Liang, and H. H. Chen. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of the AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, California, USA, March 2006.

[11] C. Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004.

[12] M. Liu, W. Li, M. Wu, and H. Hu. Event-Based Extractive Summarization using Event Semantic Relevance from External Linguistic Resource. In *Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology, ALPIT 2007*, pages 117–122, Henan, China, 2007.

[13] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. Blogs: Who Gets the Scoop? In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006*, California, USA, March 2006.

[14] A. Louis and A. Nenkova. Automatic Summary Evaluation without Human Models. In *Notebook Papers and Results, Text Analysis Conference (TAC-2008)*, Gaithersburg, Maryland (USA), November 2008.

[15] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, Gaithersburg, Maryland, USA, November 2007.

[16] G. Mishne and N. Glance. Predicting Movie Sales from Blogger Sentiment. In *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.

[17] A. Nenkova. Summarization Evaluation for Text and Speech: Issues and Approaches. In *Proceedings of Interspeech 2006*, Pittsburg, USA, 2006.

[18] A. Nenkova and R. Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the HLT/NAACL*, 2004.

[19] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.

[20] S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov. QA with Attitude: Exploiting Opinion Type Analysis for Improving Question Answering in On-line Discussions and the News. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, Colorado, USA, March 2007.

[21] H. Yang, T. S. Chua, S. Wang, and C. K. Koh. Structured use of External Knowledge for Event-based Open Domain Question Answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, Toronto, Canada, 2003.