Getting Emotional About News

Alistair Kennedy[1], **Anna Kazantseva**[1], **Saif Mohammad**[2], **Terry Copeck**[1], **Diana Inkpen**[1], **Stan Szpakowicz**[1,3]
[1]School of Electrical Engineering and Computer Science
University of Ottawa
{akennedy, anna, terry, diana, szpak}@eecs.uottawa.ca
[2]Institute for Information Technology
National Research Council Canada
saif.mohammad@nrc-cnrc.gc.ca
[3]Institute of Computer Science
Polish Academy of Sciences

Abstract

News is not simply a straight re-telling of events, but rather an interpretation of those events by a reporter, where the feelings and opinions of that reporter can often become part of the story itself. Research on automatic summarization of news articles has historically focused on facts and not emotions, but perhaps emotions can be significant in these stories too. In this article you will read about the work of several researchers, primarily from the University of Ottawa, in their attempt to identify emotions common to different kinds of news articles and incorporate this into the summarization process.

This article also describes the University of Ottawa's contribution to the Automatically Evaluating Summaries of Peers (AESOP) challenge.

1 Introduction

The Document Understanding Conference (DUC) and its successor the Text Analysis Conference (TAC) are annual competitions where researchers from around the world compete to see who has created the best query-driven multi-document news summarization system. In this task one is given a query and a set of news articles from which to construct a summary, answering whatever questions appear in the query. In 2010 and 2011 the Text Analysis Conference (TAC) altered their task of document summarization to what they called guided summarization. In guided summarization the objective is to create summaries of news articles that fall into one of five broad categories: "Accidents and Natural Disasters", "Attacks", "Health and Safety", "Endangered Resources" and "Investigations and Trials". The query, which in non-guided summarization had been different for every document set is now standardized for each of the five categories.

The researchers at the University of Ottawa regularly enjoyed participating in these competitions and 2011 was no exception. A recent research trend at both at the University of Ottawa and at the National Research Council of Canada (NRC) has been emotional analysis of texts. Given the abundance of expertise easily available to them, the summarization researchers decided to incorporate emotion into uOttawa's system for this year. They hypothesized that certain emotions will be more strongly associated with summaries for each of these five categories. By identifying these emotions within a news article they hope to select better sentences for their extractive text summarization system. Essentially they proposed to identify emotional categories that are more common to the model summaries of the five news categories than they are across the document sets that they summarize. This way they could use these emotional words to identify sentences that are more likely to be useful in a summary.

There were three possible ways they hoped this research might improve their summarization system. Firstly, it is possible that people will enjoy reading summaries with more emotion to them and so might find them to be more readable. Secondly, if their hypothesis is right and summaries of one category tend to contain more emotional words, then selecting sentences with emotional words would improve

of the summarization system on Pyramid Evaluation as well. Finally, this all could cause a higher overall responsiveness.

This paper is broken down into 7 sections. This section is the Introduction. Section 2 contains a description of 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 that uOttawa submitted to TAC. A description of the TAC evaluation and the authors conclusions can be found in Sections 4 and 5 respectively. Finally Section 6 describes uOttawa's contribution to the 2011 AESOP challenge.

2 Identifying Emotion

Human cognition is capable of many nuanced emotions, but joy, sadness, anger, fear, trust, disgust, surprise, and anticipation, have been argued to be the most prototypical (Plutchik, 1980).

The uOttawa team used the NRC Emotion Lexicon v0.5 created by the National Research Council of Canada (NRC) (Mohammad and Turney, 2011) to count both emotional and sentimental words. The words in the lexicon are marked for associations with the eight prototypical emotions and also positive and negative sentiment. In addition many words with no emotional or sentiment are labelled as such. The counts of words from the emotional/sentimental classes in this data set are as follows:

- 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

Many words were labeled with multiple emotions, and so the sum of words from all emotions is greater than the total number of words associated with emotions.

2.1 Emotions by Category

The goal of the researchers was to find emotions that were most useful when making summaries for each category. To do this they determined which of the N emotions appear more than expected in the summaries of a given category. To do this they calculated the emotion density. This is done by normalizing the count of E_i by the count of all emotional words $E_{1..N}$ and non-emotional words $\neg E$

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

They calculated the emotion densities of the model summaries $P_M(E_i)$ and the document set $P_D(E_i)$. They then calculate the ratio $\frac{P_M(E_i)}{P_D(E_i)}$ to determine which emotions are more frequent in the model summaries than the document sets. These same experiments were conducted for sentiment as well as emotion. Student's t-test was applied to measure statistical significance, at $p < 0.05$, for each emotional density. The results for emotion are in Table 1 while the results for sentiment are shown in Table 2. This evaluation took place using both the A and B datasets from TAC 2010.

In tables Tables 1 & 2 the researchers noted that for all categories the number of emotional and sentiment words in the summaries was higher than in the document set, often significantly so. Their findings were that the following summary categories are most likely to contain the following emotions:

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

The uOttawa team performed one more experiment where they did not take the summarization

	Joy	Sadness	Fear	Surprise	Disgust	Anger	Trust	Anticipation	None
Accidents	1.070	**1.349**	1.079	1.036	0.998	1.254	0.842	0.966	**0.917**
Attacks	0.801	**1.220**	**1.242**	0.996	1.201	**1.378**	**0.593**	**0.590**	0.908
Health	1.127	1.171	1.163	0.973	1.158	1.271	0.790	**0.726**	0.971
Resources	1.202	0.906	1.120	**0.622**	1.197	1.070	1.073	1.021	0.968
Trial	0.797	**1.561**	**1.157**	**1.372**	**1.453**	**1.458**	0.818	**0.841**	**0.686**

Table 1: The ratio of emotion densities across the summaries and the source documents on TAC 2010 data. Bold values are statistically significant.

	Positive	Negative	None
Accidents	1.039	**1.195**	**0.924**
Attacks	0.908	**1.323**	**0.885**
Health	0.932	**1.271**	0.951
Resources	**1.305**	1.123	**0.901**
Trial	0.999	**1.522**	**0.807**

Table 2: The ratio of sentiment densities across the summaries and the source documents on TAC 2010 data. Bold values are statistically significant.

categories into account and determined if the emotion densities in summaries are significantly higher than the emotion densities in the source documents. Their findings were that the summaries had a significantly higher number of words associated with 'sadness', 'fear', 'disgust', 'anger' and also 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. They were not surprised since they suspected news to be more strongly associated with negative events and so negative emotions.

3 The System

With these findings in mind the team from uOttawa then began putting together a system. There were two main modules to this system. The first was a clustering system (Givoni and Frey, 2009) which grouped sentences together based on topic. The purpose of the clustering module was to establish main themes of each document set independent of the query. The second module was a sentence ranker (Kennedy and Szpakowicz, 2010) which selected the sentence from each cluster that was closest to the query. Two variations on this system were attempted, the first system was a baseline that used only the queries. The second system was the emotionally aware summarization system. The emotions identified in Section 2.1 were used as query expansion terms (explained further in Section 3.2). This would create summaries that are highly emotional. They wanted to investigate if using emotion words as features improves news summarization.

3.1 Module 1: Clustering

The queries available for each document set may be used to pivot the summarization process so as to best answer information need described in the query. On the other hand, each document 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, the researchers of the uOttawa team clustered sentences of each document set into topical clusters.

The clustering algorithm they used was Affinity Propagation (Givoni and Frey, 2009) which is a loopy belief propagation algorithm for exemplar-based clustering. It takes as input a matrix of pairwise similarities between data points (in the case of this system the data points are sentences) as well as 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 that maximizes net similarity—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 the researchers pre-

processed the sentences. They 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 using the *tf.idf* metric. The similarity between sentences was computed using the well-known cosine similarity metric:

$$cos(s_1, s_2) = \frac{s_1 \bullet s_2}{||s_1|| \times ||s_2||} \quad (2)$$

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 based on prior knowledge. Usually, leading sentences in each newswire article usually summarize the entire document quite well. To reflect this, the researchers decided to adjust the preference values so as to increase the likelihood to choose those sentences as cluster exemplars.

Usually, for each document set the clusterer identified at least one 'stray' cluster - a cluster with sentences that have little similarity with any other sentence in the document set. The researchers 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. The researchers found the parameter settings that maximize the value of the objective function for clustering (net similarity) and then used those settings to run on the test data.

3.2 Module 2: Sentence Ranking

The sentence ranker the uOttawa team chose to use is the same one employed in the last two years (Copeck et al., 2009; Kennedy et al., 2010) and was published in Kennedy and Szpakowicz (2010). The sentence ranker they chose uses *Roget's Thesaurus* to measure the distance between the query and a sentence in the document.

To evaluate the sentence ranker they used a corpus labeled with Summary Content Unit (SCU) information. Sentences from previous years summaries were mapped back to the original corpus and then sentences in the corpus could be labeled as containing a SCU, containing no SCUs or unknown. Only sentences known to contain SCUs or known to contain no SCUs were used for evaluation of a sentence ranker though. The actual evaluation was done by taking the mean average precision score of the known positive and negative sentences in the SCU labeled corpus. They decided to use 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 their system.

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.

The *Roget's* based sentence ranker works as follows: For each word in the query, the most closely semantically related word in a sentence is found, giving a score from 0 to 18, 0 meaning no relation, 18 being a perfect match. Closely related words, synonyms or near synonyms were given scores of 16 or 14. These word pair scores were then summed together to give a sentence score. The sentences are then ranked by these sentence scores.

There were a number of experiments performed with this sentence ranker to find the best baseline system. For example, should the query include all the aspects[1] for each topic, or should they only include the topic statement? The experiments to establish the baseline system are in Table 3. Evaluation was performed on both the A and B data sets from 2010, though their greatest interest was in the evaluation on set A, as they were aware their work did not directly apply itself to update summaries. This table includes a random baseline and a longest sentence baseline, which are largely there for comparison sake. Their finding was that including only the topic statement for each query gave the best results.

The next question was how to incorporate the emotional/sentiment words into the sentence ranker. They found that simply adding these new words

[1] An aspect is a summary-category specific question usually pertaining to the "who", "what", "where", "when", "why" or "how" of the news article.

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.

to the query was prohibitive in terms of run time, as it would require *Roget's* to measure the distance between millions of word pairs. They also believed that grouping words by closeness of semantics does not guarantee closeness of emotion. 'Happy' and 'sad' are closely related semantically, but not emotionally, as such only exact matches were used. The researchers decided to only match emotional/sentiment words exactly, but what weight should be applied to these words? They decided to try a few variations. In all cases the topic words would carry a weight of up to 18, as described earlier.

They 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 Tables 1 & 2. These results are shown in Table 4. 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.

One drawback is that the update summary – set B – does not beat the longest sentence baseline as seen in Table 3. This is a hard baseline to beat. However often the longest sentence will be 100 words or more, as such summaries would be made up by at most 1 or 2 sentences. By comparison the summaries they generated using the *Roget's* word matching method were generally 3 or 4 sentences long.

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

Although their results are only for the TAC 2010 data set it seems that the emotional/sentiment system

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 compared between the different categories on the TAC 2010 data.

significantly outperforms the baseline frequently. Resources had the smallest improvement, though the uOttawa team noted that the only emotion/sentiment that Resources correlated with was 'positive' words. This is an extremely broad class of words and does not intuitively make much sense. The researchers suspected that this was an anomaly, however they decided not to let their suspicions influence the experiment. They were optimistic as this evaluation showed that adding emotional words would improve their sentence ranking system. In theory this could lead to a higher score in the Pyramid Evaluation, and hopefully Responsiveness too.

3.3 The Final Systems

In the final system they used their sentence clustering algorithm to assign every sentence a cluster ID. They then applied the sentence ranker to rank all sentences in the document set. Sentences that were 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.

The uOttawa submission to TAC 2011 consisted

	Joy	Sadness	Fear	Surprise	Disgust	Anger	Trust	Anticipation	None
Accidents	1.000	**3.847**	2.167	2.364	3.125	2.200	1.278	0.905	0.953
Attacks	1.667	**1.900**	**2.182**	1.125	2.500	**1.921**	1.190	1.417	0.888
Health	0.913	1.920	2.038	1.000	2.154	2.059	0.895	1.047	1.072
Resources	2.833	0.923	0.857	1.400	1.200	0.923	2.136	2.500	1.094
Trial	1.00	**2.296**	**1.596**	**2.727**	**2.368**	**1.837**	0.581	1.500	0.911

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

	Positive	Negative	None
Accidents	1.143	2.267	0.923
Attacks	1.286	1.878	0.932
Health	0.949	**2.244**	0.950
Resources	**2.310**	1.077	1.012
Trial	1.00	1.816	0.931

Table 7: Sentiment count in emotional summaries normalized by count in baseline summaries on TAC 2011 data.

of two versions of every summary. One was the baseline where the query was just the topic statement, and a second where the emotional 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" under the category of 'Accidents and Natural Disasters". This category was most closely related to the emotion 'sadness'.

The uOttawa researchers 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. Tables 6 & 7 show the proportion of emotional words, by category, found in the emotional summaries, over that of the baselines summaries. That is $\frac{emotionCount(emotionalSummaries)}{emotoinCount(baselineSummaires)}$. The emotions/sentiment that the uOttawa team used for query expansion are in bold. Not surprisingly, the emotions/sentiment that were used for query expansion tend to be more frequent than the other emotions/sentiment. It would appear that they have successfully created more emotional summaries, however there is still the matter of evaluation.

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.

Figure 1: Examples of a baseline and emotional summary for document set "D1110A: Earthquake Sichuan". This summaries category is "Accidents and Natural Disasters" which was strongly associated with 'sadness'. Words related to 'sadness' are in bold.

Set	Score	Baseline	Emotion
A&B	Responsiveness	2.27	2.34
	Readability	3.06	3.09
	Pyramid	0.28	0.28
A	Responsiveness	1.26	1.25
	Pyramid	0.16	0.16
B	Responsiveness	1.01	1.09
	Pyramid	0.12	0.11

Table 8: Evaluation scores for Responsiveness, Readability and Pyramid Evaluation.

4 TAC Evaluation

Although the experiments showed promise on the 2010 data it did not follow through to the 2011 TAC data. The results in Table 8 show that the addition of emotional information did not noticeably improve either responsiveness, readability or the Pyramid Evaluation.

To be more sure the researchers examined each category individually, however they found that no single category was significantly affected by the addition of emotional words. This came as a real disappointment to them. In fact the only measure on which there was any significant change was the number of redundant SCUs counted, which was significantly increased by the addition of emotion words. This in itself seemed strange to them as expanding the query with emotion words, should mean there will be more sentences with high scores and so less likelihood of picking sentences with redundant information.

5 Conclusion and Discussion

The summaries created by the uOttawa emotional summarizer did contain many more emotional words than the baseline system. Based on their experiments with the TAC 2010 data set they were optimistic that the emotional summaries would yield an improvement in Pyramid evaluation and possible the other evaluation metrics used by TAC. This was not to be the case. That said there are some interesting results to be taken from this experiment. The increase in emotion words, while apparently not so helpful did not hurt their system at all. This itself is an accomplishment as it is not completely intuitive that emotion would be beneficial to news summaries.

There are a number of possible reasons that a significant improvement was not seen. While Table 5 showed a significant improvement in to the sentence ranker when using emotional words, this evaluation was conducted over the entire document set and not just those sentences selected for summarization. It may be possible to greatly improve overall ranking 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 uOttawa team sees this research as a starting point towards building emotional summaries where a user may direct the system to create a summary that captures expressions of anger, joy, anticipation, or some other emotion.

6 AESOP

2011 is the third year in which TAC has set its participants the task of Automatically Evaluating the Summaries Of Peers. In AESOP's first year the University of Ottawa researchers ranked summaries according to the extent to which they used source document sentences appearing in other summaries. The hypothesis was that the group as a whole would tend to select meaningful sentences; a kind of crowd sourcing. Although the uOttawa AESOP results were in the middle of the pack, analysis of the 2009 guided summarization results showed a significant correlation between summary responsiveness and sentence frequency, suggesting that "those of us who tend to use sentences used by others tend to produce more responsive summaries" (Copeck et al., 2009).

The following year the researchers' improved on the approach by distinguishing between the model summaries written by human authors and those generated automatically by peers (Kennedy et al., 2010). Human authors tend to compose a summary from nothing based on their understanding of the facts learned from reading source documents. Automatic systems with very few exceptions summarize by extracting the sentences they deem most pertinent from those source documents. Highly-rated model summaries will therefore not reuse source document sentences and their presence in the summary data set confounds it. After restricting the first run to peer

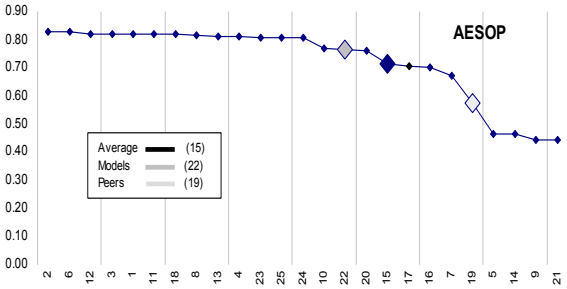

Figure 2: TAC 2011 AESOP Rating Correlation vs. a Consolidated Summarization Measure.

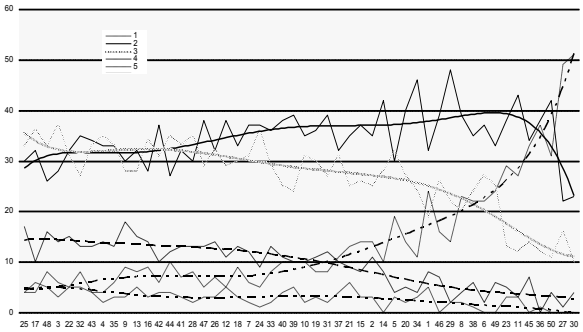

Figure 3: Summary Responsiveness Ranks vs. Peers, Least to Most, Data Points and Trend lines.

summaries, the researchers applied the hypothesis to the high quality model summaries in a second run which ranked peer summaries according to the degree to which they employed content phrases appearing in a model summary. Content phrases were taken as approximations to the concepts in the subject matter. A third run averaged the ratings of the first two approaches to address the possibility of localized data-fitting in either or both. Despite these improvements, all runs once more fell in the middle of the pack, with ratings based on model summaries outperforming those based on peers. Averaged ratings did best, suggesting that some variance in the performance of each of the two underlying measures did exist.

This year the peer summary rating procedure was recoded and slightly improvements made. The uOttawa team again submitted runs based on both sets of summaries, peers and models, and on their average.

6.1 AESOP Results

Figure 2 shows the results. They are slightly improved from the previous year, with averaged ratings appearing between the two measures on which they are based as would be expected.

The weak performance of group-based summary evaluation led the researchers this year to look more closely at the underlying data, in particular that for peer summaries. It is the more interesting case, in that given the state of the art it is hard to envision a real-world situation in which a manually-written summary would not be preferred to one generated automatically. What we discovered provides some explanation for these outcomes.

6.2 AESOP Analysis

What is the character of the set of peer summaries? What can we learn about them that would help improve an AESOP metric?

To address these questions the researchers proceeded as follows. Ratings of the key measure of summary responsiveness were tabulated for the 88 original and update summaries produced by the 50 track participants, 4400 values in all. Responsiveness is indicated by a value of 1 to 5 for least to most responsive; for each of the 88 summary topic categories[2] instances of the same value were summed. Inspection of the resulting quintuples showed that the set of topic categories can be broken down into a number of rather disjoint subsets. For instance four summaries had 5 as their most common responsiveness rating; another four had 4. Eight subsets were distinguished containing 4, 4, 12, 3, 18, 22, 8 and 17 topic categories respectively. Table 9 details their particulars. By way of contrast, a parallel analysis undertaken along the peer axis shows no similar articulation in the data. Charted from the highest scoring peer to lowest, counts of each rating value rise or fall in a generally unchanging and smooth manner as seen in Figure 3.

Table 9 reports three items of information for each subset, identified as A though H. The first column shows how many of the 88 topic categories it contains. The adjacent block of data breaks out its average responsiveness; each row totals 50, the number

[2] A topic category is either A, the original summary of the topic, or B, the update summary. The 2011 Guided Summarization track incorporated 44 topics and 88 topic categories.

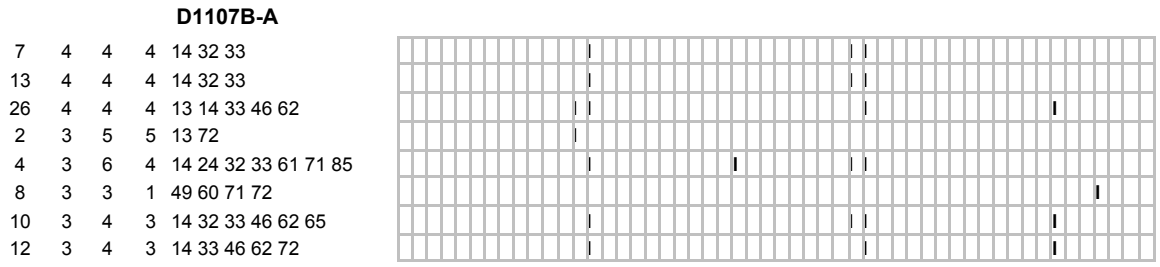

Figure 4: A Portion of the Scatterplot for Topic Category D1107B.

#		Average					Count				
		1	2	3	4	5	1	2	3	4	5
A	*4*	3	4	7	12	25	12	15	26	48	99
B	*4*	4	6	15	23	2	17	24	60	92	7
C	*12*	4	11	22	12	2	43	137	263	138	19
D	*3*	13	11	12	14	0	39	33	35	42	1
E	*18*	5	15	27	4	0	81	264	483	69	3
F	*22*	7	24	15	2	0	160	534	334	63	9
G	*8*	5	37	7	1	0	142	292	54	11	1
H	*17*	18	26	5	1	0	308	438	89	14	1
							702	*1737*	*1344*	*477*	*140*
88									**4400**		

Table 9: Emotional count in emotional summaries normalized by count in baseline summaries on TAC 2011 data.

of peer summarizers. Thus in the topic categories composing subset A, 25 participants on average received the highest possible rating; while in H, 26 peers received on average a below par rating of 2. The last data block tabulates how many individual summaries in the subset were assigned a given rating value. This block sums to 4,400, the number of summaries submitted to TAC 2011.

The researchers' next objective was to examine the individual summaries in the topic categories for sentence commonality. This directly checks our hypothesis. In order to do this the sentences in each topic category document collection were assigned indexes. Scatterplots of summary sentence indexes versus the peer summaries that use them were produced for each of the 88 topic categories. The entries in these tables were then sorted on their responsiveness rating to bring similarly-ranked summaries together. The resulting tableaux of data were inspected manually to see if highly-rated summaries used the same sentences; i.e. if they grouped together in the plot. Figure 3 shows part of the scatterplot for topic category D1107B. Its first line shows that peer #7's summary was ranked 4 for responsiveness; that this summary is composed of 4 sentences, and that each of the four was linked to a SCU. Three of the sentences could confidently be identified in the source document set as nos. 14, 32 and 33. Inspection of the plotted indexes shows peer #13's summary to be identical, while peers' #4 and #10 are supersets.

6.3 AESOP Discussion

The first stage of analysis recapped in Table 9 gives a fairly clear indication of why the researchers' scheme of ranking summaries based on the degree of their use of the same source document sentences is not more successful. The hypothesis presumes that the summaries in question contain a substantial number of well-rated ones. Table 9 shows that for subsets E through H, 65 of the 88 topic categories, the average rating is at or below the middle 'barely acceptable' rating in the scale. Of the 3,250 summaries in these categories just 14 are top ranked and only 157 rated 'good'. This is barely five percent. Thus in three-quarters of the summary data, there are almost no good sentences to identify. That fact dominates the overall performance of the approach used here.

Review of topic category scatterplots confirms this conclusion. Consider the 65 topic categories discussed above. Is it possible that the few 'good' and very few 'very good' summaries in this group use each others' sentences to such a great degree that they stand out from the poorer summaries? Inspection says otherwise. Figure 4 is representative; good summaries do tend to use the same sentences. But so do less-good ones to some degree, and summary

quality falls off imperceptibly. There is no basis in the data alone to distinguish the good from the less-good. If we look at the 23 topic categories in subsets A through D containing substantial numbers of 'good' and 'very good' summaries, we see that the same situation prevails with the added fillip that often more than one set of sentences can be used to produce a good summary.

A second observation arises from this processing of the summary data. The most important factor determining the success of automatic summarization appears to be the 'accessibility' of the topic. There is more variability in all peers' performance across topics than within the set of peers themselves.

We still have some distance to go.

Acknowledgments

Partial support for this work comes from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Terry Copeck, Alistair Kennedy, Martin Scaiano, Diana Inkpen, and Stan Szpakowicz. 2009. Summarizing with Roget's and with FrameNet. In *Second Text Analysis Conference (TAC 2009)*.

Inmar E. Givoni and Brendan J. Frey. 2009. A Binary Variable Model for Affinity Propagation. *Neural Computation*, 21:1589–1600.

Alistair Kennedy and Stan Szpakowicz. 2010. Evaluation of a Sentence Ranker for Text Summarization Based on Roget's Thesaurus. In Petr Sojka, Ales Horák, Ivan Kopecek, and Karel Pala, editors, *Text, Speech and Dialogue, 13th International Conference, TSD 2010*, pages 101–108, Brno, Czech Republic, September. Springer.

Alistair Kennedy, Terry Copeck, Diana Inkpen, and Stan Szpakowicz. 2010. Entropy-Based Sentence Selection with Roget's Thesaurus. In *Third Text Analysis Conference (TAC 2010)*.

Saif Mohammad and Peter Turney. 2011. Crowdsourcing a Word-Emotion Association Lexicon. *To Appear in Computational Intelligence*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33.