EXPLORING THE STYLE-TECHNIQUE INTERACTION IN EXTRACTIVE SUMMARIZATION OF BROADCAST NEWS

BalaKrishna Kolluru, Heidi Christensen, Yoshihiko Gotoh and Steve Renals

Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK
{b.kolluru, h.christensen, y.gotoh, s.renals}@dcs.shef.ac.uk

ABSTRACT

In this paper we seek to explore the interaction between the style of a broadcast news story and its summarization technique. We report the performance of three different summarization techniques on broadcast news stories, which are split into planned speech and spontaneous speech. The initial results indicate that some summarization techniques work better for the documents with spontaneous speech than for those with planned speech. Even for human beings some documents are inherently difficult to summarize. We observe this correlation between degree of difficulty in summarizing and performance of the three automatic summarizers. Given the high frequency of named entities in broadcast news and even greater number of references to these named entities, we also gauge the effect of named entity and coreference resolution in a news story, on the performance of these summarizers.

1. INTRODUCTION

A news broadcast is a set of stories, based on a wide variety of content, and presented in a number of styles. A broadcast news story is often a complex composition of several elements, including both planned speech (usually read) and spontaneous speech, such as a reaction or an answer. There are a number of subtle differences between spontaneous and read documents[1]; additionally, every news programme has a distinct style, including the nature of the content, the depth of information in a news story and the use of attention-grabbing headlines. For example, a weather forecast usually tends to be short and very specific as compared to an interview of a witness of an event where information could be spread out. Technically, these varied styles in broadcast news may be observed from things such as the named entity density and the rate of speaker changes [2]. This implies that the most important information, from a summarization viewpoint, is not distributed evenly across all news stories in a uniform manner. Based on these considerations we hypothesise that different summarization techniques might be best suited to different styles of news stories.

Automatic summarization of textual documents dates back to the 1950s [3], and Mani [4] presents an overview of research in this area. In recent years there has been a growing interest in the automatic summarisation of spoken language: Valenza *et al.* [5] investigated the summarization of broadcast news stories using n-gram statistics (for some smoothness), inverse document frequency (for informativeness) and speech recognition confidence measures. Zechner [6] extended these methods for spoken dialogues, using approaches such as cross-speaker information linking (eg linking questions and answers). Kikuchi *et al.* [7] have recently presented a speech summarization system based on sentence extraction and compaction of the extracted sentences.

In this paper we are concerned with interaction between the summarization technique employed and the style of news story. The automatic summarization techniques that we have investigated are based on sentence extraction, using novelty factor, content, and context as their respective criterion for summarization (explained in detail in Section 2). These automatic summarizers are compared against human generated summaries. Since we are primarily concerned with interaction between summarization techniques and broadcast style, we have used hand transcribed news broadcasts, that have been manually classified to appropriate categories, so that the potential errors of recognition can be excluded. Evaluation by human judges indicates that there is a subtle interaction between summarization technique and broadcast style. Having observed the abundance of named entities in the set of broadcast news stories that we classified, we have investigated the effect of named entities and coreference resolution on the task of summarizing broadcast news stories.

This research was supported by EPSRC grant GR/R42405 (S3L: Statistical Summarization of Spoken Language).

2. THE CORPUS AND SUMMARIZERS

We have used a portion of the hand transcripts from the Hub–4 acoustic model training data [8]. The transcripts are not case-sensitive and are devoid of any punctuation, such as sentence boundaries. For the work reported here, we manually split each segment of the transcriptions into individual news stories and marked the sentence boundaries.

2.1. Broadcast news classification

Broadcast news has been classified in a number of ways. In the Hub–4 speech recognition evaluation, focus conditions (F-conditions) were specified to categorize the acoustic conditions [8]. In the topic detection and tracking (TDT) evaluation stories were classified based on their content. For this summarization investigation we have proposed another classification mainly concerned with style, that highlights the difference between spontaneous and planned speech in broadcast news, identifying multiple speakers wherever appropriate.

News stories with spontaneous speech tend to have the summary-worthy information distributed uniformly across the document, whereas read news stories tend to start off with a "summary lead", getting into more detail as the story progresses. Also, in news stories with spontaneous speech the information layout is different if it involves only one person (typically an expert) apart from the news-reader, whose utterances usually form the core of that news story. Hence we have classified spontaneous speech stories depending on if there is an expert or not. Such an explicit classification (along the lines of zoning [9]) enables the news stories to be classified on basis of the information layout and more importantly the presentation style of a broadcast news story. News stories belonging to multiple classes can be handled by other attributes in the definitions of categories, such as number of speakers in the news story. For example, a question from the news reader during the presentation of a weather report, would mark that news story as spontaneous speech with an expert.

To evaluate and compare the performance of summarizers on different categories of news stories, we manually categorised the news stories into three classes:

(a) **Spontaneous / multiple speakers**: This category includes in it all the news stories which have both planned content and spontaneous utterances made by multiple subjects apart from the news-reader. Typically this category includes street interviews, question/answer based conversations and large group discussions.

(b) **Spontaneous / with an expert**: This category includes the news stories where a knowledgeable third party is involved in the conversation along with the reporter and/or the news-reader. It includes news stories such as interviews and individual discussions.

(c) **Reported news**: This category incorporates all the news stories whose content is pre-planned and contains no spontaneous utterance. Usually these news stories tend to be short in length compared to the other categories. Typical examples for this category are financial reports and weather reports.

2.2. Summarizers

Figure 1 illustrates the construction of extractive summaries from four spoken documents, by indicating which sentences were chosen to form part of the summary (extracted) by a human. We have used three different sentence extractive summarizers to automate this operation. The first uses a novelty factor to extract sentences for a summary, using an iterative technique that groups sentences which are similar to the document, but dissimilar to the partially constructed summary. The second selects the first line of the document (assumed to be a "summary lead") and those sentences within the document that are similar to the first sentence. The third picks up the whole chunk of text around the sentence that is most similar to the document as a whole. For all the three summarizers we apply term frequency and inverse document frequency ($tf * idf$) weighting and re-arrange the selected sentences in the order of their appearance in the original document.

2.2.1. Summarizer using novelty factor

This summarizer is based on the maximum marginal relevance (MMR) algorithm [10] proposed by Carbonell and Goldstein, and builds an extractive summary sentence-by-sentence, combining relevance (similarity to the document) with a novelty factor (dissimilarity to the partially constructed summary). At the k^{th} iteration, it chooses

$$s_k \equiv \hat{s} = \underset{s_i \epsilon D/E}{\operatorname{argmax}} \left\{ \lambda Sim(D, s_i) - (1 - \lambda) \max_{s_j \epsilon E} Sim(s_i, s_j) \right\} \tag{1}$$

where s_i is a sentence in the document, D is the document and E is the set of sentences already selected in the summary. D/E gives us the set difference, sentences not already selected. To form the summary the selected sentences $\{ s_k \}$ are re-arranged in the appearance order of the original news story. Sim is the cosine similarity measure:

$$Sim(x, y) = \frac{x \cdot y}{|x| \cdot |y|} \tag{2}$$

The constant λ decides the margin for the novelty factor, thereby having a direct impact on the nature of the summary. A $\lambda = 0.65$ was selected for experiments in this

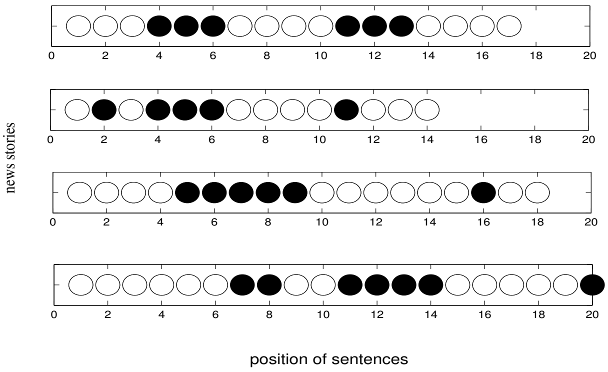


Fig. 1. Illustration showing the occurrence of sentences which are included in the summary for a given document. Each row represents a document (news story) and the sentence is represented by a circle. Each filled circle in the graph implies the sentence chosen by the human summarizer to be included in the summary.

paper, based on some preliminary experiments on another database (BBC news transcripts).

2.2.2. *Summarizer using content*

It is well-established that the first line of a textual news story is often a summary-worthy sentence (indeed, it is sometimes referred to as the "summary lead" by journalists), and this holds for some broadcast news stories. For example, in a set of 15 documents including spontaneous speech, human summarizers selected the first sentence in 11. We can use this observation to design a summarizer that extracts the first sentence, and treats it as a seed, extracting those other sentences that are most similar to it. The summary is a rearrangement of $\{ s_k \}$ that are selected by

$$s_k \equiv \hat{s} = \underset{s_i \epsilon D/E}{\operatorname{argmax}} \{Sim(s_1, s_i)\} \tag{3}$$

Sim is the cosine similarity measure of equation (2).

2.2.3. *Summarizer using context*

Another feature of extractive summarization is that highly relevant sentences tend to occur in clusters (see Figure 1). The third summarizer is based on this observation, with the sentence that is most similar to the whole document being chosen as a seed:

$$\hat{s} = \underset{s_i \epsilon D}{\operatorname{argmax}} \{Sim(D, s_i)\}, \tag{4}$$

with the summary being formed by choosing those sentences adjacent to this seed sentence, $\hat{s}$: the summary is thus the seed sentence and its context.

3. EVALUATION

Each news story was classified into one of the three categories defined in section 2.1, and four summaries (three automatic, one human) were generated. Their quality was then evaluated by human judges.

We selected 22 news stories from the corpus, which were classified into the three categories as described. Each category had 7–8 news stories and they varied in terms of size (Table 1). Each news story was summarised using each of the three automatic summarizers (novelty, content and context). The summarizers grouped the sentences forming a third of document or 100 words, whichever was larger. As a benchmark, corresponding gold-standard summaries were generated by native English speakers. For the sake of uniformity of evaluation, the human summarizers were instructed to select the sentences from the document which they would ideally include in a summary, in the order of appearance. The human summarizers were also asked to rate the degree of difficulty in summarising each document, resulting in 2 documents being classed as difficult to summarize. The four summaries for each document were then rated by a set of four human judges (different from the people who summarised the documents) using a 1–10 scale, where 10 was the best.

In order to obtain inter-judge agreement on the summarizer, we have calculated κ [11], defined by

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{5}$$

where $P(A)$ is the proportion of the times that the l judges agree and $P(E)$ is the proportion of the times we would expect the l judges to agree by chance. Given that we are looking at l judges evaluating N document/summary pairs out of a score of a maximum of M for each category, we had to calculate the κ for each category. $P(A)$ and $P(E)$ are defined as

$$P(A) = \left[\frac{1}{Nl(l-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} n_{ij}^2 \right] - \frac{1}{l-1} \tag{6}$$

where n_{ij} is the number of judges agreeing on a score of j for i^{th} summary.

$$P(E) = \sum_{j=1}^{M} p_j^2 \tag{7}$$

where p_j is proportion of the summaries assigned a score of j. If there is complete agreement then $\kappa = 1$ else if there is no agreement among the k raters $\kappa = 0$. The judges are said to be in moderate agreement when the κ is about 0.4 to 0.6. Table 2 shows the κ values for the four judges, indicating a moderate level of agreement.

categories of news stories	*number of*	*sentences*			*words*		
	documents	*min*	*avg*	*max*	*min*	*avg*	*max*
Spontaneous / multiple speakers	8	20	32	64	387	650	1562
Spontaneous / with an expert	7	17	30	54	361	630	1525
Reported news	7	16	27	48	272	570	969

Table 1. Statistics of the 22 documents (news stories) used.

Summarizer	κ	
	for all documents	*for* 20 *documents*
Human	0.49	0.50
Novelty	0.41	0.48
Content	0.39	0.45
Context	0.52	0.52

Table 2. Agreement among four judges for evaluation of various summarizers with and without the “difficult to summarize” documents

4. RESULTS AND DISCUSSION

The results of the human evaluations of the four summarizers in the three categories are shown as radar graphs in Figure 2. Each axis in a graph represents a news story while the plot is guided by the average rating for each summarizer.

The human summaries were judged to be the best for 18 out 22 stories, with the largest deviations occurring in the spontaneous/expert category, including the two documents that were classed as difficult to summarize. The human summarizers commented that these two documents were either too vague or contained a lot of information; the inter-judge agreement for these documents was low.

The automatic summarizers using novelty and content performed similar to each other for spontaneous news stories, and better than the context-based summarizer. For reported news stories, the context-based summarizer performs best on some stories, the novelty-based summarizer is best on others; on only one reported news story was the content-based summarizer the best.

The content-based summarizer performs the best in relation to all three categories, on news stories with spontaneous speech. As can be inferred from Figure 1, the insignificance of the first line in reported news stories, especially in weather reports and financial reports, it does not fare will in reported news stories. The judges point out that the summaries here lack the coherence required to form a good summary.

The context-based summarizer performs better than the other two summarizers, on the reported news category, which has a higher density of information than the other two categories. Its performance degrades on spontaneous news stories or stories with a high degree of data sparseness. The judges pointed out that this summarizer fails for spontaneous speech as it fails to highlight the real issues of the document.

The novelty- and content-based summarizers tended to lack coherence, with phenomena such as unexplained subject-object references and dangling anaphora[1]. This problem is avoided by the context-based summarizer, which produces more coherent summaries, but at the cost of occasional repetition.

5. INFLUENCE OF NAMED ENTITIES AND COREFERENCE

Given that proper names account for 9% of the total output in broadcast news and can be identified automatically with an F-measure of about 0.9 [12] (for manually transcribed data, degrading linearly with speech recognition word error rate), we wanted to see the effect of named entity identification with coreferenced names identified.

To observe the perfect named entity and coreference effect on summarisation in the three categories described before, each news story had a morphed copy, in which each named entity and its related references were manually annotated by a unique identifier. Identifying and coreferencing named entities in this way makes the coreference chains, explicit. For example:

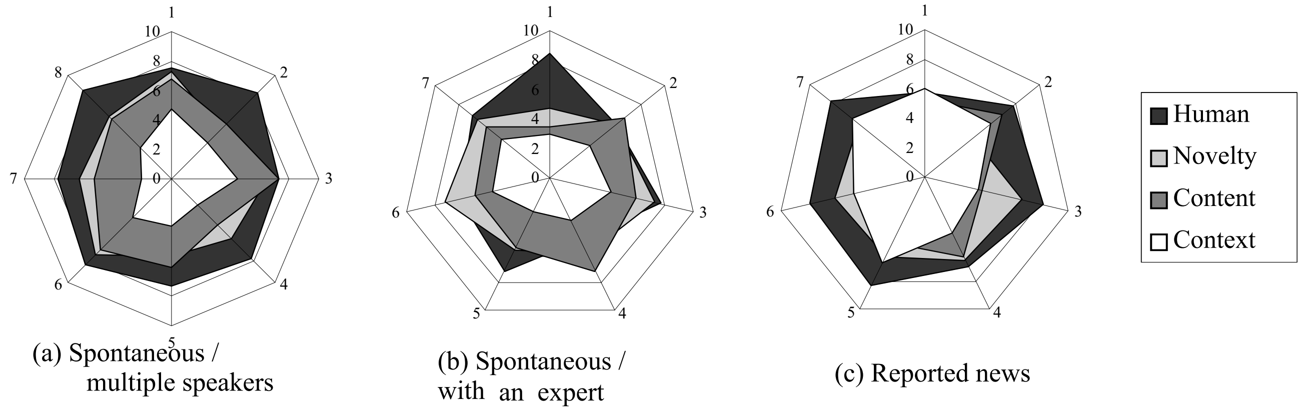
> Tony Blair said that Saddam Hussein is an evil man. He also reiterated that his evil regime must end.

Here the first “he” refers to Tony Blair and “his” to Saddam Hussein. After named entities are identified and coreferenced, the extract is transformed to:

> **tbr001** said that **shn001** was an evil man. **tbr001** also reiterated that **shn001** evil regime must end.

The summaries produced by the automatic summarizers on these coreferenced stories were evaluated by 2 human

[1] For example, *“Another reason why. . .”* in the summary without the mention of first reason, *“and it ended tragically. . .”* without mentioning what “it” was and so on.

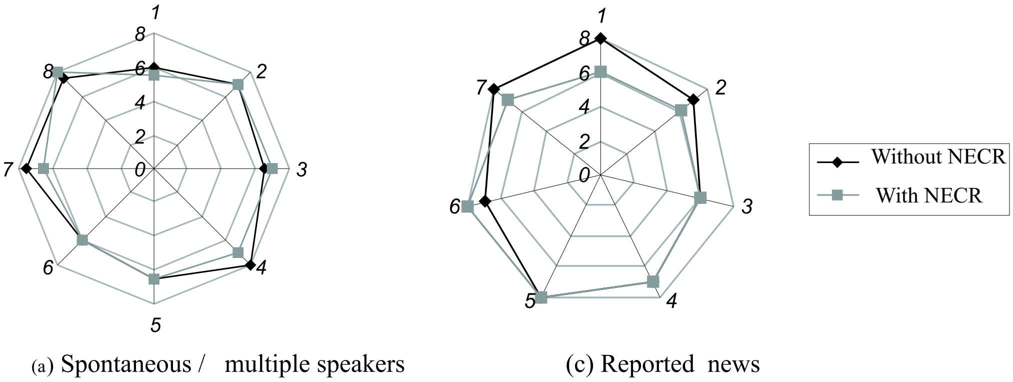


(a) Spontaneous / multiple speakers

(b) Spontaneous / with an expert

(c) Reported news

Fig. 2. Figure showing the performance of the four summarizers on all three categories. Each axis represents a news story and the plot on this axis is the average rating of the summary for that news story. Performance of the summarizers can be compared with each other by looking at the maximum overlap area.



(a) Spontaneous / multiple speakers

(c) Reported news

Fig. 3. Illustration showing the performance of novelty based summarizer on a set of documents (news stories) with and without named entity and coreferencing (NECR) marked in 2 categories with spontaneous speech and reported speech. Each axis represents a document (news story) and the plot on the axis is the average rating for the summary of that document.

judges in the same way as earlier. The radar graphs shown in Figure 3 reflect the variations in performance of same summarizer with and without coreferencing. Coreferencing of named entities has a limited impact on the summaries. Although the absolute similarities are increased, there is little change in the relative ranks of sentences. However, the automatic summarizers employed were based on unigram bag-of-words models, and coreferencing will have a limited impact in this scheme. Direct modelling of coreference chains (eg Azzam *et al.* [13] and Bergler *et al.* [14]) would be more appropriate in this situation.

6. CONCLUSION AND FUTURE WORK

The experiments reported here were performed on hand transcriptions of spoken broadcast news. The results indicate that different summarizers may be appropriate to different styles of news story, particularly considering whether the presentation consists of planned or spontaneous speech. The novelty-based summarizer performs better on spontaneous speech especially in news stories with an expert. The content-based summarizer performs consistently well on the classes with spontaneous element. Context-based summarisation technique is really limited to totally planned content.

Although there is a moderate agreement amongst the judges on the performance of three summarizers, there are certain factors that influence their scoring. Factors like personal style of scoring, which is probably not linear between the best and the poorest and personal bias towards certain events, where judges tend to look for information about what they think is right and not necessarily what the document tries to convey, are very difficult to implement by statistical means. We are considering the use of comprehension tests [15], where in the judges would have to answer questions based on the document after having read its summary, for future evaluations.

On the basis of these results, we are currently investigating a combination of coreference chains and statistical means for a summarizer. This summarizer will handle the speech recogniser output in conjunction with speaker-change detection and automatic categorisation of the document.

7. REFERENCES

[1] Sadaoko Furui, "From read speech recognition to spontaneous speech understanding," in *the Sixth Natural Language Processing Pacific Rim Symposium, Hitotsubashi Memorial Hall, National Center of Sciences*, Tokyo, Japan, November 2001.

[2] Francis Kubala, "Broadcast news is good news," in *DARPA Broadcast News Workshop*, February-March 1999, pp. 83–87.

[3] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, pp. 155–164, April 1958.

[4] Inderjeet Mani, *Automatic Summarisation*, John Benjamin's Publishing Co, 2001.

[5] Robin Valenza, Tony Robinson, Marianne Hickey, and Roger Tucker, "Summarisation of spoken audio through information extraction," in *ECSA Workshop: Accessing information in spoken audio*, April 1999, pp. 111–116.

[6] Klaus Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, vol. 28, no. 4, pp. 447–485, December 2002, Special Issue on Summarization.

[7] Tomonori Kikuchi, Sadaoki Furui, and Chiori Hori, "Automatic speech summarization based on sentence extraction and compaction," in *ICASSP*, 2003, vol. 1, pp. 384–387.

[8] "Focus conditions for broadcast news evaluation, hub4," http://www.nist.gov/speech/tests/ bnr/ hub4_96/h4spec.htm, 1996.

[9] Simone Teufel, *Argumentative Zoning: Information Extraction from Scientific Articles*, Ph.D. thesis, University of Edinburgh, 1999.

[10] Jaime Carbonell and Jade Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *the proceedings of SIGIR*, August 1998.

[11] Sidney Siegel and N. John Castellan Jr, *NonParametric Statistics for the Behavioral Sciences*, McGraw-Hill International Editions, 1988.

[12] Yoshihiko Gotoh and Steve Renals, "Information extraction from broadcast news," in *Philosophical Transactions of the Royal Society of London, series A*, vol. 358, issue 1769, pp. 1295–1310. April 2000.

[13] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas, "Using coreference chains for text summarization," in *ACL Workshop on Coreference and its Applications*, June 1999.

[14] Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz, "Using knowledge-poor coreference resolution for text summarization," in *DUC, Workshop on Text Summarization*, May-June 2003.

[15] Lynette Hirschman, John Burger, David Palmer, and Patricia Robinson, "Evaluating content extraction from audio sources," in *ECSA, ETRW Workshop: Accessing Infomation in Spoken Audio*, April 1999.