Extractive email thread summarization: Can we do better than He Said She Said?

Pablo Ariel Duboue (pablo.duboue@gmail.com)
Les Laboratoires Foulab – 999 du College, Montreal, Quebéc

Abstract

Human-written, good quality extractive summaries pay great attention to the text intermixing the extracts. In this work, we focused on the lexical choice for verbs introducing quoted text. We analyzed 4000+ high quality summaries for a high traffic mailing list and manually assembled 39 quotation-introducing verb classes that cover the majority of the verb occurrences. A significant amount of the data is covered by on-going work on e-mail "speech acts." However, we found that one third of the "tail" is composed by "risky" verbs that most likely will be beyond the state of the art for longer time. We used this fact to highlight the trade-offs of risk taking in NLG, where interesting prose might come at the cost of unsettling some of the readers.

Focus: appropriate and varied ways to quote email threads.

A seemingly simple task, this problem touches:

- Speech act detection [Searle, 1975] (question vs. announcement vs. reply)
- Opinion mining [Pang and Lee, 2008] (complained vs. thanked)
- Citation polarity analysis [Teufel, 1999]: (agreed vs. disagreed vs. added)

Our Data: Linux Kernel mailing list (LKML) summaries.

- Kernel Traffic summaries of the activities in LKML
  - LKML: extremely high traffic (300 mails a day on average)
  - For five years (since 1999), Jack Brown published a weekly summaries for hand-picked most newsworthy threads.
  - The summaries are XML-encoded available under a GPL
  - Explicit marking of all quoted text, with attribution
- The summaries are quite entertaining and with a prose quality
  - Great training data for NLG
- We study the introduction of quoted references in a rich manner
  - In the 4,253 summaries (344 newsletters) 95% contain a quote
  - An average of 3.28 quotes per summary
  - 72% of total chars in the summaries are inside quotes (including markup)

Kernel Traffic #6, Feb. 18th 1999 (excerpt).

```
<p>Gregory Maxwell replied,
  <quote who="George Maxwell">Do you see the "(sic)" That usually stands
  for "Spelling is Correct".</quote>
</p>
<p>Oliver Xymoron rejoined:</p>
<quote who="Oliver Xymoron">
  <p>I think what we have here is an ironic double typo. The message is
  actually indicating the drive is not feeling very good:</p>
  <p>+        { 0xb900, "Play operation aborted (sick)" },</p>
  <p>Hopefully this very important change will make it into 2.2.2.</p>
</quote>
<p>Brendan Cully kafloogitated:</p>
<quote who="Brendan Cully">
  <p>"sic" doesn t stand for "spelling is correct", or even "stated in
  context" (yech!).</p>
  <p>In fact, it stands for "yes, I know it looks funny, but that's how
  I want it". But people got tired of typing Y,IKILF,BTHIWI so they
  abbreviated it to SIC.</p>
</quote>
```

The loss to its readership is summed by this post to the LKML by Jean Delvare on Feb. 8th 2006:

> *I'd like to thank you a lot for all the good work you have done on kerneltraffic in the past few years. It was amazing! I am not too surprised to see you stop now - I can imagine the amount of work it was, and it's in fact impressive you have been doing it for so long.*

UIMA [Ferrucci and Lally, 2004] Processing pipeline:

- Extracted the verbs immediately before a quotation
  - For the sentence before a quotation we extracted the word marked with the POS tag 'VBD' closer to the quotation
- Processed 334 Kernel Traffic issues
- 11,634 verb occurrences
- 344 unique verbs (and verb-like errors)

Analysis of the extracted verbs.

- From the grand-total of 344 verbs (+ typos/POS-tagger errors)
  - all verbs that appeared at least 100 times (top 55 verbs)
  - expanded them from the full 344-verbs list (+ WordNet synsets)
  - grouped them into classes
- The grouping captures synonyms *for the particular task of introducing quoted text in summaries*.
  - 39 classes
  - contain 127 verbs accounting for 96% of the cases
  - the “other” class with the remaining 217 verbs accounts for 4% of the occurrences
  - Verbs included from WordNet do not appear in the corpus and thus have a count of zero
- using only ‘s/he said’ will miss many possibilities
- **17 different variations with associated likelihoods** for ‘s/he said’.

“dangerous” criteria for generation errors for a verb class:

If the automatic determination of whether the original quote fell into a particular verb class fails, would the original author take issue with the summary upon reading the misclassified verb?

- 10 classes highlighted in the table are too “dangerous”
- none of them account for more than 1% of the total occurrences
- But...
  - a few cases account for most occurrences (“said,” “asked,” and “replied” account for 2/3)
  - the top 9 classes account for 93% of the cases
  - from the rich tail the “dangerous” classes account for 35% of the cases from position 10 and onward
- Mr. Brown’s summaries were enjoyable because of such variability?

Even in the highly technical domain of operating system kernel discussions, Mr. Brown felt the need to use words such as 'groused' and 'chastised.'

- A large number of risky classes:
  - Do we enjoy text that takes a stand? That argues its points in an opinionated manner?
  - Is such the distinction between dull reports and flourish summaries?
  - How can we generate such text?
- Our culture as NLG practitioners, where we always thrive for the perfect output makes such risk taking difficult.

Our data shows that to go beyond 'He Said She Said' in a truly interesting manner we will have to be ready to make mistakes which could make some people unhappy, a trade-off that it would be interesting to see explored more often in NLG.

Our contributions:

- We bring to the attention of NLG practitioners the rich resource embodied in five years of Kernel Traffic newsletters.
- We have highlighted the richness of the problem of lexical choice for verbs introducing quotations in extractive email summarization.
- We have contributed 39 clusters manually assembled from naturally occurring verbs extracted from 4000+ high quality summaries. These clusters can enrich even the most straightforward existing systems.
- We argue that, while useful summaries might be around the corner, entertaining summaries will be well beyond the state of the art until the field is willing to take the risk involved in standing behind automatically generated prose with intrinsic value-judgments.

Acknowledgments

The author would like to thank the anonymous reviewers as well as Annie Ying for valuable feedback and insights.

He will also like to thank the Debian NYC group for bringing the Kernel Traffic summaries to his attention.

References

[Ferrucci and Lally, 2004] Ferrucci, D. and Lally, A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

[Searle, 1975] Searle, J. R. (1975). A taxonomy of illocutionary acts. In *Language, Mind and Knowledge*, pages 344–369. University of Minnesota Press.

[Teufel, 1999] Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text*. PhD thesis, University of Edinburgh, England.