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

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

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.

1 Introduction

High traffic mailing lists pose a challenge to an extended audience laterally interested on the subject matter but unable or unwilling to follow them on everyday minutiae. In this context, high-level summaries are of great help and in certain cases there are people or companies that step into the plate to provide such service. In recent years, there has been an ever increasing interest (Muresan et al., 2001; Nenkova and Bagga, 2003; Newman and Blitzer, 2003; Rambow et al., 2004; Wan and McKeown, 2004; McKeown et al., 2007; Ulrich, 2008; Wang et al., 2009) in automating this task, with many works focusing on selectively extracting quotes from key e-mail exchanges.

In this work, we focus on finding appropriate and varied ways to cite selected quotes from the 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) and citation polarity analysis (Teufel, 1999): (agreed vs. disagreed vs. added).

At this stage, we will show training data we have acquired for the task and a set of manually assembled verb clusters that show the richness of the problem. Moreover, we have used these clusters to highlight a trade-off of "risk taking" in NLG, where generating interesting prose might lead to text that can upset some readers in the presence of errors.

This paper is structured as follows: in the next section we discuss the data from where we obtained the raw verbs and then proceed to describe the manual analysis to cluster and identify "risky" verbs. We then present the whole set of clusters and conclude with a discussion of risk taking in NLG.

2 Data

This work is part of a larger effort to build automatic tools to replace a key resource that the Linux Kernel development community enjoyed for five years: the Kernel Traffic summaries of the activities in the Linux Kernel mailing list (LKML).

The LKML is of extremely high traffic (300 mails a day on average). For five years (since 1999), Jack Brown hand-picked the most newsworthy threads in a week time and published a summary for each thread. The summaries were made available (under a Free Software license) in a rich XML-based format

```
<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>
```

Figure 1: Kernel Traffic #6, Feb. 18th 1999 (excerpt).

(Figure 1) that included, among many other things, explicit marking of all quoted text, with attribution.

These summaries were in general followed by a much larger audience than the mailing list itself due to a number of factors including the fact that they make for quite an entertaining read. Mr. Brown's prose was high quality and quite consistent in style,[1] which highlights its potential as training material for NLG. As Reiter and Sripada (2002) pointed out, learning tasks in NLG profit from training data of the highest possible quality in terms of prose and consistency (as compared with training data for NLU, where robustness comes from exposing the system to a variety of malformed texts).

In our journey to approximate Mr. Brown's work by automatic means, we decided to start on a relatively unstudied problem: introducing quoted references in a rich manner. In the 4,253 hand written summaries by Mr. Brown (made available in 344 newsletter issues) 95% contain a quote, with an average of 3.28 quotes per summary. Moreover, 72% of the total characters in the summaries are inside quotes (including markup).

2.1 Processing

We employed a processing pipeline implemented in the UIMA framework (Ferrucci and Lally, 2004) to extract the verbs immediately before a quotation. We used annotators from the OpenNLP project (Apache, 2011) implementing Maximum Entropy models for NLP (Ratnaparkhi, 1998). For the sentence before a quotation we extracted the word marked with the POS tag 'VBD' closer to the quotation. Processing the 334 issues available for Kernel Traffic resulted in 11,634 verb occurrences extracted for 344 verbs (and verb-like errors). These verbs are the ones we employ for the analysis and inferences drawn in the next section.

3 Analysis

From the grand-total of 344 verbs (including typos and POS-tagger errors), we took all the verbs that appeared at least a hundred times (the top 55 verbs) and expanded them from the larger list (plus WordNet synsets (Miller, 1995)), grouping them into classes. The grouping captures synonyms *for the particular task of introducing quoted text in summaries*. The resulting 39 classes (Table 1) contain 127 verbs accounting for 96% of the cases (the table contains an "other" class with the remaining 217 verbs that account for 4% of the occurrences). The verbs included from WordNet do not appear in the corpus and thus have a count of zero. This large set of verbs highlights the many possibilities a system that chooses to go just with 's/he said' will be missing. Moreover, such a system can be immediately enriched with 17 different variations with associated likelihoods.

We determined whether or not generation errors for a given verb class would be "dangerous" using the following criteria:

> *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?*

That is, if the system decides that Brendan Cully (from the example in the introduction) has indeed kafloogitated[2] with his reply but such decision was made in error (and Mr. Cully was just remarking or explaining), would Mr. Cully take issue with the summary? As with any automated system, the possibility of automated mistakes should make its designers err on the side of making more conservative decisions. Under such desiderata, we think the 10

[1] A quality of prose that continues with his editorial contributions to Linux Journal and Linux Magazine.

[2] That word has been invented by Mr. Brown and was used only once within the five years of Kernel Traffic.

classes highlighted in Table 1 are thus too "dangerous" to be addressed currently by automated means.

Initially, that might not appear such a big loss, as none of them account for more than 1% of the total occurrences. However, as with many other phenomena in NLP, a few cases account for most occurrences: the clusters for "said," "asked," and "replied" account for 2/3 of the total occurrences and, overall, the top 9 classes account for 93% of the cases. From the rich tail that encompasses Mr. Brown prose, the "dangerous" classes account for 35% of the cases from position 10 and onward. It is our opinion that such cases were the reason Mr. Brown's summaries were enjoyable to read and are only a small example of the humor and piquancy behind his prose. Now, it might be the case such quality will be beyond the state of the art of NLG for quite some time.

In that sense, we consider the prevalence of risky classes as a negative result that highlights a problem for NLG well beyond the task at hand: we, as humans, enjoy text that takes a stand, that argues its points in an opinionated manner.[3] Such is the distinction between dull reports and flourish summaries. Even in the highly technical domain of operating system kernel discussions, Mr. Brown felt the need to use words such as 'groused' and 'chastised.'

The problem might as well be cultural, with opinionated prose paradigmatic to the Western world. It might also be related to our culture as NLG practitioners, where we always thrive for perfect output. 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.

4 Related Work

Since the seminal work by Muresan et al. (2001), email summarization and in particular email thread summarization has spanned full dissertations (Ulrich, 2008). Existing resources for email summarization (Ulrich et al., 2008), however, do not emphasize explicitly the type of quotes being used.

Understandingly, most of the work has been devoted to selecting the particular words, sentences or paragraphs to extract from the original e-mails. either by distilling terms or topics (Muresan et al., 2001; Newman and Blitzer, 2003) or finding a representative example (Nenkova and Bagga, 2003; Rambow et al., 2004; Wang et al., 2009).

The issue of choosing how to introduce the extracted text has only been studied in the context of speech act detection (Cohen et al., 2004; Wan and McKeown, 2004) within emails or within threaded discussions (Feng et al., 2006), which is limited to questions, replies and the like (a very important case which covers 2/3 of our available data). The problem of detecting question / answer pairs in e-mails is by far the one who has received the most attention in the field (Bickel and Scheffer, 2004; Shrestha and McKeown, 2004; McKeown et al., 2007).

The verbs in each of the classes in Table 1 have a near-synonym relation:[4] even though "recommended" and "urged" share most of their meaning, the differences in style, color and subtle meaning need to be further elucidated for successful lexical choice. This topic has started to be explored in detail recently (Edmonds and Hirst, 2002).

Our work falls in the larger field of summarization by using NLG means, a discipline that has received significant attention of late (Belz et al., 2009).

5 Conclusion

In this paper, we have brought to the attention of NLG practitioners the rich resource embodied in five years of Kernel Traffic newsletters. We had also highlighted the richness of the problem of lexical choice for verbs introducing quotations in extractive email summarization.

Moreover, we 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. Finally, we argued 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.

In our ongoing work, we are targeting the creation

[3] Not unlike this discussion.

[4] Thanks to an anonymous reviewer for bringing this fact into our attention.

Table 1: Quotation introducing verb classes, with counts. The "other" class appears in row 7. Lines in bold are considered "dangerous." The last column is the author's opinion about which type of technology is more relevant for choosing that class (speech act detection (A), opinion mining (O) or citation link analysis (C)). Verbs missing due to space restrictions are in the appendix.

#	Top Verbs	# verbs	Total Counts		Accum.	Type
1	said (2726) remarked (361) posted (163) pointed out (148)	17	3531	(30.35%)	30.35%	A
2	replied (3476) responded (21) answered (11)	3	3508	(30.15%)	60.50%	A
3	added (1059) included (13) followed (10)	3	1082	(9.30%)	69.80%	C
4	announced (902) declared (1)	2	903	(7.76%)	77.56%	A
5	asked (509) inquired (0)	2	509	(4.37%)	81.94%	A
6	explained (427)	1	427	(3.67%)	85.61%	A
7	FELT (21) MADE (21) WANTED (8) BROKE (8)	217	403	(3.46%)	89.07%	-
8	reported (254) detailed (1)	2	255	(2.19%)	91.26%	A
9	suggested (188) proposed (35)	2	223	(1.91%)	93.18%	O
10	**objected (90) protested (5)**	2	95	(0.81%)	94.00%	O
11	concluded (48) ended (5) finished (4) closed (2)	5	59	(0.50%)	94.50%	C
12	offered (52) volunteered (6)	2	58	(0.49%)	95.00%	O
13	confirmed (44) supported (4) affirmed (3) reasserted (1)	7	52	(0.44%)	95.45%	C
14	summed up (21) summarized (18)	2	39	(0.33%)	95.78%	A
15	agreed (37) concurred (1) concorded (0)	3	38	(0.32%)	96.11%	C
16	described (33)	1	33	(0.28%)	96.39%	A
17	**took issue (17) disagreed (11) dissented (2) differed (1)**	4	31	(0.26%)	96.66%	O
18	**complained (22) sounded off (2) kicked (1) groused (1)**	7	29	(0.24%)	96.91%	O
19	**argued (28) contended (0) debated (0)**	3	28	(0.24%)	97.15%	O
20	listed (27) enumerated (0)	2	27	(0.23%)	97.38%	A
21	continued (25) kept (1)	2	26	(0.22%)	97.61%	A
22	clarified (25) elucidated (0)	2	25	(0.21%)	97.82%	C
23	recommended (17) urged (4) advised (2) advocated (1)	4	24	(0.20%)	98.03%	C
24	**speculated (16) mused (2) guessed (2) supposed (1)**	6	22	(0.18%)	98.22%	O
25	elaborated (11) expanded (7) expounded (2)	3	20	(0.17%)	98.39%	C
26	**corrected (18) chastised (1) rectified (0) righted (0)**	4	19	(0.16%)	98.55%	O
27	**exclaimed (6) called out (5) cried out (4) shouted (2)**	5	18	(0.15%)	98.71%	O
28	quoted (15) cited (2)	2	17	(0.14%)	98.85%	C
29	**warned (8) cautioned (6) admonished (2)**	3	16	(0.13%)	98.99%	O
30	interjected (11) sprung (1) interposed (1)	3	13	(0.11%)	99.10%	O
31	**quipped (10) joked (1) chuckled (1) cracked (1)**	4	13	(0.11%)	99.21%	O
32	requested (12)	1	12	(0.10%)	99.32%	A
33	tried (9) attempted (2) tested (1)	3	12	(0.10%)	99.42%	O
34	acknowledged (8) admitted (3) recognized (0)	3	11	(0.09%)	99.51%	A
35	countered (10)	1	10	(0.08%)	99.60%	C
36	found (7) discovered (2) launched (1)	3	10	(0.08%)	99.69%	A
37	reiterated (9) repeated (1)	2	10	(0.08%)	99.77%	C
38	started (9) began (1)	2	10	(0.08%)	99.86%	A
39	**rejoined (6) retorted (2) returned (1)**	3	9	(0.07%)	99.93%	O
40	chimed (7)	1	7	(0.06%)	100%	O

of a systemic fragment for the quotation-introducing verbs, in the style of KPML (Bateman, 1995).

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.

Appendix

The verbs omitted for reasons of space in Table 1 are the following: for the "said" cluster, mentioned (34), commented (25), wrote (20), noticed (17), spoke (9), expressed (6), showed (5), observed (5), stated (5), asserted (4), referred (1), noted (1), declared (1); for the "concluded" cluster, resolved (0); for the "confirmed" cluster, corroborated (0), sustained (0), substantiated (0); for the "complained" cluster, hollered (1), ranted (1), kvetched (1); for the "speculated" cluster, theorized (1), conjectured (0); for the "exclaimed" cluster, sputtered (1).

References

Apache. 2011. OpenNLP $http://incubator.apache.org/opennlp$.

John A. Bateman. 1995. KPML: The KOMET–Penman multilingual linguistic resource development environment. In *Proc. of EWNLG*, pages 219–222.

Anja Belz, Roger Evans, and Sebastian Varges, editors. 2009. *Proc. of the 2009 Workshop on Language Generation and Summarisation (UCNLG+Sum 2009)*. ACL, Suntec, Singapore, August.

Steffen Bickel and Tobias Scheffer. 2004. Learning from message pairs for automatic email answering. In *ECML*, volume 3201 of *LNCS*, pages 87–98. Springer.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proc. of EMNLP*, volume 4.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. Learning to detect conversation focus of threaded discussions. In *Proc. HLT-NAACL*, pages 208–215.

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

Kathleen McKeown, Lokesh Shrestha, and Owen Rambow. 2007. Using question-answer pairs in extractive summarization of email conversations. In *CICLing*, volume 4394 of *LNCS*, pages 542–550. Springer.

G.A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Smaranda Muresan, Evelyn Tzoukermann, and Judith L. Klavans. 2001. Combining linguistic and machine learning techniques for email summarization. In *Proc. of the 2001 workshop on Computational Natural Language Learning-Volume 7*, page 19. ACL.

Ani Nenkova and Amit Bagga. 2003. Facilitating email thread access by extractive summary generation. In *RANLP*, volume 260 of *Current Issues in Linguistic Theory (CILT)*, pages 287–296.

Paula S. Newman and John C. Blitzer. 2003. Summarizing archived discussions: a beginning. In *IUI*, pages 273–276. ACM.

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

Owen Rambow, L. Shrestha, J. Chen, and C. Lauridsen. 2004. Summarizing email threads. In *Proc. of HLT-NAACL 2004: Short Papers*, pages 105–108. ACL.

Adwait Ratnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Ehud Reiter and S. Sripada. 2002. Should corpora texts be gold standards for NLG? In *Proceedings of Second International Conference on Natural Language Generation INLG-2002*, pages 97–104, Arden House, NY.

John R. Searle. 1975. A taxonomy of illocutionary acts. In *Language, Mind and Knowledge*, pages 344–369. University of Minnesota Press.

Lokesh Shrestha and Kathleen McKeown. 2004. Detection of question-answer pairs in email conversations. In *Proc. of ACL*, page 889. ACL.

Simone Teufel. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh, England.

Jan Ulrich, Gabriel Murray, and Giuseppe Carenini. 2008. A publicly available annotated corpus for supervised email summarization. In *AAAI08 EMAIL Workshop*, Chicago, USA. AAAI.

Jan Ulrich. 2008. Supervised machine learning for email thread summarization. Master's thesis, Computer Science.

Stephen Wan and Kathleen McKeown. 2004. Generating overview summaries of ongoing email thread discussions. In *Proc. of ACL*, page 549. ACL.

Baoxun Wang, Bingquan Liu, Chengjie Sun, Xiaolong Wang, and Bo Li. 2009. Adaptive maximum marginal relevance based multi-email summarization. In *AICI*, volume 5855 of *LNCS*, pages 417–424. Springer.