Speaker Role Based Structural Classification of Broadcast News Stories

BalaKrishna Kolluru        Yoshihiko Gotoh

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

Abstract

This paper is concerned with automatic classification of broadcast news stories based on speaker roles such as anchor, reporter and others. The story classification is the first step for many related tasks such as browsing, indexing, and summarising the news broadcast. We use broadcast news audio and its automatic speech recogniser transcripts to implement the classification system. It builds on speaker segmentation and identification, story segmentation and named entity identification. It has achieved 92% accuracy when individual stories were provided manually. The performance declined to 67% and 51%, of precision and recall related measures respectively, when combined with automatic story boundary segmentation.

Index Terms: structural classification, speaker role

1. Introduction

A news broadcast is a continuous stream of many individual stories of varying length. Several factors account for the length of a news story such as the news value of the story (i.e., relevance to its listeners/viewers), the amount of air-time available, the order of its appearance in that newscast, and availability of the news story. For news stories with known for a story). The information layout is composed based on the nature and content of the stories. For news stories with lower content (e.g., a daily update of financial market data) editors may rely on a short story with a few sentences, simply read out by the anchor. For longer news stories, audio content (sound bites) from the actual scene may be used to grab listeners’ attention, or the reaction of a witness can provide more information about the incident. All these aspects implicitly create a genre according to their presentation styles. Journalists’ handbooks (e.g., [1]) define news structures based on speaker roles, sound bites and other aspects of the news stories.

The layout of news stories is a combination of the narration style and information spread. Structural classification is the first step when processing broadcast news for tasks such as summarisation. Our previous experiments indicate that different categories of news stories warrant different summarisation techniques to yield better summaries [2]. For example, short stories read by the anchor, can be summarised using the observation that most important information is likely to occur in the first few sentences. On the other hand, longer stories involving specialist speakers may warrant a different summarisation technique because the useful information is often spread over the entire story. As these stories are type-casted into various categories at the time of production, such a classification can also be used for automatic browsing and indexing of individual news. It is also useful when studying the time-line of a news story — e.g., how a breaking news story progresses from a short 3-sentence snippet to a 15-minute full length feature.

There have been many studies in ‘topic’ based classification of news broadcasts, however the number of works concerned with structural classification are relatively sparse. Barzikay et al. implemented a classifier that identified speaker roles in a news program into anchor, journalist and guest based on certain cue phrases [3]. They have achieved 77% accuracy using a maximum entropy method. Their work was built on automatic speech recogniser (ASR) transcripts with individual speaker segments being provided manually, whereas we attempt to build a system that classifies stories within the news program, when story boundaries are not provided, using both broadcast audio and its ASR transcripts. Maskey and Hirschberg used structural features such as a position of a sentence and an identification of a speaker in order to build a summary which was independent of speech recognition errors [4]. However, the purpose of their speaker identification was to find the contribution of a certain speaker to the broadcast rather than to assign any special role.

In this paper, we address the problem of automatically classifying broadcast news stories into predefined categories based on speaker roles. It builds on various processing components, such as speaker segmentation and identification, story segmentation, and named entity (NE) identification, all of which are later combined using a set of rules. The story structure classification system has achieved 92% accuracy when individual stories were provided manually. The performance declined to 67% and 51%, of precision and recall related measures respectively, when combined with automatic story boundary segmentation.

2. Structural Categories

2.1. Categories from Journalists’ Vocabulary

The following descriptions are found in various books for journalists (e.g., [5], [6], and [7]): Reader is a very short snippet of news, typically not more than a few sentences, usually placed in the middle of a news broadcast and read out by the anchor. Voicer has a pattern of speaker turns. Typically the anchor introduces the story, while the reporter, usually at the scene of an incident/event, presents the detail. Wrap involves other speakers apart from the anchor and the reporter. This style grabs listeners’ attention by adding the actualities. Typically these stories are presented by the anchor and taken by the reporter on the ground along with a few spontaneous comments from the news maker (e.g., embedding Clinton’s speech in a news story).

2.2. Structural Classification Task

We modified the journalists’ categories so that the news layout can be identified using a simple set of rules. Given the definition below, the task is reduced to identification of speaker roles:

Reader is a news story where the anchor is the only speaker. The length of the story is not considered, although the majority are short stories.

Voicer is presented by the anchor and the reporter.
In order to test the idea of structural classification, a small experiment was conducted by two human subjects ($S_A$ and $S_B$) before implementing the automatic classifier. They manually classified 102 news stories into proposed categories after reading hand transcribed closed-captions with manually segmented story boundaries. Speaker turns were shown, but their identities were removed prior to the experiment. For each story, the nature of speakers were identified first and the story was classified based on instances of speaker roles. The process was then repeated while listening to the corresponding audio broadcast.

Table 1 shows the confusion matrices. For the classification solely based on the closed-captions (on the left side panel of Table 1), two subjects did not agree on categories for 11 out of 102 stories. The largest disagreement was observed between Voicer and Wrap; after the experiment human subjects commented that it was often difficult to identify ‘other’ speakers based on a closed-caption alone, due to the lack of obvious clues. The disagreement was alleviated when the corresponding audio broadcast was available (right side panel). Subjects $S_A$ and $S_B$ revised their classification for 13 and 9 stories, respectively, after listening to the audio. They had differed only on one news story, which was due to a very ambiguous nature of the story with varying acoustic conditions and no lexical cues. Further, there were several stories, which both of them initially classified as Voicers, actually were Wraps according to extra information found in the audio.

The above result indicates that, although it is possible to arrive at consistent categories based on instances of speaker roles, the task is difficult when relying on textual materials alone. Even for human subjects who worked on closed-captions with a relatively small amount of transcription errors, it was often not possible without additional information such as speaker identities. The above observation has led to our implementation of the automatic classifier in Section 3; it utilises an audio stream and its ASR transcript. For the rest of the experiments (Section 4), Subject $S_A$’s classification was used as a reference.

### 3. Implementation of a Classifier

#### 3.1. Classification Procedure

The classifier involves several components: ASR, speaker segmentation and identification, story boundary detection and NE identification — all of which are combined using a set of rules.

1. Software developed at the Carnegie Mellon University (CMU) is used to segment and to identify various speakers from broadcast news audio stream [10]. It segments audio into predefined classes derived from acoustic conditions. It also clusters the segments spoken by the same person and assigns a unique identification (ID).
2. The ASR was previously used for the SDR task [11]. The WER was 32.0% for the TDT–2 broadcast news data.
3. A ‘news story’ corresponds to the individual topic in the

<table>
<thead>
<tr>
<th>$S_A$</th>
<th>$S_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>V</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>9</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) CC only

<table>
<thead>
<tr>
<th>$S_A$</th>
<th>$S_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>V</td>
</tr>
<tr>
<td>R</td>
<td>41</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
</tr>
</tbody>
</table>

(b) CC & audio

Table 1: Individual stories were classified by two human subjects ($S_A$ and $S_B$) after reading the closed-captions (CC), with and without listening to the corresponding audio broadcasts. ‘R’, ‘V’, ‘W’, and ‘I’ indicate Reader, Voicer, Wrap, and Interview categories, respectively.

---

2.4. Manual Classification Experiment

In order to test the idea of structural classification, a small experiment was conducted by two human subjects ($S_A$ and $S_B$) before implementing the automatic classifier. They manually classified 102 news stories into proposed categories after reading

Figure 1: This figure shows the length distribution of news stories for each structural category. Story lengths for 102 closed-caption transcripts were measured by the number of sentences. Please note that the ordinate is scaled according to the number of stories in each category.

**Wrap** involves instances of the anchor, the reporter and the other speaker(s) in the news story.

**Interview** is a conversation between the anchor and the other speaker(s) without involvement of a reporter.

**Reader**, **Voicer**, and **Wrap** categories correspond to reported news, spontaneous news with a reporter, and ones with multiple speakers in our earlier work [2]. In that study, context-based, novelty-based, and relevance/novelty-based summarisation schemes, respectively, were found to perform well for these categories. In this paper, the **Interview** category is additionally introduced in order to cover news stories without a reporter.

#### 2.3. Data Set

For the experiments presented in this paper, we use a subset of broadcast news stories from the TDT–2 corpus [8]. They were used for NIST TDT evaluations and the TREC–8 and TREC–9 spoken document retrieval (SDR) evaluations. They consist of 114 ABC news programs; a typical program contains seven to eight long news stories and several more short stories. Each program spans 30 minutes, reduced to around 22 minutes once ad- vert breaks are removed. In addition to audio broadcast streams, manually generated closed-caption transcripts, with the word error rate (WER) of 13.5%, are available [9]. Closed-captions are provided with manually annotated story boundaries.

Nine news programs, accounting for 102 individual stories, are randomly selected for testing the structural classification scheme. The rest of programs are used for construction of statistical models for story segmentation and NE identification. Figure 1 shows statistics for story. **Readers** were short stories with a high peak at around five sentences. **Wraps** were longer stories with roughly 25 sentences and the larger variance. **Voicers** and **Interviews** occurred only a few times in the test set.

#### 2.4. Manual Classification Experiment

In order to test the idea of structural classification, a small experiment was conducted by two human subjects ($S_A$ and $S_B$) before implementing the automatic classifier. They manually classified 102 news stories into proposed categories after read-

2. The ASR was previously used for the SDR task [11]. The WER was 32.0% for the TDT–2 broadcast news data.

3. A ‘news story’ corresponds to the individual topic in the
ASR transcripts. Prosodic and lexical cues are combined using the maximum entropy framework [12].

4. A trainable, statistical finite state model is used to identify NE’s in the ASR transcripts [13]. The NE model used here was initially developed for the 1999 Hub4 IE-NE task [14], with additional data being manually annotated on ‘anchors’ and ‘reporters’ on the training set so that these two categories can also be identified.

In the experiment, a story is reconstructed using the automatic speaker and story segmentation components. It is then classified based on identification of speakers and their roles (i.e., anchor, reporter, and other). Improvement achieved by the use of NE information is also investigated.

3.2. Rules for Speaker Role Identification

After a careful examination of news programs in the training data, the following rules are established. The first three rules rely on ASR transcripts with speakers identified.

**Rule 1:** The first speaker in a news story is the anchor. Occasionally the anchor returns in the same news story — it may be identified by examining the speaker ID.

**Rule 2:** A speaker immediately after anchor’s cue phrases is the reporter. The following cue phrases are used: ‘here’s ABC’s < . . . >’, or ‘ABC’s < . . . > reports’ where < . . > can be any word (sequence). It is typically a reporter’s name however, in this particular rule, we did not examine whether it is really a reporter. The return of the reporter can be identified by the speaker ID.

**Rule 3:** If a speaker is neither an anchor nor a reporter, (s)he is marked as another speaker.

The above rules assign one role to each speaker in the ASR transcripts. It is then straightforward to make a structural classification of that story using the definitions in Section 2.

The approach may be refined using NE information. To this end, the following two rules are implemented, where <reporter> and <location> are any word (sequence) in the ASR transcript, that are identified as a reporter and a location.

**Rule 1a:** The anchor may state cue phrases: ‘here’s ABC’s <reporter>’, or ‘ABC’s <reporter>’.

**Rule 2a:** The reporter may state cue phrase: ‘<reporter> ABC news <location>’.

Using all five rules, anchors are identified by rules 1 and 1a, then reporters are identified based on rules 2 and 2a. Finally, other speakers are marked according to rule 3. The literature review suggests that the above rules are not very far from those implemented by the news producers [5, 6, 7].

There are two notes relating to the implementation. Firstly we were aware that, for this particular data set, the story length could have been a good indicator for classifying Reader and Wrap categories (see Figure 1). However we did not exploit this feature due to the existence of Voicer and Interview whose story lengths overlapped with the other two categories. Secondly in this story classification task, our assumption is that the source of the broadcast (i.e., ‘ABC news’) is known, and the classifier utilises manually selected, source specific rules. Although this process can be automated, it is out of the scope of this paper. (In our earlier work, we have shown that cue words around the story boundary could be selected automatically using the maximum entropy approach [12].)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>A</th>
<th>R</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) without NE

(b) with NE

Table 2: For 1726 speaker turns, the above confusion matrices compare speaker roles identified by Subject S_A on the closed-captions and the automatic classifier on the ASR transcripts. The left side panel is the outcome based on rules 1, 2, and 3 only, while the right side panel used all rules (1, 1a, 2, 2a, 3).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>R</th>
<th>V</th>
<th>W</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) without NE

(b) with NE

Table 3: 102 individual stories from the ASR transcripts were classified into one of four categories by the automatic classifier, then compared with the classification by Subject S_A on the closed-captions. ‘R’, ‘V’, ‘W’, and ‘I’ indicate Reader, Voicer, Wrap, and Interview categories, respectively.

4. Automatic Classification Experiments

4.1. Classification of Manually Segmented Stories

In the first set of experiments, news stories in the ASR transcripts were manually segmented by alignment with the corresponding closed-captions. Table 2 shows confusion matrices for speaker role identification by Subject S_A on the closed-captions and the automatic classifier on the ASR transcripts.

There were 1726 instances of speaker turns, out of which 58% of speaker roles were identified correctly based on rules 1, 2, and 3 only. The number has improved to 64% when NE information was integrated by using all rules (1, 1a, 2, 2a, 3).

Failures in anchor identification were usually caused by speaker identification errors. Occasionally the anchor returned in the middle of the broadcast without any textual cue: such instances could not be identified when it was not enumerated with the same speaker ID as the first speaker of the story. Identification of reporter relied on occurrences of specific cue phrases (rule 2), thus susceptible to ASR errors. It was also caused by speaker identification errors. Distinction between reporters and other speakers was the most difficult to implement, due to a lack of appropriate cue phrases. However, integration of NE information did help improvement in this aspect. It resulted in reduction of mis-classification between reporters and other speakers.

Table 3 shows the confusion matrices for structural classification of news broadcasts, resulting from speaker role identification in Table 2. Out of 102 stories, 82% and 92% were correctly identified based on rules 1, 2, 3 and on all five rules, respectively. The scheme was relatively robust; failure in speaker role identification was not always translated into story classification errors. Not surprisingly, the performance was negatively affected by ASR errors and speaker identification errors. However, it was alleviated by use of NE information, that was often able to re-install the correct role for speakers.
4.2. Classification of Automatically Segmented Stories

In the second set of experiments, we run the classifier on the ASR transcripts with automatically segmented story boundaries. The maximum entropy based segmentation scheme resulted in 78 news stories (in comparison to 102 stories by hand) from 9 news programs [12]. It was found particularly difficult to identify short stories. Figure 2 illustrates the alignment of stories between the automatically segmented ASR transcript and the manually segmented closed-caption (reference). Two measures exist; the first is to examine whether the category identified by the automatic classifier on m1, m2, m3, m4, while the ‘recall counts’ examines whether reference categories m1, m2, m3, m4 are correctly identified by the automatic categories a1, a2, a4, a4.

Table 4 shows the precision and recall counts when all five classifiers are compared with reference categories a1, a2, a3, a4. We defined the structure of news stories derived from journalistic analysis of speakers and their roles. It was shown that, even for human subjects, speaker role identification was not always an easy task when it was solely dependent on ASR transcripts. The classifier was designed to utilise broadcast audio stream. It built on speaker segmentation and identification, story segmentation, and NE identification, all of which were later combined using a set of rules. The classifier has achieved 92% accuracy when individual stories were provided manually, and 67% and 51% of precision and recall when combined with automatic story segmentation. The improvement of the story segmentation component is the key step to further refine the structural classification.

5. Conclusion

We experimented the structural classification of broadcast news stories based on speaker roles such as anchor, reporter and others. We defined the structure of news stories derived from journalists’ classification, which was closely related to the appearance of speakers and their roles. It was shown that, even for human subjects, speaker role identification was not always an easy task when it was solely dependent on ASR transcripts. The classifier was designed to utilise broadcast audio stream. It built on speaker segmentation and identification, story segmentation.

6. Acknowledgements

This work was funded by UK EPSRC grant GR/R42405, Statistical Summarisation of Spoken Language (S3IL).

7. References