ARE EXTRACTIVE TEXT SUMMARISATION TECHNIQUES PORTABLE TO BROADCAST NEWS?

Heidi Christensen, Yoshihiko Gotoh, BalaKrishna Kolluru, Steve Renals

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

ABSTRACT

In this paper we report on a series of experiments which compare the effect of individual features on both text and speech summarisation, the effect of basing the speech summaries on automatic speech recognition transcripts with varying word error rates, and the effect of summarisation approach and transcript source on summary quality. We show that classical text summarisation features (based on stylistic and content information) are portable to broadcast news. However, the quality of the speech transcripts as well as the difference in information structure between broadcast and newspaper news affect the usability of the individual features.

1. INTRODUCTION

If possible, we would like to reuse textual summarisation techniques when summarising spoken language. However, speech transcripts differ from text documents in both structure and language, warranting an investigation of several issues concerning this knowledge transfer to the speech domain [1]. In general, how well do the information extraction techniques found the most meaningful for text documents fare on speech, and in particular what is the effect of applying a text inspired summariser to erroneous speech recogniser transcripts?

For text it has been found that good extractive summarisers depend heavily on features relating to the content of the text [2] and on the structure and style of the text [3, 4]. Content-based features are clearly vulnerable to errors introduced by a speech recognisers, and in this paper we present experiments that quantify the effect of recognition errors on summarisation. Structural and stylistic features are likely to be more robust to automatic speech recognition (ASR) errors, although they may be heavily dependent on other components such as sentence boundary detection and topic segmentation. Furthermore there are large differences in structure and style between newspaper text and a transcript of a TV news programme.

We have used two corpora to compare and contrast the summarisation of broadcast news with the summarisation of newspaper text. The broadcast news corpus was a subset of the TDT2 corpus of North American broadcast news [5], containing 114 TV news programmes (about 43 hours of speech). A comparable collection of summarised newspaper text was obtained through the Document Understanding Conferences (DUC), containing 144 news stories, each with a multiple-line extractive summary obtained from human summarisers. These datasets are described in more detail in section 2.

In section 3 we report on a series of experiments which compare the effect of individual features on both text and speech summarisation, the effect of basing the speech summaries on ASR transcripts with varying word error rates (WERs), and the effect of summarisation approach and transcript source on summary quality.

2. DATA

2.1. Broadcast news data

Research on spoken language summarisation has been limited by suitably annotated data. For this work, we have annotated a portion of the TDT–2 broadcast news corpus with human-generated extractive summaries. The TDT–2 [5] corpus has been used in the NIST Topic Detection and Tracking evaluations and in the TREC–8 and TREC–9 spoken document retrieval (SDR) evaluations. We selected a set of 114 ABC news broadcasts (ABC_SUM) from this corpus, totalling 43 hours of speech. Each programme spanned 30 minutes as broadcast, reduced to around 22 minutes once advert breaks were removed, and contains on average 7–8 news stories, giving 855 stories in total. In addition to the acoustic data, both manually-generated "closed-caption" transcriptions and a set of transcriptions from six different ASR systems, used in the TREC–8 cross-recognition experiments, are available [6]. The estimated WERs of the ASR transcripts range from 20.5% to 32.0%.

All ABC_SUM transcripts have been segmented at three levels: 1) sentence boundaries (manually segmented), 2) speaker turns (produced by LIMSI [7] for TREC/SDR) and 3) story boundaries (the individual news stories were hand-segmented as part of the TREC/SDR evaluations). The speaker turns and story boundaries were imposed on the manually-generated transcripts through alignment with the ASR outputs. In turn the sentence boundaries were imposed on the ASR outputs by aligning the manually-generated transcript to the ASR outputs. All alignments were done automatically but manually checked to limit segmentation errors.

The 114 broadcasts in ABC_SUM span the period from February to June 1998. They have been randomly divided into three separate datasets for training, development and testing. Table 1 presents the statistics for the partitions.

For each segmented story in the ABC_SUM data, a human summariser selected one sentence as a "gold-standard", one-sentence extractive summary. These one-sentence summaries were all produced by the same human summariser, and an evaluation experiment was carried out to check their consistency and qual-

This research was supported by EPSRC grant GR/R42405 *S3L: Statistical Summarization of Spoken Language*. We would like to thank the participants of the TREC SDR track in the late 1990s for making their ASR transcripts available at that time.

part	# words	# sentence	# documents	# hours
All	283.419	21589	114	43.1
Train	235.593	17.948	95	33.8
Dev	24.996	1966	10	3.9
Test[a]	22.871	1679	9	3.4

[a]This partition was not used for the work presented here.

Table 1. Statistics for the train, development and test part of the ABC_SUM

$\hat{\kappa}$	All	Native	Non-native
All stories	0.56	0.55	0.59
Long stories	0.34	0.34	0.39
Short stories	0.82	0.79	0.81

Table 2. $\hat{\kappa}$ values indicating above chance agreement between the 6 summarisers (4 native and 2 non-native), estimated for all stories, stories longer than the median and stories shorter than the median.

ity. Five additional human summarisers each produced summaries for four of the broadcasts (44 news stories), randomly chosen and spread reasonably in time throughout the corpus. To assess the level of agreement between the summarisers the κ statistic was used [8]. κ is a measure of agreement between judges, taking into account the agreement one would expect to see from pure chance. An estimate of κ, can be found as follows

$$\hat{\kappa} = \frac{\hat{P}(A) - \hat{P}(E)}{1 - \hat{P}(E)} \quad (1)$$

where $\hat{P}(A)$ is the estimate of the proportion of inter-summariser agreements and $\hat{P}(E)$ is the estimate of the expected proportion of chance agreement. $\hat{\kappa}$ is 1 if the judges are in perfect agreement, and $\hat{\kappa}$ is 0 when there is only chance agreement. Table 2 shows the estimated $\hat{\kappa}$ values quantifying the degree of agreement for the six human summarisers. In general the $\hat{\kappa}$ values indicate that the summarisers' agreement is above chance level, with the $\hat{\kappa}$ estimate equal to 0.56 for all summarisers. Looking at the native (4 persons) and non-native (2 persons) summarisers separately, there seems to be no noticeable difference in agreement.

Calculating the $\hat{\kappa}$ value according to (1) assumes that the number of categories each judge can choose from is constant. However, in this case, the categories (ie. the number of sentences to choose from in a given story) vary from story to story. The story length over the entire ABC_SUM corpus adheres to a rather bimodal distribution with many very short stories (typically 2-5 sentences) and many longer stories (20-30 sentences), the latter of course leaving room for much more disagreement between summarisers. A separate $\hat{\kappa}$ measure was therefore calculated for stories below and above 16 sentences (the median story length). Analysing the short and long stories separately gives a more nuanced picture of the constituents of the $\hat{\kappa}$ value and indeed as expected, on the short stories the summarisers have a very high agreement ($\hat{\kappa}$ values around 0.80), whereas for longer stories the $\hat{\kappa}$ drops to 0.34 - 0.39 indicating a far more difficult task. The summarisers commented that the task of picking a single sentence to represent an entire news story was at times very difficult, and the observed disagreements can partly be attributed to the difficulty in defining (and interpreting) what is meant by a suitable one-sentence summary of a news story.

The average pair-wise $\hat{\kappa}$ calculated between the main summariser and each of the other summarisers in turn, was 0.57. Overall, the analysis of inter-summariser agreement shows above chance agreement for the tested documents, and therefore the summarisation work done by the main summariser is considered of sufficient quality and consistency, and the summaries are reasonable candidates for a "gold-standard" set of summaries.

Extracting one-sentence summaries is closely related to automatic headline generation [9, 10]. The role of a headline will typically differ from that of the one-sentence summary; headlines tend to be more compact (no length restriction was put on the summaries for ABC_SUM) and can be classified as being either *eye-catchers*, *indicative* or *informative*. The ABC_SUM summaries are all sought to be as informative about their corresponding news story as possible.

2.2. Newspaper data

We have used text data obtained from the DUC-2001[1] text summarisation evaluation. This data consists of newspaper stories originally used in the TREC–9 question answering track, totalling 144 files (132 for training, 12 for testing) from the Wall Street Journal, AP newswire, San Jose Mercury News, Financial Times, and LA Times, together with associated summaries[2]. Each document comprises a single news story topic, and the data is from the period 1987-1994. Although the speech data is from 1998, the broad topics covered in the two data sets are very similar.

3. EXPERIMENTS

The summarisation task is to automatically generate an extractive summary for a spoken or printed news story. Our approach uses a trainable, feature-based model which assigns a score to each sentence that indicates how suitable that sentence is for inclusion in a summary. When generating an N-line summary, the summary is comprised of the N highest-scoring sentences.

A set of features are extracted for each sentence. The summariser is based around a set of multi-layer perceptron (MLP) classifiers [11]:, one for each feature (*feature-MLPs*) and a second level MLP (*merger-MLP*) which combines the outputs of the *feature-MLPs* (figure 1). This feature-based approach is somewhat similar to that employed by [12]; that approach discretised the features and was based on a Naive Bayes classifier. The training set for each *feature-MLP* consists of a set of single feature inputs, together with the summarisation label from the "gold-standard" (1 or 0), for each sentence. Thus each *feature-MLP* is trained to optimise summarisation for that feature alone. Given a set of trained *feature-MLPs*, a *merger-MLP* may be obtained from a training set in which each sentence is represented as the vector of *feature-MLP* outputs. This two level architecture was primarily chosen because

[1]url = http://www-nlpir.nist.gov/projects/duc/index.html

[2]Extractive summaries for this data were contributed by John Conroy (IDA) as an addition to the non-extractive summaries distributed with the original DUC-2001 data, and were derived to cover the same content as the non-extractive summaries.

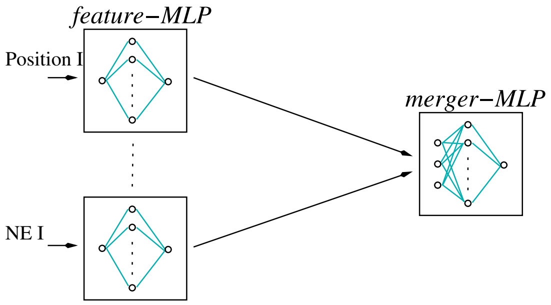

Fig. 1. Summariser architecture. All MLPs used in this work had 20 hidden units in a single hidden layer.

Feature	: Description
POSITION I	: Reciprocal position from the start.
POSITION II	: Sentence position from the start.
LENGTH I	: Length of sentence in words.
TF.IDF I	: Mean of normalised *tf.idf* terms.
COSINE I	: Cosine similarity measure of *tf.idf* terms.
NE I	: Number of NEs.
NE II	: Number of first occurrences of NEs.
NE III	: Proportion of different NEs to number of NEs.

Table 3. Description of sentence-level features. The ’start’ and ’end’ are relative to the boundaries of the particular news story topic. NE = named entity. Counts of NEs are per sentence.

it facilitates the analysis of the contribution of each features, by sampling the performance of the *feature-MLPs*.

We investigated a large set of candidate features, which could be divided into four categories: position of the sentence in the story, length of the sentence, similarity of the sentence to the overall document, and distribution of named entities (NEs) within the sentence. After some preliminary experiments, we settled on the set of eight features listed in table 3. The first three features can be classified as structural features, and are concerned with length and position. The remaining features concern the content of the sentence. TF.IDF I and COSINE I are based on traditional information retrieval term weights comprising information about *tf* (*term frequency*) and *idf* (*inverse document frequency*) [13]. The COSINE I is the cosine similarity measure of the *tf.idf* term vector to the document term vector. The final three features all concern the NE distribution in the sentence. For the text data NE annotations from the DUC evaluations have been used. The speech data has been run through an automatic NE recogniser [14].

We have used both automatic evaluations (with respect to the “gold-standard” summaries) and user-tests with human judges to investigate the portability of text document summarisation approaches to broadcast news. Section 3.1 looks at the difference between the contribution of the individual features for printed and spoken data, section 3.2 presents results from experiments with summarisation of ASR transcripts with different WERs and finally section 3.3 describes the outcome of the human evaluation of the perceived quality of the different speech summaries.

3.1. The contribution of the individual features

We assessed the contribution of an individual feature by basing a summariser on the relevant *feature-MLP* alone. Figure 2 shows the ROC curves for four of the single feature summarisers and a summariser combining the whole feature set, each operating on both text and speech data. For both text and speech the summariser based on the full feature set had the best performance characteristics. For text, the positional feature POSITION I is clearly the most informative for summarisation; for speech there is no similarly dominant feature. This is linked to the stylistic differences between print and broadcast media. Printed news stories typically present the most important facts in the opening line, with subsequently related facts presented in the order of decreasing importance (the “inverted information pyramid”): indeed the opening line is often referred to as the “summary lead”. Broadcast news is rather different: it is written to be heard, and the lead sentence(s) often aim to capture the interest of the viewer or listener, without summarising the main facts in the opening sentence. Furthermore, the information density within the story is not uniform, and depends on the style, for example the news anchor may speak information-rich sentences, compared with an interview.

These stylistic differences are also reflected in the contribution of the last style feature, the length feature (LENGTH I). For text, the sentence length is of less importance, but for speech it contains a lot of discriminative information about whether a sentence is summary-worthy. In the speech domain, the high information regions in the stories are often from the anchor in the studio, the main reporter or the occasional expert. It is often well-formed speech with longer sentences (either read or partly scripted speech). In contrast short sentences tend to be less information-rich.

The conclusions are similar when looking at the other main group of features, the content features. In text none of these features have been able to compete with the simple, yet very effective position features. In the speech domain, the content features contribute significantly. A very noticeable difference is for the named entity based features. Their performances in the text domain are relatively poor, but again the uneven information distribution in speech means that named entities become much stronger indicators of fact filled sentences. The *tf.idf* based features tell much the same story.

A final point to note is that for text the combination of the complete eight features added only minimal improvement to the performance of the best single feature summariser - based on the simple position feature. In the speech domain, the single feature summarisers are more complementary and their combination is significantly better than any of them on their own.

Although the newspaper text and the broadcast news speech data are chosen so as to be as closely matches as possible, one crucial difference is the type of evaluation summaries: multi-line summaries for the text and one-sentence summaries for the speech. This discrepancy between data sets adds a level of complexity when drawing conclusions from these experiments. In terms of the contribution of the individual features it is likely that the apparent lack of contribution from some of the content features on the text data is partly down to the fact that when creating a multi-line summary any sentence candidate must not only be high in information relevant to the content of the story but also be a complementary match to sentences already selected in the summary (a novelty factor is incorporated into approaces such as the MMR [2]).

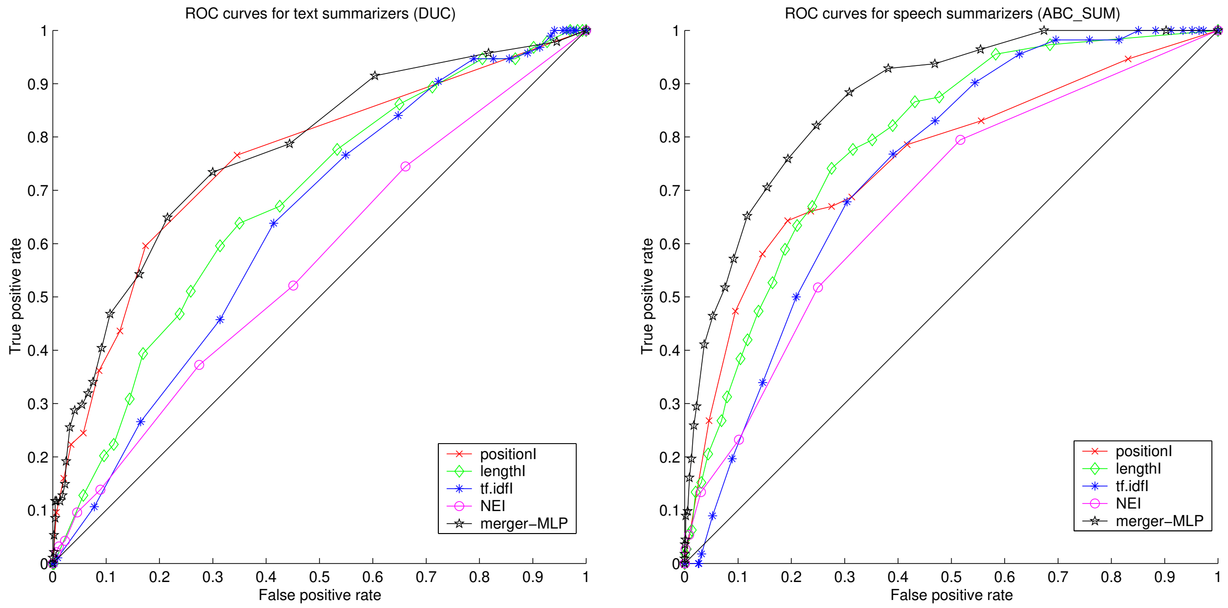

Fig. 2. . Influence of the various features on the text and speech summarisers - ROC curves for the individual features and their combination to newspaper summarisation (DUC; left) and broadcast news summarisation (ABC_SUM; right).

3.2. Summarising ASR transcripts

The above experiments on broadcast news were carried out on manual, closed-caption transcriptions. Although these transcripts are not error-free (WER of 14.5%) they are still far better than transcripts from ASR systems. However, applications for automatic summarisation of spoken news story would have to make do with transcripts output from automatic speech recognisers. This section reports on experiments carried out to assess the effect of using ASR transcripts of varying WER.

Figure 3 shows the ROC curves for speech summarisers based on transcripts from six different ASR systems (produced for the TREC–8 SDR evaluation), along with the manual transcript. Each summariser was trained and tested on transcripts from the same source. The curves for the different transcripts are very close, and these results indicate that there is relatively little difference due to WER, although the summariser based on the recogniser with the highest WER does show some degradation in performance. This relative indifference to WER, similar to that observed in spoken document retrieval using this data, can be explained, at least in part, by the structure of a typical broadcast news story. The most information rich parts of a broadcast news story tend to correspond to planned studio speech; spontaneous speech in variable acoustic environments is less information rich, from the point of view of summarisation—and harder to recognise. Zechner *et al.* report an increase in summarisation accuracy and a decrease in WER on broadcast news summaries by taking into account the confidence score output by the ASR system when producing the summary, and thereby weighting down parts of speech with potentially high WERs [16].

Clearly, factors such as the structure of the news story, the WER of the transcript and the types of feature do have an effect on the summary. For ABC_SUM, the structure of the news stories varies: some are organised like typical broadcast stories with the characteristic diffuse spread of information, others are more reminiscent of newspaper stories. A thorough investigation of the interaction of these factors would require a proper categorisation of the news stories. We have carried out preliminary experiments based on a very simple categorisation, whereby long and short news stories were processed separately. The majority of the short stories of around three sentences are almost always read speech, spoken by the anchor in the studio and the information is presented in a straightforward manner with the most important facts introduced first. The long stories, on the other hand, tend to have the content rich sentences spread throughout the news story.

Figure 4 shows four plots arising from doing summarisation on long and short news stories based on high WER (shef-s1), low WER (cuhtk-s1) and manually-generated transcripts. Each plot shows ROC curves from four *feature-MLP* summarisers as well as from the *merger-MLP* combining all eight features.

Comparing plots for the long and short stories (left-hand column to right-hand column) shows that the different types of feature perform differently depending on the style of the news story. On the long stories the position feature is much less important than for the short stories. The sentence length and the *tf.idf* based features, on the other hand, are far more important in longer stories. This further confirms the found link between feature contribution and structure of news story, and is in line with the conclusions drawn in section 3.1.

Only subtle differences in summarisation accuracy arise from an increasing WER. The curves for the manual and cuhtk-s1 transcripts are very similar. For the long/shef-s1 combination the area under the ROC curves is smaller, reflecting the increased number of errors in the transcripts. A larger difference is observed for the short/shef-s1 stories where the length and content based features have dropped in performance, in contrast to the position feature (which is not directly dependent on the speech recognizer).

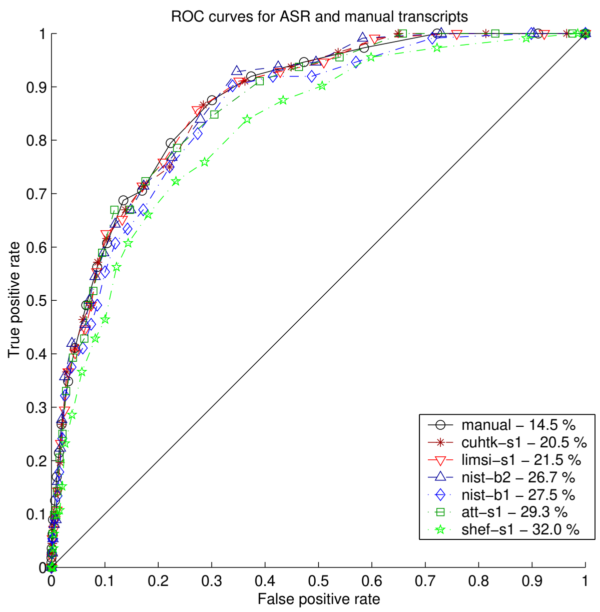

Fig. 3. . The influence of various WERs on the speech data summarisers - ROC curves for summarisers corresponding to different quality ASR transcripts plus the manually-generated transcript. WERs for the ASR and manual transcript are measured against 10 hours of TREC-8 reference data [15].

The experiments have shown that the optimal choice of features when transfering text features to broadcast news, is susceptible to both the structure of the news story and the quality of the transcripts.

3.3. Human perception of different summaries

The proximity of the curves in Figure 3 indicates that only for relatively high WERs was there any degradation to be found in the ability of the summarisers to pick the "gold-standard" sentence. However, this type of evaluation fails to investigate the effect of the WER on the quality of the summary. It is easy to imagine that even a single word substitution, with a small impact on the WER, can change the meaning of a sentence completely.

Another issue is the difference between the expected nature of the errors found in the "closed caption" transcripts as opposed to the ASR induced errors. Errors in the manual transcripts are made by humans and often occur because the steno-captioner has misheard or possibly is unable to keep up. However, the resulting transcripts are in general grammatical and therefore more readable than erroneous ASR transcripts.

To better address this point, a final set of experiments was conducted, where 8 judges were asked to give a utility score to various summaries. They were each given the full manually-generated transcript for each of 44 news stories (from the same 4 broadcasts as the human summarisers were asked to check, see Section 2.1), and for each story they were asked to evaluate 9 different summary candidates. Each summary candidate was given a score between 1 and 10 (the best).

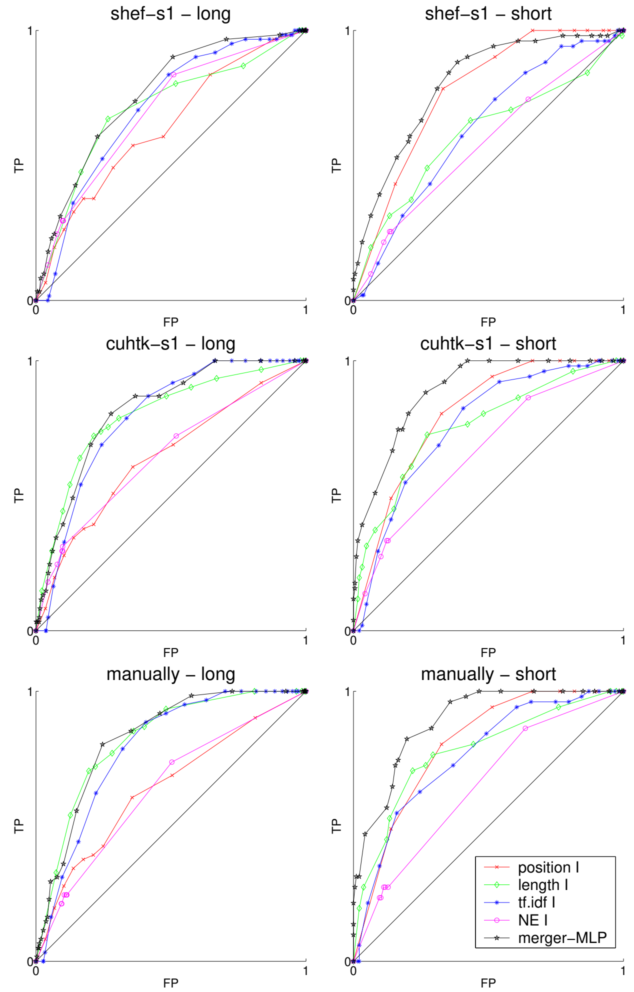

Fig. 4. The performance of summarisers based on all features and four typical single feature summarisers on long and short news stories and high WER (shef-s1), low WER (cuhtk-s1) and manually-generated transcripts.

The 9 summary candidates were obtained by generating three different types of summary on three different types of transcript. The three summarisation approaches were:

- **"gold-standard"** , as generated by the human summariser.
- **few_features**, a summariser combining a reduced set of features: POSITION I, LENGTH I, TF.IDF I, COSINE I, and NE III.
- **all_features**, a summariser combining the full feature set (all eight features from Table 3).

The three different transcripts were: 1) manually-generated (closed captions), 2) from the cuhtk-s1 ASR (relatively low WER ASR), and 3) from the shef-s1 ASR (relatively high WER).

Table 4 shows the average score for each of the summariser/transcript combinations. Comparing the effect of applying the summarisers to different types of transcripts, we found that the human judges had a preference to the summaries based on the manual transcripts. The scores for the manually-generated transcripts are all higher than for the ASR transcripts. This is in con-

Summariser /Transcript	gold standard	all features	few features	avr.
manual	6.96	6.63	5.80	6.46
cuhtk-s1	5.73	5.45	4.53	5.24
shef-s1	4.98	3.34	4.31	4.21
avr.	5.89	5.14	4.88	

Table 4. Average score from human judges. Each score is the average of the scores of 8 judges for 44 news stories. The scores for each human judge have all been normalised to lie within the full range 1-10.

trast to the results from the automatic evaluations that failed to detect any noticeable difference in performance between manually-generated transcripts and the cuhtk-s1 transcripts.

The quality of the summaries produced by the three types of summariser were also perceived to be different. The "gold-standard" type summary has the highest average score (as expected). The feature based summaries scored lower when using the reduces feature set compared with the full feature set, apart from when using the high WER transcripts.

It is important to note that often not even the "gold-standard" summary for the manually-generated transcripts was thought by the judges to be a very good summary for the particular news story. More research is needed to investigate whether a one-sentence extractive summary is only useful for particular types of news stories. Since the news stories the judges were given to evaluate were the same as those the test summarisers worked on, we know for certain that some of these stories were very difficult to summarise; for example for one 29 sentence story each of the six human summarisers chose a different summary sentence. Therefore some of the "gold-standard" summaries must be considered as less than definitive.

Both the ROC curve inspection and scoring using human judges belong to the category of *intrinsic* evaluation (concerned with comparing summaries using different criteria, [17]). In future work we plan to investigate more *extrinsic* evaluation methods directed towards assessing how useful a summary is for carrying out a particular task (eg. comprehension).

4. CONCLUSIONS

We have investigated the portability of extractive text summarisation techniques to broadcast news. We assessed the contribution of individual features (stylistic and content-based) by investigating ROC curves for summarisers based on newspaper data and broadcast news data respectively. It was found that for text the position feature is very dominating, and features containing content information are less important. For speech however, the stylistic features and the content features were all significant. Looking at the effect of using transcripts of speech obtained from ASR systems, the automatic evaluation showed only a degradation in performance when the WER became high. However, a test with human judges made clear that even small errors in the transcripts can reduce the perceived quality of the summary.

We have shown that classical text summarisation features (based on stylistic and content information) are largely portable to the domain of broadcast news. However, the experiments reported here also made evident that the different characteristics of a broadcast news story, such as the different information distribution and the effect of different types of transcript error, warrant more sophisticated information extraction techniques, where the organisation of summary-worthy information in the news story is more explicitly taken into consideration.

Future work will investigate in more details the effect on the summary quality of using automatically detected sentence boundaries, speaker changes and news story segments rather than manual segmentations as in the data employed in this work, along with automatic feature selection algorithms for this task.

REFERENCES

[1] K. Zechner, "Spoken language condensation in the 21st century," in *Proceedings of EUROSPEECH*, Geneva, Switzerland, 2003.

[2] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries.," in *Proceedings of SIGIR*, 1998.

[3] H. P. Edmundson, "New methods in automatic extracting," *Journal of the ACM*, vol. 16, no. 2, pp. 264–285, 1969.

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

[5] C. Cieri, D. Graff, and M. Liberman, "The TDT-2 text and speech corpus," in *Proceedings of DARPA Broadcast News Workshop*, 1999.

[6] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proceedings of RIAO-2000*, April 2000.

[7] J. Gauvain, "The LIMSI SDR system for TREC," in *Proceedings of TREC-9*, 2000.

[8] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, 1960.

[9] W. Kraaij, M. Spitters, and A. Hulth, "Headline extraction based on a combination of uni- and multidocument summarization techniques," in *Proceedings of DUC-2002*, 2002.

[10] B. Dorr, D. Zajic, and R. Schwartz, "Hedge trimmer: A parse-and-trim approach to headline generation," in *Proceedings of Workshop on Automatic Summarization*, May 2003.

[11] C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, England, 1995.

[12] J. Kupiec, J. O. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceesings of SIGIR*, 1995.

[13] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 2001.

[14] Y. Gotoh and S. Renals, "Information extraction from broadcast news," *Philosophical Transactions of the Royal Society of London, series A*, vol. 358, pp. 1295–1310, April 2000.

[15] S. E. Johnson, P. Jourlin, K. Spärck Jones, and P. C. Woodland, "Spoken Document Retrieval for TREC-8 at Cambridge University," in *Proceedings of TREC-8*, 2000.

[16] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *Proceedings of NAACL-ANLP-2000*.

[17] I. Mani and M. T. Maybury, Eds., *Advances in Text Summarization*, MIT Press, Cambridge, MA, 1999.