Getting Emotional About News Summarization

Alistair Kennedy[1], Anna Kazantseva[1], Diana Inkpen[1], Stan Szpakowicz[1,2]

[1] School of Electrical Engineering and Computer Science
University of Ottawa, Ottawa, Ontario, Canada
{akennedy,ankazant,diana,szpak}@eecs.uottawa.ca
[2] Institute of Computer Science
Polish Academy of Sciences, Warsaw, Poland

Abstract. News is not simply a straight re-telling of events, but rather an interpretation of those events by a reporter, whose feelings and opinions can often become part of the story itself. Research on automatic summarization of news articles has thus far focused on facts rather than emotions, but perhaps emotions can be significant in news stories too. This article describes research done at the University of Ottawa to create an emotion-aware summarization system, which participated in the Text Analysis Conference last year. We have established that increasing the number of emotional words could help ranking sentences to be selected for the summary, but there was no overall improvement in the final system. Although this experiment did not improve news summarization as evaluated by a variety of standard scoring techniques, it was successful at generating summaries with more emotional words while maintaining the overall quality of the summary.

1 Introduction

1.1 Text Analysis Conference

Research on text summarization goes back to the early days of Artificial Intelligence [1]. Summarization has been applied to a variety of domains, but one of the most popular domains has been news documents. In recent years there has been a trend towards summarizing opinions in blogs [2–4]. We discuss experiments to move this line of research towards summarizing news articles using emotion.

The Document Understanding Conference (DUC) and its successor, the Text Analysis Conference (TAC), are annual shared evaluation exercises. Researchers get a chance to see who has created the best-evaluated query-driven multi-document news summarization system. In the recent years, summaries of 100 words have been targeted: given a query and a set of news articles, construct a summary which addresses the points raised by the query. In 2010 and 2011, TAC changed the task to what is called guided summarization. The objective is to create summaries of news articles which fall into one of five news categories: “Accidents and Natural Disasters”, “Attacks”, “Health and Safety”, “Endangered Resources” and “Investigations and Trials”. Within each category, there are a number “aspects”, questions which the system should discuss and answer. For example, the aspects for “Accidents and Natural Disasters” are:

- WHAT: what happened

- WHEN: date, time, other temporal placement markers
- WHERE: physical location
- WHY: reasons for accident/disaster
- WHO AFFECTED: casualties (death, injury), or individuals otherwise negatively affected by the accident/disaster
- DAMAGES: damages caused by the accident/disaster
- COUNTERMEASURES: countermeasures, rescue efforts, prevention efforts, other reactions to the accident/disaster

Five different sets of questions, mostly pertaining to the "who", "what", "where", "when", "why," and "how", are standardized for each news category. Each query also contains a topic such as "Giant Panda", "Columbine Massacre" or "South Korean Wire Tapping".[3] The 2010 data were used for tuning, the 2011 data – for testing.

As in previous years at TAC, participating systems generated one normal summary and one update summary for each query. The original and update summaries were generated from two 10-document sets, referred to as document sets A and B respectively. The intent of the update summary was to select from document set B only new information, not present in document set A. In this paper we focus on our experiments meant to improve the creation of summaries for set A.

Four kinds of evaluation are performed at TAC. One is to measure the readability of the summaries by manual annotation. The second is Pyramid Evaluation [5], a manual procedure which follows a strict sequence. Summary Content Units (SCUs) are extracted from human-written *model* summaries. A SCU is a factoid weighted by the number of model summaries in which it appears. An annotator then goes through all the machine-generated *peer* summaries and marks them with SCUs. The summaries are given a score based on their recall of SCUs.

The third kind of evaluation is for overall responsiveness. This too is a manually assigned score, meant to be a balance of readability and content. The last form of evaluation uses an automatic method called ROUGE [6]. ROUGE finds N-grams in model summaries and counts how many of them were found in each peer summary. This is a heuristic method of approximating responsiveness. Since responsiveness is calculated by hand for each of our summaries, ROUGE is presented for the sake of comparison.

1.2 Our Motivation

Recently much research in Natural Language Processing has been devoted to emotions. Given the recent successes of summarizing for opinions, it was natural to assume that the next step would be to summarize for emotion. Our hypothesis has been that certain emotions will be more strongly associated with summaries for each of the five categories in TAC 2011. By identifying these emotions in a news article we wanted to select better sentences for our extractive text summarization system. We proposed to identify emotional categories which are more common to the model summaries of the five news categories than they are across the document sets which they summarize.

[3] See http://www.nist.gov/tac/2011/Summarization/Guided-Summ.2011.guidelines.html for a detailed description.

Emotional words could thus help identify sentences more likely to be useful in a summary. Emotional summaries may also be desirable to readers if they wish to know more about how the author or the subjects of the news article felt about the event described. To our knowledge this has been the first attempt to employ emotions in summarization.

Three ways of improve our summarization system presented themselves. First, people might enjoy reading summaries with more emotion and so find them to be more readable. Second, if our hypothesis was right and summaries of one news category tended to contain more emotional words, then selecting sentences with emotional words could improve the summarization system on Pyramid Evaluation as well. Third, all this could increase overall responsiveness and potentially the ROUGE score as well.

This paper has four more sections. Section 2 describes the word-emotion association lexicon and how it was used to identify important emotions for news articles. Section 3 describes the baseline and emotion-aware summarization systems. A description of the TAC evaluation and our conclusions can be found in Sections 4 and 5 respectively.

2 Identifying Emotion

Human cognition is capable of many nuanced emotions, but it has been argued that joy, sadness, anger, fear, trust, disgust, surprise and anticipation are the most prototypical [7]. We worked with the NRC Emotion Lexicon v0.5 created by the National Research Council of Canada (NRC) [8] to count both emotional and sentimental words.

The words in the lexicon are marked for associations with the eight prototypical emotions, and with sentiment of positive or negative polarity. Many words not labelled with any emotion or sentiment. Here are the counts of words from the emotional and sentimental classes in this data set (many words were labeled with multiple emotions):

- Emotion: 2283
  - Joy: 353 • Sadness: 600 • Fear: 749 • Surprise: 275 • Disgust: 540 • Anger: 647 • Trust: 641 • Anticipation: 439 • No emotion: 4808
- Sentiment: 2821
  - Positive: 1183 • Negative: 1675 • No sentiment: 4270

2.1 Emotions by Category

Our goal was to find emotions most useful when making summaries for each news category. We assumed that emotional words that appear more frequently in the human-written model summaries than in the document sets for summarizing should also be more numerous in an automatically generated summary. To that end, we determined which of the N emotions appear more than expected in the summaries of a given news category. We calculated the emotion density ED by normalizing the count of emotion E_i by the count of all emotional words $E_{1..N}$ and non-emotional words $\neg E$.

$$ED(E_i) = \frac{count(E_i)}{count(E_{1..N}) + count(\neg E)} \quad (1)$$

	2010	2011
Accidents	7	9
Attacks	7	9
Health	12	10
Resources	10	8
Trial	10	8
Total	46	44

Table 1. Count of all five news categories in the 2010 and 2011 TAC data sets.

	Emotion									Sentiment		
	Joy	Sad	Fear	Surprise	Disgust	Anger	Trust	Anticip	None	Pos	Neg	None
Accidents	1.070	**1.349**	1.079	1.036	0.998	1.254	0.842	0.966	**0.917**	1.039	**1.195**	**0.924**
Attacks	0.801	**1.220**	**1.242**	0.996	1.201	**1.378**	**0.593**	**0.590**	0.908	0.908	**1.323**	**0.885**
Health	1.127	1.171	1.163	0.973	1.158	1.271	0.790	**0.726**	0.971	0.932	**1.271**	0.951
Resources	1.202	0.906	1.120	**0.622**	1.197	1.070	1.073	1.021	0.968	**1.305**	1.123	**0.901**
Trial	0.797	**1.561**	**1.157**	**1.372**	**1.453**	**1.458**	0.818	**0.841**	**0.686**	0.999	**1.522**	**0.807**

Table 2. The ratio of emotion/sentiment densities across the model summaries and the source documents on TAC 2010 data. Boldface signals statistical significance at $p < 0.05$.

We calculated the emotion densities of the model summaries $ED_M(E_i)$ and of the document set $ED_D(E_i)$. We then calculated the emotion ratio:

$$\frac{ED_M(E_i)}{ED_D(E_i)} \tag{2}$$

This determined which emotions are more frequent in the model summaries than the document sets. The same experiments were run for sentiment and emotion. Student's t-test measured statistical significance, at $p < 0.05$, for each emotion ratio. Table 2 shows the results. This evaluation used both the A and B datasets from TAC 2010. Four model summaries were generated for sets A and B, giving 8 model summaries per topic. Each query had 10 news articles, each set A and B giving 20 articles per topic. Table 1 shows the number of topics for each news category.

The results in Table 2 show that for all news categories, from the TAC 2010 data, there were more emotional and sentiment words in the summaries than in the document set, often significantly so. The human summarizers appear to favour emotional content when generating summaries. It is difficult to explain precisely why this is so; perhaps some aspects of the queries are more naturally answered with emotion. We hoped to find emotions with strong positive connections to each news category, but in cases where no emotion had a strong positive connection we considered strongly positive sentiment. The findings were that the following news categories were most likely to contain the following emotions and sentiments:

- Accidents: Sadness
- Attacks: Sadness, Fear & Anger

- Health: None, but strongly Negative
- Resources: None, but strongly Positive
- Trials: Sadness, Fear, Surprise, Disgust & Anger

In another experiment we did not take the news categories into account to determine whether the emotion densities across all summaries were higher than the emotion densities in the source documents. The findings were that the summaries had a significantly higher number of words associated with *sadness*, *fear*, *disgust*, *anger* and with negative sentiment. There was a significant negative correlation with *trust*, *anticipation*, non-emotional words and non-sentimental words. *Joy* and *surprise* words were not strongly associated either way. That was not a surprise because we suspected news to be more strongly associated with negative events and so negative emotions.

3 The System

With these findings in mind, we began putting together a summarization system. It had two main modules. A clustering system [9] grouped sentences by topics discovered within the documents. The purpose of the clustering module was to establish main themes of each document set independent of the query. It would also be useful in minimizing redundancy in the summaries. The second module was a sentence ranker [10] which selected from each cluster the sentence closest to the query.

Two variations were attempted. A baseline system used only the queries. The second system was the emotionally aware summarization system. For each news category we attempted to boost the emotional/sentimental words which had the strongest association with that category (as seen in Table 2). These emotions were used for query expansion for each news category. This would create summaries which are highly emotional.

3.1 Module 1: Clustering

The queries for each document set may be used to guide the summarization process so as to best answer the information need described in the query. On the other hand, each set of documents is rather self-sufficient in that it is possible to produce an informative summary even without the query. From reading the documents alone one may infer the important subtopics and include only the most relevant ones in the summary. To utilize this information, we cluster sentences of each document set into topical clusters.

The clustering algorithm is Affinity Propagation [9], a loopy belief propagation algorithm for exemplar-based clustering. It takes two inputs: a matrix of pair-wise similarities between data points (here the data points are sentences); and a vector of preference values corresponding to *a priori* beliefs of how likely each data point is to be a cluster centre. The algorithm chooses a set of cluster centres – exemplars – and assigns all data points to the best-fitting exemplar in a way which maximizes net similarity. That is to say, the total sum of similarities between all data points and their respective cluster centres (the same objective function as in the well-known k-means algorithm).

In order to perform clustering we pre-processed the sentences. We chose to represent each sentence as a bag of words, with stop words removed. Each sentence was

represented as a vector of type-token frequencies, weighted by the *tf.idf* metric. The similarity between sentences was computed by the usual cosine similarity metric:

$$cos(s_1, s_2) = \frac{s_1 \bullet s_2}{||s_1|| \times ||s_2||} \tag{3}$$

The result was a pairwise similarity metric between all sentences in each document set.

One of the parameters for Affinity Propagation is a vector of preference values (one for each data point) which reflects how likely each data point is to be an exemplar, given prior knowledge. Lead sentences in a newswire article summarize the entire document quite well. To reflect this, we decided to adjust the preference values to increase the likelihood of choosing those sentences as cluster exemplars.

Usually, for each document set the clusterer identified at least one "stray" cluster – a cluster with sentences which have little similarity with any other sentence in the document set. We identified such clusters by their low net similarity value and discarded them. The topical clusterer then returned at most 50 central sentences for each good cluster, along with their scores.

The clustering module was fine-tuned using the TAC 2010 dataset. We found the parameter settings which maximize the value of the objective function for clustering (net similarity) and then used those settings to run on the TAC 2011 test data.

3.2 Module 2: Sentence Ranking

We chose the same sentence ranker as the one we employed in the last two years [11, 12] and further discussed in [10]. The ranker uses the 1911 edition of *Roget's Thesaurus* [13] to measure the distance between the query and a sentence in the document.[4]

To evaluate the sentence ranker we used a corpus labeled with SCUs [5]. Sentences from summaries from previous years were mapped back to the original corpus and then sentences in the corpus could be labeled as containing a SCU, containing no SCUs or of unknown SCU status. Only sentences known to contain SCUs or known to contain no SCUs were used to evaluate a sentence ranker. The actual evaluation took the mean average precision score of the known positive and negative sentences in the SCU labeled corpus. We use the macro average of the mean average precision, calculated for both the A and B sets on the 2010 TAC data in order to determine the best parameters for the system. This process is further described in [10].

The *Roget's*-based sentence ranker works as follows. For each word q in the query Q, the most closely semantically related word w in a sentence S is found, giving a score from 0 to 18 (0 means no relation, 18 – a perfect match). Closely related words or near-synonyms were given scores of 16 or 14. This scoring function is known as $semDist$ [14]. The word pair scores were then summed up to give a sentence score:

$$score(S) = \sum_{q \in Q} max(SemDist(w, q) : w \in S) \tag{4}$$

The sentences are then ranked by sentence score.

[4] A Java implementation of *Roget's Thesaurus* can be found at: http://rogets.eecs.uottawa.ca/

Method	Set A	Set B
Random	*0.430*	*0.352*
Longest Sentence	*0.541*	***0.465***
Topic	**0.580**	0.433
Topic & Aspects	0.549	0.435

Table 3. Mean average precision for the baseline sentence ranker on the TAC 2010 data.

We performed two experiments to identify how to format the query for the baseline system. One version is to use the query topic and the aspect, the other is to use the query topic alone. In theory, the aspects should add useful information, but it may well be the case that all they do is introduce noise. Two other baselines were also tested: ranking sentences randomly and ranking sentences by length. The longer the sentences is, the more likely it is to contain a SCU. The experiments to establish the baseline system are reported in Table 3.

Evaluation was performed on both the A and B data sets from 2010, though our interest was in the evaluation on set A – our work did not directly apply itself to update summaries. We found that including only the topic statement for each query gave the best results for set A, while for set B the longest sentence baseline was superior. For set B the difference between using the topic alone or with aspects was very small. Set A was the data set we were most interested in, so we decided to use only the topic as the query. Despite the longest sentence baseline working well, we noted that the longest sentence often had close to 100 words, so selecting it might create a one-sentence summary. By comparison, the summaries generated using the *Roget's semDist* sentence ranker generally had around 4 sentences each.

The next question was how to incorporate the emotional/sentimental words into the sentence ranker. We found that simply adding these new words to the query was prohibitive in terms of run time, because it would require *Roget's* to measure the distance between millions of word pairs. We also believed that grouping words by closeness of semantics did not guarantee closeness of emotion. *Happy* and *sad* are closely related semantically, but not emotionally, so only exact matches were used. We decided only to match emotional/sentimental words exactly, but what weight should be applied to these words? Several different weighting variations were tested.

We tried giving each emotion/sentiment word a weight of 1, 2, 4 and a *Ratio-Weight* corresponding to the score for each emotion/sentiment from Table 2. *Ratio-Weight* gives a different weight depending on how strongly associated each emotion/sentiment is with the news category. Table 4 shows the results. A clear winner is the *Ratio-Weight* method though in general one can see that scores in the range of 1 or 2 gave strong results.

We also wanted to see how the *baseline* and *emotional* sentence rankers would perform on the five news categories. To do this we calculated the mean average precision on each of the news categories for the A and B data sets. The mean average precision scores and p-values are shown in Table 5.

Although the results are only for the tuning data set from TAC 2010, it seems that the emotion-aware summarization system can often significantly outperform the base-

Method	Set A	Set B
Weight 1	0.611	0.460
Weight 2	0.612	**0.462**
Weight 4	0.610	0.457
Ratio-Weight	**0.616**	**0.462**

Table 4. Mean average precision for the emotional sentence ranker on the TAC 2010 data.

Set	Category	Baseline	Emotion	p-value
A	All	0.580	0.616	0.002
	Accidents	0.661	0.691	0.062
	Attacks	0.515	0.575	0.096
	Health	0.478	0.542	0.074
	Resources	0.584	0.594	0.486
	Trial	0.687	0.701	0.425
B	All	0.433	0.462	0.007
	Accidents	0.545	0.528	0.088
	Attacks	0.522	0.528	0.693
	Health	0.366	0.410	0.115
	Resources	0.373	0.376	0.885
	Trial	0.431	0.480	0.106
A&B	All	0.506	0.539	0.000
	Accidents	0.603	0.637	0.008
	Attacks	0.519	0.552	0.087
	Health	0.422	0.476	0.014
	Resources	0.479	0.485	0.562
	Trial	0.559	0.591	0.065

Table 5. Mean average precision for the different news categories on the TAC 2010 data.

line. Resources had the smallest improvement, though we noted that the only emotion/sentiment with which Resources correlated was *positive* words. This is a very broad class of words and does not intuitively make much sense. We suspected that this was an anomaly, but we decided not to let our suspicions influence the experiment. We were optimistic, because this evaluation showed that adding emotional words would improve our sentence ranking system. In theory, this could lead to a higher score in the Pyramid evaluation, and hopefully in responsiveness too.

3.3 The Final Systems

In the final systems we used the sentence clustering algorithm to assign every sentence a cluster ID. The sentence ranker then ranked all sentences in the document set. Sentences closest to the query were then added to the summary under the condition that it did not exceed 100 words and that the summary never contained two sentences with the same cluster ID.

Baseline Summary:
The quake, with a magnitude of 7.8, struck close to densely populated areas in Sichuan province, including the capital Chengdu, shortly before 2:30 pm (0630 GMT) on Monday. Chinese authorities did not detect any warning signs ahead of Monday's earthquake that killed more than 8,600 people, state media reported. The State Ethnic Affairs Commission decided on Tuesday to grant 2 million yuan (about 285,000 U.S. dollars) to its provincial branch in the southwestern Sichuan Province for **disaster**-relief work.

Emotional Summary:
China has allocated 200 million yuan (29 million dollars) for **disaster** relief work after an earthquake rocked the country's southwest **killing** more than 8,700 people, state press reported Tuesday. The **disaster** areas of Sichuan will see moderate to heavy rainfall in the next two days, tailing off Wednesday, said a statement released by the World Meteorological Organization here. The ASEAN Inter-Parliamentary Assembly (AIPA) on Wednesday expressed its condolence and **sympathy** to China following the **devastating** earthquake in Sichuan province.

Fig. 1. Examples of a baseline and emotional summary for document set "D1110A: Earthquake Sichuan". This news category is "Accidents and Natural Disasters", strongly associated with *sadness*. Words related to *sadness* are in bold.

The systems, based on TAC 2011 data, produced two versions of every summary. One was the baseline summary in which the query was just the topic statement. The other was an emotional summary in which emotional/sentimental words were used for query expansion. An example of a baseline and emotional summary can be seen in Figure 1. These summaries are for news articles on the topic of "Earthquake Sichuan" in the news category "Accidents and Natural Disasters". This category was most closely related to the emotion *sadness*.

We decided to examine the number of emotional words in the baseline and emotional summary systems in other to confirm that their query expansion was working. Table 6 shows the proportion of emotional words, by news category, found in the emotional summaries, over that of the baselines summaries. This is calculated as follows:

$$\frac{emotionCount(emotionalSummaries)}{emotionCount(baselineSummaries)} \tag{5}$$

The emotions/sentiment used for query expansion are in bold. Not surprisingly, those emotions/sentiment tend to be more frequent than the other emotions/sentiment. In a few cases, emotions not used for query expansion were also boosted, for example "disgust" in the accident, attack and health categories. Words may be marked with multiple emotions, so it is likely that words related to "disgust" are also found in other emotional categories. Nonetheless the summaries produced using our emotion-aware summarization system clearly enhanced emotions in these news categories. The experiment has thus been successful, but there is still the matter of evaluation.

	Emotion									Sentiment		
	Joy	Sad	Fear	Surprise	Disgust	Anger	Trust	Anticip	None	Pos	Neg	None
Accidents	1.000	**3.847**	2.167	2.364	3.125	2.200	1.278	0.905	0.953	1.143	2.267	0.923
Attacks	1.667	**1.900**	**2.182**	1.125	2.500	**1.921**	1.190	1.417	0.888	1.286	1.878	0.932
Health	0.913	1.920	2.038	1.000	2.154	2.059	0.895	1.047	1.072	0.949	**2.244**	0.950
Resources	2.833	0.923	0.857	1.400	1.200	0.923	2.136	2.500	1.094	**2.310**	1.077	1.012
Trial	1.00	**2.296**	**1.596**	**2.727**	**2.368**	**1.837**	0.581	1.500	0.911	1.00	1.816	0.931

Table 6. Emotional count in emotional summaries normalized by count in baseline summaries on TAC 2011 data.

Set	A & B		A		B	
Score	Baseline	Emotion	Baseline	Emotion	Baseline	Emotion
Responsiveness	2.273	2.341	2.523	2.500	2.023	2.182
Readability	3.057	3.091	3.136	3.091	2.977	3.091
Pyramid	0.283	0.277	0.329	0.326	0.237	0.227
ROUGE-SU4	-	-	0.122	0.114	0.097	0.100
ROUGE-2	-	-	0.083	0.073	0.055	0.059

Table 7. Evaluation scores for Responsiveness, Readability and Pyramid Evaluation on the TAC 2011 data.

4 TAC Evaluation

Although the experiments showed promise on the 2010 data, there was less success on the 2011 data. The results in Table 7 show that the addition of emotional information did not noticeably improve responsiveness, readability or the Pyramid Evaluation. Also present are the ROUGE-SU4 and ROUGE-2 scores for the A and B document sets. ROUGE-2 measures bigram overlap, ROUGE-SU4 – skip bigrams with up to 4 spaces between words. Responsiveness and readability are measured out of 5, Pyramid Evaluation and ROUGE out of 1. We examined each news category individually; no single news category was significantly affected by the addition of emotional words.

These results show that the difference in terms of responsiveness, readability and Pyramid scores is almost indistinguishable between the baseline and emotion-aware systems. Overall, a small increase in responsiveness and readability was measured but this is too small to be considered statistically significant. Despite this, the emotion-aware summarization system did produce summaries with considerably more emotional words than the baseline system.

ROUGE scores are only provided for the A and B data sets. These scores are also quite close, though for set A the evaluation measures a small drop, while for set B there is a small increase for the emotional summarizer. As ROUGE is meant to estimate responsiveness, which is measured manually, these scores are of much less importance in this evaluation.

News Category	Emotions – 2011	Emotions – 2012
Accidents	Sadness	None
Attacks	Sadness, Fear & Anger	Fear & Anger
Health	None, but strongly Negative	None, but strongly Negative
Resources	None, but strongly Positive	None, but strongly Negative
Trials	Sadness, Fear, Surprise, Disgust & Anger	Sadness, Fear & Anger

Table 8. Comparison of the significantly associated emotions/sentiment for each news category between the 2011 and 2012 data.

5 Discussion and Conclusion

The summaries created by our emotion-aware summarizer contained many more emotional words than the baseline system. Based on the experiments with the TAC 2010 data set we were optimistic that the emotional summaries would yield an improvement in Pyramid evaluation, readability and responsiveness. Our hope did not materialize, but even so there are some interesting lessons in this experiment. The increase in emotion words did little to help but it also did not hurt our system at all. This could be a starting point for future research on emotion in summarization.

There may be many reasons why a significant improvement was not found. Table 5 shows a significant improvement in the sentence ranker when using emotional words, but this evaluation was conducted over the entire document set and not just the sentences selected for summarization. It may be possible to improve overall ranking a lot, yet not have a measurable difference when a small summary is generated. Perhaps any improvement cannot be measured on summaries of just 100 words. The intermediate results – experiments ranking sentences – did benefit from using emotional query expansion. While this alone does not guarantee improved summaries, it does suggest that using emotional words in the sentence ranking process may improve future summarization systems.

Errors may also arise if the emotions most strongly associated with each news category have changed between the 2010 and 2011 data. To verify this we repeated the experiments from Section 2 using the TAC 2011 documents and model summaries. A comparison of the significantly associated emotions appears in Table 8. The emotions for Resources and Accidents were completely different from 2010 to 2011, but for Attacks, Health and Trials the associated emotions/sentiments are quite similar. Even so, we did not find any improvement on these three news categories individually.

Finally, the evaluation performed at TAC has no measure specific to the emotional content of the summary. Emotional summaries may be desirable if one wishes to identify how the authors or subjects of news articles actually felt about events, but this could not be identified using the current evaluation techniques.

We see this research as a starting point towards building emotional summaries where a user may direct the system to create a summary which captures expressions of anger, joy, anticipation, or some other emotion. This project is also a first step towards moving beyond summarizing opinion and moving into the domain of summarizing emotion. This line of research may bring the highest benefit to summarization in domains other

than news, particularly product reviews, conversations or stories where emotion may play a stronger role. It may also be preferable to create summaries which contain absolutely no emotion. Although this is the opposite of what we present here it would still require the summarizer to be aware of emotions in text.

Acknowledgments

Partial support comes from the Natural Sciences and Engineering Research Council of Canada. Thanks to Saif Mohammad for his help with the NRC Emotion Lexicon v0.5 and to Terry Copeck for generating the SCU-labeled corpus year after year.

References

1. Luhn, H.P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development **2** (1958) 159–165
2. Mithun, S., Kosseim, L.: Summarizing Blog Entries Versus News Texts. In: Proceedings of the Workshop on Events in Emerging Text Types. eETTs '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 1–8
3. Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., Martínez-Barco, P.: Summarizing Threads in Blogs Using Opinion Polarity. In: Proceedings of the Workshop on Events in Emerging Text Types. eETTs '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 23–31
4. Feng, S., Wang, D., Yu, G., Li, B., Wong, K.F.: Summarizing and Extracting Online Public Opinion from Blog Search Results. In: Database Systems for Advanced Applications (DASFAA), 15th International Conference. (2010) 476–490
5. Nenkova, A., Passonneau, R.J.: Evaluating Content Selection in Summarization: The Pyramid Method. In: Proceedings of HLT-NAACL, Human Language Technology conference / North American chapter of the ACL annual meeting. (2004) 145–152
6. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of summaries. In: Proceedings of the ACL workshop: Text Summarization Branches Out. (2004) 74–81
7. Plutchik, R.: A General Psychoevolutionary Theory of Emotion. Emotion: Theory, Research, and Experience **1**(3) (1980) 3–33
8. Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon. To Appear in Computational Intelligence (2012)
9. Givoni, I.E., Frey, B.J.: A Binary Variable Model for Affinity Propagation. Neural Computation **21** (2009) 1589–1600
10. Kennedy, A., Szpakowicz, S.: Evaluation of a Sentence Ranker for Text Summarization Based on Roget's Thesaurus. In: Text, Speech and Dialogue (TSD), 13th International Conference, Brno, Czech Republic (2010) 101–108
11. Copeck, T., Kennedy, A., Scaiano, M., Inkpen, D., Szpakowicz, S.: Summarizing with Roget's and with FrameNet. In: Second Text Analysis Conference (TAC). (2009)
12. Kennedy, A., Copeck, T., Inkpen, D., Szpakowicz, S.: Entropy-Based Sentence Selection with Roget's Thesaurus. In: Third Text Analysis Conference (TAC). (2010)
13. Kennedy, A., Szpakowicz, S.: Evaluating Roget's Thesauri. In: ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, The Association for Computer Linguistics (2008) 416–424
14. Jarmasz, M., Szpakowicz, S.: Roget's Thesaurus and Semantic Similarity. In: Recent Advances in Natural Language Processing III. Selected papers from RANLP-03. CILT vol. 260. John Benjamins, Amsterdam (2004) 111–120