



# Summarizing Blog Entries versus News Texts

---

Shamima Mithun  
Leila Kosseim

Concordia University  
Montreal, Canada



# Outline

---

- Motivation
- Goal
- Background
- Error Analysis
  - Error Identification
  - Error Categorization
  - Comparison of blog summarization with news texts summarization based on these errors
- Related Work
- Conclusion



# Motivation

---

- People express their opinions in blogs.
  - Automatically mining and organizing these opinions is very useful.
  - NLP tools to process and utilize information from texts are available, BUT:
    - Most of these systems are targeted for news texts.
    - and are not as useful for blogs because blogs and news texts are much different in style and structure.
- > Adaptation of NLP approaches for news texts to process blogs is an interesting and challenging task.



# Goal

---

- The first step towards this adaptation is to identify the differences between news texts and blogs.
- We compared automatically generated summaries of blog entries VS of news texts.
  1. identified the types of errors that typically occur in query-based opinionated summary for blog entries,
  2. then categorized these errors according to their sources,
  3. and compared these errors to news texts summaries.



# Background: Characteristics of Blogs

---

## **Blogs:**

- are online diaries that appear in chronological order.
- reflect personal thinking and feelings on all kinds of topics including day to day activities of bloggers.

## **Characteristics:**

- Subjective in nature.
- Written in casual and informal language.
- Usually contain unrelated information to the main topic.
- May contain spelling and grammatical errors.
- Punctuation and capitalization are often missing.



# Background: Blog Summarization

---

## TAC 2008 Opinion Summarization:

- In 2008, the Text Analysis Conference (TAC) introduced a query based opinion summarization track.
- TAC provided:
  - 22 target topics
  - For each topic:
    - 2 questions (on average)
    - 9 to 39 relevant blog entries
    - optionally, sample answer snippets extracted from the participating QA systems at the TAC 2008 QA track.



# Background: TAC 2008 Opinion Summarization

---

- **Goal:**
  - For each question, generate a summary from the specified sets of blog entries about the target that answers the question.
- **Corpus:**
  - Source: subset of Blog06 collection.
  - Size: 537 blogs of average length of 1888 words.
- **Evaluation:**
  1. summary's *content*
    - use the pyramid method for scoring [0-1]
  2. summary's *linguistic quality*
    - manual subjective score [0-10]
  3. summary's *overall responsiveness* score [0-10] which reflects both content and readability.



# Background: TAC 2008 Opinion Summarization

---

## Topic:

*UN Commission on Human Rights*

## Questions:

*What reasons are given as examples of their ineffectiveness?*

*What steps are being suggested to correct this problem?*

## Optional snippets:

*Replace it with a more credible body.*



# Background: TAC 2008 Update Summarization

---

- TAC provided:
  - 48 target topics
  - For each topic:
    - 1 question
    - 20 relevant documents divided into 2 sets:
      1. Document Set A (10 docs)
      2. Document Set B (10 docs)
- **Goal:**
  - Generate 2 summaries:
    1. one from Set A: a simple query-focused summary.
    2. one from Set B: also query-focused but should be written under the assumption that the reader of the summary has already read the documents in Set A.



# Background: TAC 2008 Update Summarization

---

- **Corpus:**
  - Source: Subset of AQUAINT-2 collection.
  - Size: 960 news articles of average length of 505 words.
- **Evaluation:**
  - Similar evaluation metrics as of opinion summarization.
- **Example:**
  - Topic:** *Airbus A380*
  - Question:** *Describe developments in the production and launch of the Airbus A380.*



## Background: News Text summarization vs. Blog Summarization

---

- The performance of news summarization systems are generally better than blog summarizers.
  - Blog Track, 45 runs from 19 teams
  - News Track, 71 runs from 33 teams

<b>Genre</b>	<b>Pyramid Score</b>	<b>Linguistic Score</b>	<b>Resp. Score</b>
Blogs (Average)	0.21	2.13	1.61
News (Average)	0.27	2.33	2.32
Blogs (Best)	0.49	2.26	2.88
News (Best)	0.36	3.25	2.79

Table 1: TAC-2008 summarization results – blogs vs. News.



# Error Analysis

---

- To identify the errors which typically occur in 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 update summarization tracks.
- Even though there are several differences between the summarization approaches, these two datasets are the most comparable datasets for our task.



# Error Types

---

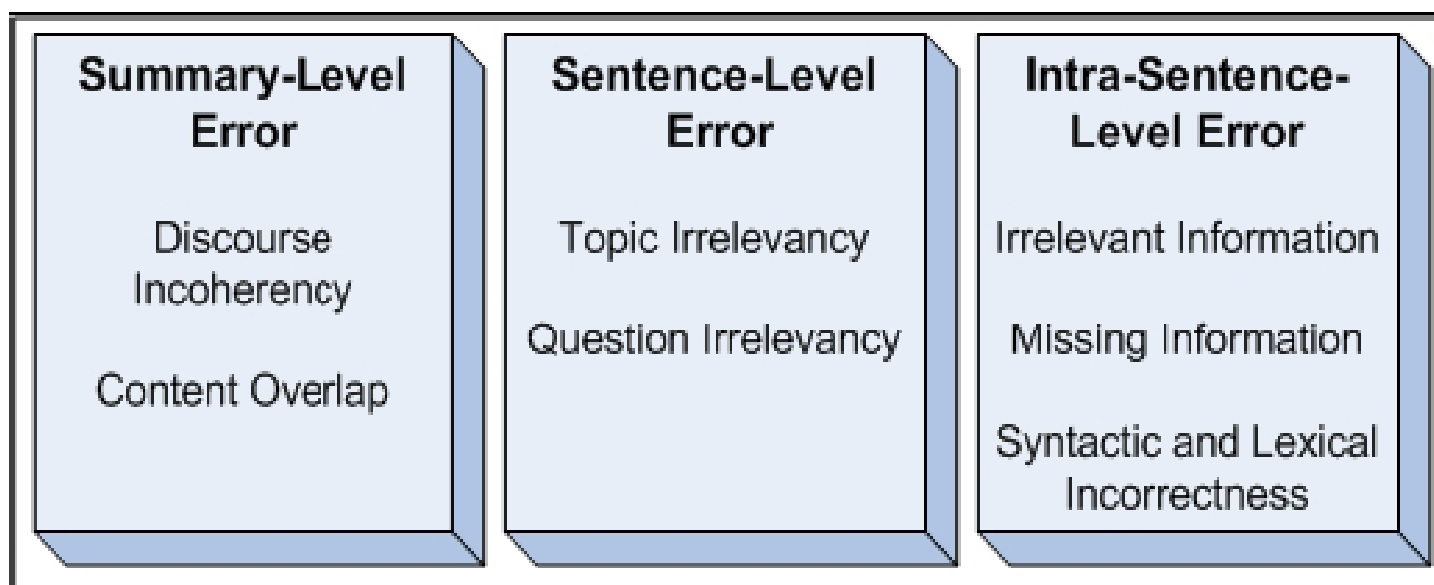


Figure 1: Types of errors in Automatic Summarization



# Summary-Level Errors

---

- Discourse Incoherency:

**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.*



# Summary-Level Errors

---

- Content Overlap

**Topic:** *China's one-child per family law*

**Question:** *What complaints are made about China's one-child per family law?*

**Summary:** *[...] If you have \$6400 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. [...]*

# Summary-Level Errors

Error Type	Blogs	News	Blogs-News
Discourse Incoherency	30.44%	10.66%	19.78%
Content Overlap	19.14%	14.66%	4.48%

Table 2: Summary-Level Errors – Blogs vs. News

may be due to the informal nature of blogs.

could be that input documents contain the same information multiple times.



# Sentence-Level Errors

---

- Topic Irrelevancy

**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. [...]*



# Sentence-Level Errors

---

- Question Irrelevancy

**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. [...]*

# Sentence-Level Errors

Error Type	Blogs	News	Blogs–News
Topic Irrelevancy	41.67%	5.86%	35.81%
Question Irrelevancy	47.87%	16.67%	31.20%

Figure 3: Sentence-Level Errors Blogs vs. News

The summary evaluation scheme.

The informal style and structure of blog entries.

Incorrect opinion identification.



# Intra-Sentence-Level Errors

---

- Irrelevant Information

**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! [...]*



# Intra-Sentence-Level Errors

---

- Missing Information

**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. [...]*



# Intra-Sentence-Level Errors

---

- Syntactic and Lexical Incorrectness

**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.. [...]*

# Intra-Sentence-Level Errors

Error Type	Blogs	News	Blogs-News
Irrelevant Information	30.91%	15.66%	15.25%
Missing Information	9.33%	2.33%	7.00%
Syntactic & Lexical Incorrectness	18.79%	4.00%	14.79%

Figure 4: Intra-Sentence-Level Errors – Blogs vs. News

informal nature of blogs explains these difference.



# Related Work

---

- Some work (e.g. [Lloyd et al. and Godbole et al.]) handle news text and blog entries but their application domains are different from ours.
- Somasundaran et al.:
  - compared their question answering approach for blogs and news texts on the basis of subjectivity information.
  - we compare summaries of both text types on the basis of typical errors.



# Related Work

---

- Ku et al.'s work
  - Most similar to our work
  - Developed a document based opinion summarization approach.
  - Found that blog entries contain more topic irrelevant information compared to news texts.
  - Also analyzed effects of the size of vocabulary of the input documents in case of relevance assessment and polarity identification.

We identified a larger number of errors of summarization and compared blog summaries with news text summaries on the basis of these errors.



# Conclusion

---

- In general, all types of summary related errors occur more often in blog summarization than news texts summarization, however:
- Much greater problem for blog summarization than news texts:
  - Topic irrelevancy (41.67% vs. 5.86%) and
  - Question irrelevancy (47.87% vs. 16.67%)
- Only slightly more frequent in blog than news texts:
  - Content overlap (19.14% vs. 14.66%) and
  - Missing information (9.33% vs. 2.33%)



## Future Work

---

- Identify the sources of these errors:
  - Input document sets
  - Summarization systems
- Our findings can be used to prioritize these error types and give clear indications as to where we should put effort to improve blog summarization.