

# EVALUATING CONTENT EXTRACTION FROM AUDIO SOURCES

Lynette Hirschman, John Burger, David Palmer, Patricia Robinson  
{lynette,john,palmer,parann}@mitre.org

The MITRE Corporation  
202 Burlington Rd.  
Bedford, MA 01730, USA

## ABSTRACT

This paper discusses evaluation of content extraction from audio sources. The most straightforward approach is to adapt existing methods for *written* sources to handle audio input. A transcription then becomes the representation of the audio source in written form; it must capture the word stream, but also other information that aids in decoding the overall structure and content of the audio source, e.g., music, speaker changes, and speech repairs. The transcription must also support content annotation superimposed on the underlying speech transcription. When automated speech recognition is used to generate the transcription, there is the additional problem of evaluating content extraction from a noisy transcription. In addition, audio sources differ from their written counterparts in genre and therefore in structure, vocabulary, and even in how names are used. If the audio includes spontaneous conversational speech, as opposed to planned speech, these differences become still more pronounced. We discuss how these differences affect the adaptation of text-based extraction evaluation to audio input. In addition, we describe two new content extraction evaluations that have been designed for use with both audio and written materials.

## 1 INTRODUCTION

If we wish to evaluate content extraction from audio sources, the most obvious approach is to apply methods developed for written sources. We can draw heavily on previous experiences in the Message Understanding Conference (MUC) evaluations[10][11].

These MUC-style extraction tasks have formed the basis for recent work in audio content extraction in the DARPA Broadcast News evaluations[8][9]. This work has shown that it is possible to adapt written evaluations to audio sources, provided that we pay explicit attention to certain differences.

- A *transcription* mediates between the audio source and its written representation. This means it must represent not only the words, but also the other information contained in the audio stream, e.g., speaker change, speech repairs, and structural clues, such as music, noise, or commercial breaks.

- Content annotation conventions become an important issue, because these must co-exist with the audio transcription annotations.
- It is important to preserve cross-indexing between the (time-stamped) audio stream and its representation in the transcription. This permits alignment of a noisy automated transcription with a high-fidelity human-prepared transcript, which makes possible re-use of many of the text-based evaluation techniques.
- Content and genre differences may require adjustments in the evaluation when applied to audio data, in terms of task appropriateness and/or amount of data required for test and training.

In the following sections, we review these issues. In the final section, we present a summary of two new evaluation approaches that have been proposed specifically with audio content in mind.

## 2 EXISTING EVALUATIONS

The Message Understanding Conferences were launched in 1987 to evaluate information extraction on written text[4]. Over this period, the written text research community developed a suite of content evaluations, including:

- Identification of named entities (e.g., person or organization names) via SGML mark-up in text;
- Identification of the list of unique entities in a document;
- Extraction of relations among these entities;
- Creation of a database of task-specific information about a particular topic, such as terrorist attacks; this was defined in terms of a “scenario template” containing slots for the type of attack, perpetrator, victim, damage, date, location, etc.

Early evaluations (through MUC-5 in 1993) focused on the most difficult task, namely database creation. However, in 1995, MUC-6 introduced several simpler “component” extraction tasks, including Named Entity. Since then, Named Entity extraction has been evaluated as part of the Multilingual Entity Tasks (MET-1 and 2), MUC-7, and most recently, in the DARPA Hub4 Broadcast News evaluation.

The Named Entity task is defined in terms of SGML annotations in the text, indicating names, times and numerical expressions.

Name expressions consist of named persons (<PERS>*Madeline Albright*</PERS>), locations (<LOC>*Lebanon*</LOC>), and organizations (<ORG>*Federal Reserve*</ORG>). Numerical expressions consist of monetary and percent expressions, e.g., *ten million dollars* and *6.5 percent*. Time expressions designate absolute temporal expressions, such as dates, e.g., *Wednesday* and times, e.g., *10:00 GMT*.

Performance on the Named Entity task is determined by comparing an automatically generated system response (a.k.a. hypothesis) against a human generated key (a.k.a. reference). This comparison is straightforward for written sources, since it can be assumed that only the annotations differ, not the underlying text. The score is calculated in terms of F-measure, which is a combination of precision and recall.<sup>1</sup> Systems must correctly label the *type* of expression – whether it is a person or a location, for example – as well as determine the *extent* of the phrase, i.e., define its boundaries, in order to get full credit.

Named Entity extraction was declared a solved problem for the domain of written well-structured text<sup>2</sup> in MUC-7. High performing MUC systems had scores from between 93% and 96% precision and recall, which compared favorably with human performance (94-97%)[11]. Due in part to this success, the Named Entity evaluation was chosen as a content evaluation task for the DARPA Broadcast News audio data.

In order to adapt the MUC evaluation to this new domain, a number of changes were required. First, content annotation guidelines had to be adapted to apply to audio transcriptions. Unlike journalistic texts, broadcast news transcriptions do not have case or punctuation to help with the detection of named entities. For example, given a mention of *SOUTH PHILADELPHIA*, it is not clear whether the entire phrase should be marked as a location, or only *PHILADELPHIA*. This required some refinement of the guidelines for what should be classified as a named entity given audio source data. The current guidelines are available at the NIST Web site [http://www.nist.gov/speech/hub4\\_98.htm](http://www.nist.gov/speech/hub4_98.htm).

Second, the content mark-up conventions had to co-exist with the transcription conventions used for the underlying audio data, specifically disfluencies, such as repetitions, restarts, and hesitations. Finally, a new scorer had to be built to address the noisy input by speech recognition systems.

Figure 1 shows the results from the Named Entity task from MUC, the MET multilingual evaluation (for Chinese and Japanese) and for broadcast news transcription at 0% word error and at 15% word error. These results illustrate that named entity performance degrades slightly for “clean” transcribed text (0% word error rate) – to 90%, compared to 93% for written journalistic text; and it degrades to 82% for automatic

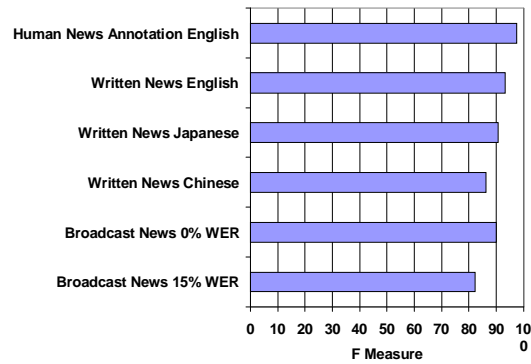


Figure 1: Named Entity Results for Written and Audio Sources

transcription (15% word error rate). Nonetheless, these results indicate that useful content (named entities) can be extracted automatically from broadcast news audio sources.

### 3 TRANSCRIPTION CONVENTIONS

The transcription is the textual representation of the audio stream. As such, it must represent the information conveyed by the audio stream. This obviously includes a representation of the actual words uttered, but also information such as speaker change, background noise, and – particularly important for later content annotation – disfluencies. Example 1 shows such a transcription for the Hub4 Broadcast News data, displaying some of the transcription conventions for indicating pronunciation of acronyms (\_C\_N\_N), filled pauses (%uh) and word fragments (C-). Typically, content annotations such as named entities have also been represented in-line with the source text – as described above, MUC used SGML mark-up. Therefore, transcription must carry a second level of information related to content. Example 2 shows the same transcription, hand-annotated for named-entity (PERson, LOCation, ORGANization).

Hello from the \_C\_N\_N center in Atlanta, welcome to the second hour of C- of %uh \_C\_N\_N morning news, I'm Donna Kelly.

Example 1: Transcription Conventions for Audio Sources

---

Hello from the <ORG>\_C\_N\_N</ORG> center in <LOC>Atlanta</LOC>, welcome to the second hour of C- of %uh <ORG>\_C\_N\_N</ORG> morning news, I'm <PER>Donna Kelly</PER>.

Example 2: Transcription with Content Annotation

<sup>1</sup> Recall is number of correct annotations found divided by the total number of correct annotations; precision is number of correct annotations found divided by the number of annotations actually found. Balanced F-measure is computed as 2\*P\*R/P+R.

<sup>2</sup> These findings were specifically for journalistic text, which has relatively few spelling errors, careful typographic conventions and regular punctuation.

|           |      |     |       |          |          |          |        |        |        |       |         |         |         |     |     |
|-----------|------|-----|-------|----------|----------|----------|--------|--------|--------|-------|---------|---------|---------|-----|-----|
| Hello     | from | the | <ORG> | <u>C</u> | <u>N</u> | <u>N</u> | </ORG> | center | in     | <LOC> | Atlanta | </LOC>  | welcome | to  | the |
| <u>OH</u> | FROM | THE | <ORG> | <u>C</u> | <u>N</u> | <u>N</u> | </ORG> | CENTER | </ORG> | IN    | <LOC>   | ATLANTA | </LOC>  | WAS | THE |

Figure 2

```
<speaker_1> <PERS>William Jefferson </speaker_1>
<speaker_2> right </speaker_2>
<speaker_1> Clinton</PERS></speaker_1>
```

Example 3: Named Entity Across Speaker Turn Boundary

```
OH FROM THE <ORG>_C_N_N CENTER</ORG> IN
<LOC>ATLANTA</LOC> WAS THE SECOND HOUR
OF THE <ORG>_C_N_N</ORG> MORNING NEWS IM
<PER>DONNA KELLEY</PER>
```

Example 5: System Generated Transcription

These two levels of annotation create superficial problems due to incompatible mark-up conventions. For example, both speaker turns and named entities are indicated by SGML. When a speaker utters a name that is interrupted by another speaker's back channel response (see Example 3), this creates a mark-up problem: how should the name be marked, without including the back channel response? It also creates a "proper bracketing" problem, since normally, names are inside of speaker turns, not across them.

Another serious problem is the need to mark disfluencies for intended content. For example, if a person stumbles in saying a location and then repairs it, how does this get annotated? This is illustrated in Example 4: should all the instances of "Lebanon" be annotated? Which are required and which are optional?

```
Transcription:
I ONLY GO GO TO LEBANON INTO LEBANON
LEBANON LEBANON LEBANON
Content Annotation:
I ONLY GO GO TO <ORG>LEBANON</ORG>
INTO <ORG>LEBANON LEBANON LEBANON
LEBANON</ORG>
```

Example 4: Disfluencies and Content Annotation

One of the challenges of adapting text-based content measures to audio involves coordinating and reconciling these layers of annotation, to make them compatible and sufficiently informative for the task. In addition, there is an urgent need for tools to support simultaneous transcription and SGML-based mark-up for audio sources.

Our experience with the Named Entity evaluation for Hub4 has led us to believe that explicit annotations for disfluencies and the associated repairs would be useful. This would allow an annotator to view a "repaired" version of the transcription for content annotation and would provide a more flexible approach to scoring content annotation, especially for highly disfluent speech, such as found in the various spontaneous corpora (Switchboard, Call Home). Explicit annotation disfluencies (and the associated repairs) would also provide a rich corpus for research in sentence repair phenomena.

## 4 ALIGNMENT AND SCORING

The scoring of an extraction system response against the key presents a different set of problems, since the scoring procedure must allow for differences not only in tags but also in the underlying word stream. A system-produced transcription is adversely affected by such factors as noise in the production, transmission and capture of the audio signal and an errorful transcription process. Example 5 shows a real example of such a transcription, automatically produced from the audio signal corresponding to Example 1 by an ASR system, then tagged for Named Entity by MITRE's Mesa-Phrag system[7].

We have previously reported on MITRE's mscore program[1], which makes use of an alignment procedure[2] to mediate between system transcription and reference transcription. This alignment step is crucial in such an evaluation, because of the differences in the underlying text stream just discussed. Figure 2 shows a portion of the reference and system transcriptions, aligned phonetically. The transcription errors made by the ASR system are underlined. This alignment enables mscore to determine how to compare reference named entities and hypothesis (system) named entities.

Mscore compares the extracted information in the aligned data in three dimensions: extent (locating the correct region), type (assigning the right class to the extracted elements) and content (getting the underlying words right). For audio data, it is important to separate a content score (closely related to word error) from scores for extent and type. This permits a system to get credit for identifying a region in which there is something interesting (e.g., a name), even where the speech recognizer mis-transcribes the region. For example, we can see that the extent of the named entity C\_N\_N CENTER is incorrect, although its type, an ORG, is correctly annotated. Separating the type, extent and content measures supports research in audio browsing (can the system identify the key region to listen to?) and it supports research in use of prosodic information independent of transcription accuracy. Mscore was rewritten by SAIC and used to evaluate Named Entity content extraction at the 1999 DARPA Broadcast News workshop.

As an additional evaluation tool for content extraction, we have also developed a modification of mscore that computes a "targeted" word error rate (TWER) metric. The TWER is the

word error metric calculated only for regions in a document which contain certain types of target constructions, for example, named entities, or noun groups (and head nouns). The TWER provides another diagnostic tool to determine the contributions of speech recognizer characteristics, such as lexicon size and language model grammar, to content extraction performance. We initially introduced the TWER at the September 1998 DARPA Conversational Speech (Hub5) workshop, and it has since become part of the standard measurements computed for the Broadcast News Named Entity extraction task[13].

## 5 GENRE AND CONTENT ISSUES

A major factor in the evaluation and comparison of content extraction is the style of the written representation. This is an important consideration when porting existing techniques from journalistic text data to spoken language transcriptions; it is also important when comparing different spoken styles, such as broadcast news vs. conversational speech.

Content extraction from journalistic (newspaper and newswire) texts can be assisted greatly by the well-behaved nature of the data. For example, news texts conform to consistent punctuation and capitalization conventions, and these features can be useful both in segmentation and in classification. On the other hand, most speech recognition systems output an unformatted, single-case (all uppers case or all lower-case) stream of the words spoken, without punctuation. In addition, speech output contains disfluencies not present in news text, as discussed above.

Speech corpora can vary greatly in content. The North American Business News corpus consists of read Wall Street Journal data, so the content (and style) is consistent with most newspaper texts. Broadcast news contains similar content, in that it is primarily about news events, but the style contains multiple speakers and varying acoustic conditions. Conversational speech, on the other hand, is much less “fluent” and contains frequent topic shifts. In addition, conversational speech varies depending on the familiarity of the speakers, as can be seen in comparing the Switchboard corpus, in which strangers talk about a prescribed topic, and the Callhome (or Callfriend) corpus, in which friends and family members speak about any topic.

The density of the phenomena to be extracted affects both performance and evaluation of content extraction tasks, specifically for trainable extraction systems whose performance depends on the quantity of training examples available. For example, we have found a significant difference in the density of person, location, and organization phrases in a variety of corpora. We compared the density of phrases in journalistic text (Wall Street Journal from MUC) to three corpora of spoken data (Broadcast News from Hub-4, Callhome and Switchboard from Hub-5). Table 1 shows, for each corpus, the ratio of total words to total number of name phrases. The density of name phrases in Wall Street Journal texts can be seen to be significantly

higher than any of the speech corpora, but there is also a significant variation between speech corpora.

|                |       |
|----------------|-------|
| Switchboard    | 109.3 |
| Callhome       | 67.8  |
| Broadcast News | 24.6  |
| WSJ            | 13.9  |

Table 1

In order to compare extraction systems fairly across text genres, the density differential has two practical consequences. First, for comparable amounts of training data to be available to trainable systems, more words of spoken language data must be annotated. Similarly, in order to evaluate systems on similar-sized test sets, larger test sets are required for spoken data.

Furthermore, the type of data in a corpus is a significant factor in assessing the appropriateness of a task. For example, we have found Named Entities to contain significantly more information in broadcast news data than in conversational data, such as the Switchboard corpus.

## 6 NEW EVALUATIONS

The previous sections have discussed the adaptation of written text content extraction metrics to audio sources. However, there are also several proposed evaluations that have been developed specifically with audio sources in mind. These are briefly described in this section.

### 6.1 Event Extraction

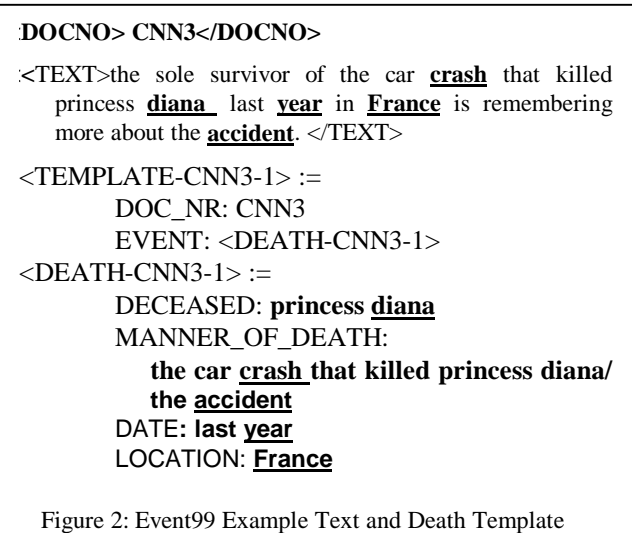
As discussed in section 2, the early MUC evaluations defined a high-level information extraction task based on filling a domain-specific *template*. However the definition of a task-specific template was complex, and the scoring of the templates required a lengthy document specifying the legal fills for each template slot.

As a result, there was a demand for a new domain-independent event-level evaluation that could accommodate audio sources. This gave rise the Event99 evaluation proposed at the March 1999 DARPA Broadcast News Workshop[3]<sup>1</sup>. The goal of this task was to evaluate event identification, based on filling a shallow template that included pointers back into sources, including news wire, radio, and television. Event99 builds on several evaluation tasks, including the text-based MUC information extraction tasks, the DARPA Broadcast News Named Entity extraction task, and the Tipster Summarization (SUMMAC) evaluation.

<sup>3</sup> The Event99 task was developed by a committee consisting of participants from MITRE (L. Hirschman, P. Robinson, L. Ferro), SAIC (N. Chinchor, E. Brown, A. Douthat), SPAWAR Systems Center San Diego (B. Sundheim), NYU (R. Grishman), chaired by MITRE.

The Event99 evaluation defines shallow templates for certain classes of “interesting” events, such as natural disasters, deaths, or financial fluctuations. Each type of template has a small number of slots for the event, typically agent, object, effect, time and location of the event (see Figure 2). This event extraction task differs from event-level extraction in MUC because the templates are shallow and it is easy to define new event types; also the Event99 task focuses on indexing into the audio to support audio browsing, whereas the MUC extraction task was focused on database creation. It uses a set of general guidelines that specify event-independent rules for filling templates, and limits event-specific rules to a minimum.

The current guidelines include sample event definitions for death and natural disaster events, with other event types under development.



The “answer key” prepared by human annotators includes alternative fills (paraphrases or casually related events) that occur in the story. The scoring algorithm is designed to give credit for filling a slot correctly if the system identifies any one of the alternative fills for the slots. The task is primarily designed to support event-based browsing and search, but the human-prepared key also provides a kind of topic-specific summarization. In addition, to allow for some leeway in what constitutes the correct fill for a slot, minimal and maximal extents are defined for each slot fill. A system receives credit for identifying any region that overlaps with the minimal fill and does not exceed the maximal fill. Slot fills may consist of names, nouns, noun phrases, prepositional phrases, adjective phrases, verbal forms, or clauses. Typical minimal/maximal fill pairs might be the head noun of a phrase and the associated

maximal noun phrase, or the content-bearing verb of a clause and the associated clause. To support annotation and evaluation of event indexing, we have prepared a set of general guidelines for event mark-up and have tested these guidelines for interannotator agreement.

## 6.2 Standardized Comprehension Tests

We have also been exploring a different approach to evaluation using “found” (pre-existing) test material. Standardized reading and listening comprehension tests are good candidates; they range in level from tests for 7-8 year olds to tests for the post-bachelor's level, e.g., Graduate Record Examination (GRE) or Testing of English as a Foreign Language (TOEFL) examination[12]. They also come with human benchmarks, so that they provide some easily understood metrics – for example, if a system were to pass such a test, it might be certified to understand at the level of an average 8 year old.

Typically, such tests ask the student to read or listen to a story or article and to demonstrate her/his understanding of that article by answering questions about it. These tests may take a number of forms, including multiple choice and short answer tests, such as the listening test item shown in Example 6.

For these reasons, standardized comprehension tests offer an interesting alternative to the kinds of carefully constructed, special-purpose evaluations described so far. In particular, standardized comprehension tests are domain-independent, they assume an age-appropriate command of the language, and also of world knowledge and subject matter needed to answer the questions. To determine whether these tests might serve as a reasonable evaluation for content extraction, we have done an initial pilot study on written materials[5]. We identified a small corpus of training and test materials (some hundred stories) consisting of remedial reading materials for 7 to 10 year olds. These materials are simulated news stories, followed by short answer who, what, when, where, and why questions.

Using these materials, we developed a simple system that uses limited linguistic processing and pattern matching techniques to select the sentence from the text that best answers the query. This has served as our baseline system and has allowed us to assess the difficulty of the overall problem. The baseline system is able to identify the sentence containing the answer to a given question about 30% of the time. With the addition of a named entity detection algorithm (for person, location and time), it improves to 37%. We plan to build on this modular architecture to incorporate incremental addition and testing of new sources of linguistic and world knowledge.

S1: Want to go to the library now, Betty?  
S2: How about in an hour or so?

Question: What does the second speaker mean?

Example 6: Sample Question from a TOEFL  
Listening Test

We are now exploring how to assemble significant training and test sets. We will be talking to some providers of educational test materials in the near future. Also, a Web search has turned up several daily on-line news quizzes that could serve as training materials. We also are extending the linguistic capabilities of the system and adapting it to take standardized multiple choice tests. We believe that our preliminary work has established that this problem is tractable, challenging, and readily evaluated using automated techniques.

## 7 REFERENCES

- [1] Burger, J., Palmer, D., and Hirschman, L. "Named Entity Scoring for Speech Input," *COLING-98*, Montreal, 1998.
- [2] Fisher, W. and Fiscus, J. "Better Alignment Procedures for Speech Recognition Evaluation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1993, Vol. II.
- [3] Hirschman, L., Brown, E., Chinchor, N., Ferro, L. Grishman, R., Robinson, P., and Sundheim, B. "EVENT99: A Proposed Event Indexing Task for Broadcast News," *Proc. of the DARPA Broadcast News Workshop*, February 1999.
- [4] Hirschman, L., "The Evolution of Evaluation: Lessons from the Message Understanding Conferences," *Computer Speech and Language* 12, pp. 281—305, 1998.
- [5] Hirschman, L., Light, M., and Breck, E., "Deep Read: A Reading Comprehension System," submitted for publication in *Proc. of the 37<sup>th</sup> Annual Meeting of the Assoc. for Computational Linguistics*, College Park, MD, June 1999.
- [6] Kubala, F., Schwartz, R., Stone, R., and Weischedel, R., "Named Entity Extraction from Speech," *Proc. of the Broadcast News Transcription and Understanding Workshop*, pp. 287—292, Morgan Kaufmann Publishers, San Francisco, CA, February 8-11, 1998.
- [7] Palmer, D., Burger, J., and Ostendorf, M., "Phrase Language Models for Named Entity Tagging," to appear in *Proc. of the Broadcast News Transcription and Understanding Workshop*, February 28 – March 3, 1999.
- [8] *Proc. of the Broadcast News Transcription and Understanding Workshop*, pp. 287—292, Morgan Kaufmann Publishers, San Francisco, CA, February 8-11, 1998.
- [9] *Proc. of the Broadcast News Transcription and Understanding Workshop*, February 28 – March 3, 1999.
- [10] *Proc. of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann Publishers, San Francisco, CA, November 1995.
- [11] *Proc. of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, April 29–May 1, 1998. <http://www.muc.saic.com>.
- [12] Testing of English as a Foreign Language Sample Test. Available from Educational Testing Service Web site at <http://www.toefl.org>.
- [13] Przybocki, M., Pallett, Fiscus, J., Garofolo, J., "1998 Broadcast News Evaluation Information Extraction Named Entities," to appear in *Proc. of the Broadcast News Transcription and Understanding Workshop*, February 28 – March 3, 1999.s