Multi-Stage Compaction Approach to Broadcast News Summarisation

BalaKrishna Kolluru  Heidi Christensen  Yoshihiko Gotoh

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

Abstract

This paper presents a fully automatic, multi-stage compaction approach to broadcast news summarisation, targeting transcripts from automatic speech recognition (ASR) systems. It employs a network of multi-layer perceptrons to remove incorrectly transcribed words based on confidence scores, and to select significant chunks at multiple stages based on tf.idf scores and named entity frequency. The resulting summaries are assessed using a combination of cross comprehension test and a fluency test, finally compared with an automatic evaluation scheme. The experimental results show the approach can produce summaries with good information content.

1. Introduction

A news broadcast is a fine blend of content and style: in contrast to the rather chaotic mixture observed in conversational speech. While content is information about a certain story, such as a description of events, details of a weather forecast or a scientific discovery, style is the information layout: how a news story is presented over the course of a broadcast. Disfluencies in a broadcast are observed infrequently; the word error rate (WER) from state-of-the-art automatic speech recognition (ASR) systems is near 10–20% level. It is thus an ideal testbed for porting technologies developed for written text into the speech domain, and for developing new approaches for processing spoken data like summarisation [1].

Speech summarisation is a developing area of research. To date, there have been several studies mainly in the broadcast news domain. Using lexical and prosodic information, most approaches have tried to extract important sentences from transcripts. Valenza et al. used multiple features such as n-gram statistics, inverse document frequency (idf) and confidence measures [2]. Kikuchi et al. investigated sentence extraction and compaction techniques on the basis of acoustic score and lexical information [3]. Stokes et al. used lexical chains to build a very short gist from a news caption [4]. Koumpis and Renals employed the PARCEL algorithm with ROC (receiver operating characteristics) analysis in voice-mail summarisation [5].

The aim of our current research is to develop an approach to produce generative (i.e., non-extractive) summaries from ASR transcripts. In this paper, we present a first step towards the goal — namely, to generate a one line (12–15 words) gist for a broadcast news story using an automatic, multi-stage, progressive, compression technique. Our approach attempts to capture the context rather than simply extracting a list of significant keywords. Thus, using this approach, we are able to cover the content while not losing too much of the style. This may be achieved by retaining the most relevant phrases in a news story at each stage of compaction. We use a multi-layer perceptron (MLP) to arrive at an optimum combination of lexical features such as the term frequency-inverse document frequency (tf.idf) scores and the named entity (NE) frequency (Section 2). We also present an evaluation scheme based on a comprehension test to determine the information coverage with reference to each news story. The scheme is designed to account for the human subjectivity when producing summaries (Section 3). The experimental results show the approach can produce summaries with good information content from ASR transcripts in comparison to a simple sentence extraction method. We are currently looking at ways to improve fluency (Section 4).

2. Approach

2.1. Compaction Procedure

We use a combination of MLPs, part-of-speech (POS) tagger, partial parser, etc., in order to select important chunks automatically from ASR transcripts. The procedure is illustrated in Figure 1, which is described below:

1 Using the confidence scores, MLP(A) identifies the words that are incorrectly transcribed by a speech recogniser.

2 Sentence boundary markers are set using a statistical model1 — a lack of segmentation at various levels (e.g., sentence, paragraph, topic) for ASR transcripts is one of the most striking differences from written texts, and is a non-trivial problem.

3 A sentence boundary closest to the first 500 characters of the ASR transcript is identified — in text processing, it is widely acknowledged that the first few paragraphs of newspaper/newswire articles are information rich zones [7]. This aspect has also been observed in broadcast news [1]. Given the absence of paragraphs in speech, we assume that the most significant information pertaining to a news story occurs in the first 500 characters (typically 60 to 70 transcribed words).

4 Abney’s partial parser [8], in conjunction with the Brill tagger [9], is used to split a sentence into chunks at multiple levels with increasing granularity (from a coarse to a fine chunk level). The Brill tagger is trained with the Switch Board Corpus obtained from Penn Tree Bank.

5 Disfluencies are removed based on POS tags [10].

6 NEs are identified using a finite state model approach2.

---

1 This approach to sentence boundary detection integrates scores from language modelling and pause duration. It resulted in precision and recall (P&R) score of 70% for ASR transcripts of BBC news [6].

2 This approach to NE identification resulted in P&R score of 90%
Figure 1: Compaction procedure: MLP(A) identifies incorrect words using the confidence scores. MLPs (B1), (B2), and (B3) select significant chunks using tf.idf scores and NE frequencies.

7 Important chunks are identified by MLP(B1), (B2), (B3) — MLP(B1) works at a coarse level; it removes less significant chunks so that the ASR transcript is reduced to approximately 300 characters. It is then reduced to 200 characters and finally to 120 characters using (B2) and (B3). An example of how this multi-stage compaction works is illustrated in Figure 2.

2.2. Dataset and MLP Training

Dataset. In the experiment, we used a subset of broadcast news stories from the TDT-2 corpus [12]. They were used for NIST TDT evaluations and the TREC-8 and TREC-9 spoken document retrieval (SDR) evaluations. They consist of 43 hours of speech; each programme contains 7 to 8 news stories on average, spanning 30 minutes as broadcast which may be reduced to 22 minutes once advert breaks are removed. 542 news stories were manually selected according to the following criteria:

- being sufficiently long (average length of their hand transcriptions: 487 words in 25 sentences);
- having news-print style layout (no metaphorical references or innuendos at the beginning of a news story).

They were split randomly into 426, 65, and 51 stories for training (of MLPs and other statistical components such as sentence boundary detection and NE identification), development, and testing dataset. For the training stories, we have made hand annotations including the following categories: story boundary, sentence boundary, named entity, coreference, and important chunks at various levels.

MLP(A). The purpose of MLP(A) is to filter out the incorrectly transcribed words. It is trained with confidence scores obtained from outputs of the ASR system. Towards this end, ASR transcripts are aligned with the corresponding hand transcriptions, then each word is marked with either ‘CORRECT’ or ‘INCORRECT’. Figure 3(a) shows the ROC curve by MLP(A) using the development data.

for hand transcriptions of broadcast news. The performance declines linearly with the increasing amount of WER in ASR transcripts [11].

3. Evaluation Schemes

Evaluation is a non-trivial task, requiring a very careful design. For the manual evaluation, we devised a cross comprehension test and a fluency test. The former is an extension of work done by Hirschmann [13] for assessing the information spread and the latter is to see the grammatical correctness of a summary.

3.1. Reference Summaries and Questions

Four sets of human authored summaries were prepared. For each story from the test set a 12–15 word summary was produced separately by four people. They were then used as references for the manual and automatic evaluations. Furthermore, these four people prepared a questionnaire for each summary

MLP(B). MLP(B)’s task is to select significant chunks based on the sum of tf.idf scores and the NE frequency for each chunk. For hand transcriptions from the training dataset, chunks were manually annotated at multiple levels if they contain significant information. The annotation was tailored to the purpose of the experiment so that three stage reductions (approximately 500 → 300 → 200 → 120 characters) were made using MLP(B1), (B2), and (B3). Figure 3(b) shows the ROC curve by MLP(B1) using the development data.
they produced. The questionnaire typically consisted of two to four questions; the correct answer can be drawn from the corresponding summary. An example of a question set and the corresponding reference summary is shown below:

**Reference**
British millionaire’s attempt to circumnavigate the world in hot-air balloon has failed.

**Questions**
1. What was attempted?
2. Who was involved?
3. What was the outcome?

These four human-authored summaries will be referred to as ‘reference A, B, C, D’ and ‘question set A, B, C, D’ in the following.

### 3.2. Manual Evaluation

**Cross comprehension test.** This test measures the informativeness of summaries. Three different types of summaries are tested in the experiment:
- ‘baseline’: the first informative sentence of a news story (this excludes greeting messages such as ‘Good Evening’);
- ‘compaction’: the approach described in Section 2;
- ‘reference A’: a human authored summary from a randomly chosen set from the four sets.

Each of the above summaries is read in turn and a judge examines whether each question from ‘question set B’ (that corresponds to ‘reference B’) can be answered by that summary. The judge will be a different person from the reference summary authors. When the answer is found, it may be relevant, partially relevant, and totally irrelevant to the one expected from ‘reference B’. Thus, the decision is made from the following four options:
- **correct**: a relevant answer is found in the summary;
- **partially correct**: a partially relevant answer is found;
- **incorrect**: an irrelevant answer is found;
- **not found**: no answer is found.

The procedure is illustrated in Figure 4. Human authored summaries are not treated as a gold standard in its traditional sense, but instead the scheme accounts for human subjectivity.

**Fluency test.** A judge also scores the fluency of summaries between 1 to 5, based on the following criteria:
- 5: fluent, grammatical and sensible;
- 4: between 3 and 5;
- 3: understandable, not fluent, some grammatical errors;
- 2: between 1 and 3;
- 1: incomprehensible, factual errors.

In this paper, the cross comprehension test and the fluency test together give a measure of quality for a summary.

### 3.3. Automatic Evaluation

For automatic evaluation, we employ an n-gram recall technique as proposed by Lin and Hovy [14]. Their tool, ROUGE, essentially compares the number of n-gram matches in the reference and automatic summaries. Recently, ROUGE was used for evaluation of summarisation from newspaper/newswire articles [15]. ROUGE simulated manual evaluation well for that task, although it is still unclear how closely it scales to other tasks.

### 4. Experiments

Shown below are the average lengths of summaries generated from each of 51 test stories:

<table>
<thead>
<tr>
<th></th>
<th>#words</th>
<th>#characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>19.7</td>
<td>141.2</td>
</tr>
<tr>
<td>compaction</td>
<td>21.7</td>
<td>132.0</td>
</tr>
<tr>
<td>reference A</td>
<td>16.3</td>
<td>113.3</td>
</tr>
</tbody>
</table>

Although 120 characters were targeted by the compaction algorithm, it resulted in slightly longer summaries because we did not impose a strict limit.

**Cross comprehension test.** Figure 5(a) shows the average for four answer options for three different types of summaries. Unsurprisingly the human authored ‘reference A’ achieved the best among the three, and the ‘baseline’ was the lowest. The ‘compaction’ approach performed much better than the ‘baseline’ but not as well as the ‘reference A’.

However, the most striking aspect of this outcome is that, even for human authored summaries, less than 50% of questions were answered correctly and no answer was found for more than 40% of questions. This illustrates the human subjectivity involved in producing summaries. Furthermore, almost uniform number of incorrect answers were observed for three different sets of summaries. We surmise that this was also due to human subjectivity; occasionally summary authors arrived at contradictory summaries, resulting in opposing views of the same story. This is possibly due to the ambiguous nature of the stories. In such cases, a questionnaire was produced from that
author’s rather subjective point of view, and this certainly af-
affected answers from not only reference A’, but the ‘baseline’
and the ‘compaction’ as well.

It should also be noted that, for nearly 80% of questions for
the ‘baseline’ and about 60% of questions for the ‘compaction’ approach, no answer was found. It was partially be-
cause the ‘baseline’ was the single sentence picked up from
25 sentences (average) in a story. In that sense, the informa-
tion concentration in the first sentence is still useful. The com-
paction algorithm was applied for the first few sentences only
(approximately 500 characters), thus there were a significant
amount of transcripts totally ignored from the processing. This
may be alleviated by applying the scheme to the entire story.

**Fluency test.** Figure 5(b) shows the average fluency scores for
the three different types of summaries. This time, the human
authored ‘reference A’ was the most fluent, followed by the
‘baseline’. Most ‘baseline’ summaries were readable, however
they sometimes suffered by disfluencies of the original speaker
and/or speech recognition errors. Although the ‘compaction’
summaries did not perform as well as the other two, most of
them were found readable and understandable. We are studying
additional steps to improve the fluency.

**Automatic evaluation.** Figure 5(c) shows the average uni-
gram recall scores calculated by ROUGE. Unigram match for
two types of summaries was tested against three more human
authored summaries (‘reference B’, ‘C’, and ‘D’). The results
were roughly comparable to those by the cross comprehension
test (i.e., the ‘reference A’ achieved the best, followed by the
‘compaction’, then the ‘baseline’ in the bottom. This seems to
indicate that this automatic scheme is able to measure the infor-
mative nature of the summary to some extent, but fails to evaluate
its fluency.

5. Conclusion

In this paper, we have presented a fully automatic, multi-stage,
compaction approach to broadcast news summarisation, target-
ing transcripts from ASR systems. It was evaluated using the
cross comprehension test, the fluency test, and an automatic
approach. The results demonstrate the ability of a network of
MLPs to identify the most informative chunks from ASR tran-
scripts, however there is room for improvement in fluency.

We are currently investigating the following directions:

- The compaction could be extended from the first 500
  characters to the entire story of the ASR transcripts.

- Addition of different types of features may improve the
  MLP classifiers’ performance. One may be the corefer-
  ence — important links to the proper names were often
  lost by the current algorithm. The length of each word
  may also be an interesting feature — the average length
  for a word in the ‘compaction’ summaries was 5.1 char-
  acters, noticeably shorter than that in the ‘baseline’
  (6.2) and the ‘reference A’ (6.0). We assume longer words

---

This requires further study because, although ‘roughly compara-
bale’, the unigram scores were not simulating the cross compre-
henension test well — e.g., the ‘baseline’ score (0.139) was 75% of the ‘refer-
ence A’ (0.186) by ROUGE, while the correct answers for the ‘baseline’
(0.166) was 34% of the ‘reference A’ (0.485) by the cross compre-
henension test.

**6. References**

- [1] H. Christensen, Y. Gotoh, B. Kolluru, and S. Renals,
  “Are extractive text summarisation techniques portable to
  broadcast news?,” *ASRU Workshop*, U.S. Virgin Islands,
  2003.

  “Summarisation of spoken audio through information ex-

  ization based on sentence extraction and compaction,”
  *ICASSP*, Hong Kong, 2003.

  “Broadcast news gisting using lexical cohesion analysis,”
  in *Advances in Information Retrieval*, Springer-Verlag,
  2004.

  voicemail messages using lexical and prosodic features,”
  *ACM Transactions on Speech and Language Processing*,
  Submitted for publication.

  in broadcast speech transcripts,” *ISCA Workshop*, Paris,
  2000.

  sation of electronic publications by sentence selection,”
  *Information Processing and Management*, vol. 31, no. 5,
  1995.

- [8] S. Abney, “Parsing by chunks,” in *Principle-Based Pars-
  ing: Computation and Psycholinguistics*, Kluwer Aca-

- [9] E. Brill, “Some advances in transformation-based part of

  of spoken dialogues in unrestricted domains,” *SIGIR*, New

  broadcast news,” *Philosophical Transactions of the Royal

  S. Strassel, “The TDT-2 text and speech corpus,” *DARPA

  “Evaluating content extraction from audio source,”

  using n-gram co-occurrence statistics,” *HLT-NAACL*, Ed-

  trinsic evaluation of generic news text summarization sys-