

EVENT99: A PROPOSED EVENT INDEXING TASK FOR BROADCAST NEWS

Lynette Hirschman (1), Erica Brown (2), Nancy Chinchor (2), Aaron Douthat(2),
Lisa Ferro(1), Ralph Grishman (3), Patricia Robinson(1), Beth Sundheim (4)

- (1) The MITRE Corporation
202 Burlington Rd.
Bedford, Massachusetts 01730
- (2) Science Applications International Corporation
10260 Campus Pt. Dr. M/S A2-F
San Diego, California 92121
- (3) New York University
715 Broadway, 7th Floor
New York, New York 10003
- (4) SPAWAR Systems Center, Code D44208
53140 Gatchell Road
San Diego, California 92152-7420

ABSTRACT

The goal of the proposed Event99 task is to evaluate event-level indexing into news stories, including news wire, radio, and television sources. The Event99 task is distinguished from earlier, related evaluations in its focus on indexing, its application to multiple media, and the relative extensibility of its task definition to new event types. The task involves identifying instances of high-interest event types, along with some basic attributes of the events (e.g., actors, effects, date, location). The current guidelines include sample event definitions for death, natural disaster and bombing events, with other event types under development. The output from 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. This paper provides background concerning the task design, an overview of the definition, annotation and scoring of the task, results of an interannotator agreement study, and information concerning current status and schedule.

1. BACKGROUND

The goal of the proposed Event99 task is to evaluate event-level indexing into news stories, including news wire, radio, and television sources. Event99 builds on the experience of several evaluation tasks, including the text-based Message Understanding Conference (MUC) information extraction tasks[1,2], the DARPA Broadcast News Named Entity extraction task[3], and the Tipster Summarization (SUMMAC) evaluation[4].

The MUC evaluations had defined a high-level information extraction task (the *scenario template* task) oriented towards creating a database of events. This required a separate domain-specific, multi-level template for each area of interest, e.g., terrorist events, or management succession, or airplane crashes. Creating a template definition was complex because of the need to anticipate the types of slots (data attributes) to accommodate the many kinds of information of interest. Automated scoring required the creation of a human-generated “gold standard” (answer key). Thus for each template definition, a set of “fill rules” had to be created, describing the legal fills for each slot. The gold standard was then used to score the templates generated by the extraction systems under evaluation. The high degree of domain-specificity and complexity of the MUC templates limited the scope of the event evaluations and required significant overhead to acquire the human expertise for each new domain. Also, the MUC evaluations were focused exclusively on newswire (text) sources.

The Broadcast News Named Entity extraction task is derived from the MUC Named Entity task, generalized to apply to broadcast news (audio) sources. This task requires in-line tagging of certain types of names, temporal references and numeric items found in transcripts of broadcast news stories. This is a low-level type of information extraction task that is viewed as a largely domain-independent element of more ambitious information extraction tasks such as the scenario template task. Extensive task guidelines and manually-produced answer keys are still required, but there is less effort involved on the part of both the evaluation designers and the participating named entity system developers.

One of the SUMMAC evaluation tasks, the Question-and-Answer (Q&A) task, exhibited characteristics of each of the above types of evaluation. For example, it was like the Named Entity task in that the answer key was represented as in-line tags on the source document, and it was like the scenario template task in that it was domain-dependent. However, unlike both tasks, there were no extensive guidelines -- it was felt that the nature of summarization made it impossible to define what should constitute a single "gold standard." The summarization systems were expected to produce summaries that were succinct and that contained enough extracts from the source document to be fully "informative" about a particular topic description that was input to the system. Scoring was done manually, because there were scoring criteria that could not be fully captured by the answer key notation.

The Event99 committee was formed to design an evaluation that could accommodate audio sources and that would respond to two competing trends:

1. The emerging interest within the spoken-language community to take on challenges that require greater degrees of language understanding;
2. The growing desire in both the written- and spoken-language communities to minimize the domain dependence of the tasks so that evaluations would be easy (and inexpensive) to implement and extend.

The Event99 Committee was formed in March 1998 and met regularly over the year to define the Event99 evaluation proposal for discussion at the March 1999 DARPA Broadcast News Workshop.

2. THE EVENT99 TASK

The Event99 evaluation task is known informally as the "templette" task. It uses a set of general guidelines that provides event-independent rules for generating the output (called an event report), to minimize the number of event-specific rules that must be defined. The typical event report structure includes slots for the main event, actors and effects, as well as associated date and location information. The date and location slots are common across all event types. The current guidelines include sample event definitions for death, bombing and natural disaster events, with others under development. Subsuming the event report structure is the "template" structure, which serves to group relevant events together that are reported in a single story. (For example, a single news story could provide news on both a natural disaster event and a death event.)

The general form of the output of this task is a simple, two-level structure, with the template-level structure pointing to one or more templette-level, event report structures. The template contains slots for identifying the news story and any reportable events that come from that story.

Although the output format is in the form of a template, the Event99 task focuses on *indexing* into the underlying sources, as opposed to *extracting* from them. The templette fills consist of excerpts (strings) from the text with pointers back into the text. This has several advantages. It supports browsing and machine learning by hyperlinking the filled slots to the source passages. It

also makes it possible to score templates created from automatically transcribed audio, where the automatic transcription (and thus the template fills) may differ from the "truth". This is similar to the procedure for Named Entity scoring and the calculation of Targeted Word Error rates used in Hub4[3].

Figure 1 shows a short text and the associated answer key template. Bold face has been added to highlight the event slot fills in the template. The square brackets indicate a *minimal* slot fill, i.e., the shortest fill that qualifies as a correct fill for the slot. An entire, bold-face phrase is a *maximal* slot fill, i.e., the longest fill that qualifies as correct. A fill produced by a system is scored as correct as long as its fill includes the minimal fill from the key and does not contain anything beyond what's in the maximal fill in the key. The task guidelines define the grammatical classes or syntactic phrase types that are expected as maximal fills for the various slots, and they also define what constitutes the minimal fill for each kind of maximal fill.

```

<DOCNO> CNN3</DOCNO>

<TEXT>the sole survivor of the car crash that killed princess
diana and dodi fayed last year in France is remembering
more about the accident. </TEXT>

<TEMPLATE-CNN3-1> :=
    DOC_NR: CNN3
    EVENT: <DEATH-CNN3-1>
<DEATH-CNN3-1> :=
    DECEASED: princess [diana]
                / [dodi fayed]
    MANNER_OF_DEATH: the car [crash]
                    that killed princess diana and dodi
                    fayed
                    / the [accident]
    DATE: last [year]

```

Figure 1

Two of the fills in the example are preceded by a forward slash. The forward slash in the answer key indicates an *alternative* fill; the system-generated fill is expected to be any one of the listed alternatives. Alternatives include all coreferential phrases (as in the MANNER_OF_DEATH slot in the example) and all distinct, non-coreferential fills for a slot (as in the DECEASED slot in the example). Note that there is no notation to indicate whether one alternative fill is in any sense "better" than another. 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. Scoring of the system output based on time alignment as well as on content alignment is addressed in the next section.

3. SCORING AND ALIGNMENT

Scores will be assigned to a system's output based on how well it agrees with the human-generated key. Measuring this agreement involves:

1. Normalizing indices into the underlying sources.
2. Mapping the system output data with the key data.
3. Tallying the number of correct, incorrect, missing, and spurious slot fills.
4. Calculating various metrics based on the above tallies.

Index normalization is required before indices in slot fills may be compared. To refer to a point in a text, *byte offsets* are often used. These are usually the numbers of characters between the start of a story and the point in question. With byte offsets, as long as both the system and the annotators are reading from the same text, it is easy to determine whether they are referring to the same point in the data stream: simply compare the character counts. However, when the texts differ due to transcription errors, byte offset indices are not comparable. They must be transformed, or normalized, before the comparison takes place.

The index normalization in the Hub 4 Named Entity Task used *content alignment* as a way to make indices comparable [5]. With this method, the underlying texts were aligned using dynamic programming techniques [7,8]. Byte offsets were replaced with counts of the points in the texts that were deemed to be “in the same place” by the alignment program. Current plans are to use this normalization method in the template task also.

It is possible that when audio transcripts contain enough timing information (usually in the form of timestamp-containing SGML tags present in the text [9]), byte offset indices could be converted to time indices, which could then be compared easily without a dynamic programming step. This *time alignment* method will be experimented with in the template task.

After index normalization, mapping of the system data to the key data is the next step in the scoring. This mapping will take place at two levels. At the slot level, one alternative fill from a key slot is paired with the one fill from the corresponding system output slot. At the event report level, for each event type in a story, the set of event reports in the key is matched with the set of event reports in the system output. Matching at the event report level will be so that the slot error (see below) is minimized.

Once mapping key and system data is completed, tallies of correct, incorrect, missing and spurious slot fills are made. If two aligned fills are judged to point to the same place in the data source, one *correct* point is counted. If the two aligned fills point to different places, one *incorrect* point is counted. If a key slot was not aligned with a response slot, one *missing* point is counted, and if a response slot was not aligned with any key slot, a *spurious* point is counted.

After the above four counts are determined, several metrics are calculated from them. Let **C** be the number of correct points, **I** the number incorrect, **M** the number missing, and **S** the number spurious. Further, let **K** be the total number of slots in the key, and **H** the total number of slots in the system output. Then

the *slot-error E* is given by [6]

$$E = (I+M+S)/K,$$

the *recall, R*, is calculated as

$$R = C/K,$$

the *precision, P*, is

$$P = C/H,$$

and the *F-measure, F*, is

$$F = 2C/(K + H).$$

(**F** is used to combine **P** and **R** into one measure [10]. Its full definition includes another parameter, **α**, used to give different weights to the **P** and **R** values. The above equation is a simplification of the case when **P** and **R** are weighted equally.)

4. INTERANNOTATOR AGREEMENT

We have completed preliminary interannotator agreement experiments, which tested both the Event99 general guidelines and the guidelines that pertain specifically to the NATURAL_DISASTER event type. The human-generated answer keys for natural disasters reported in the training and test data were produced by two annotators, one experienced (annotator 1) and one less experienced (annotator 2). The work was done independently by the annotators in three stages: relevancy judgments, maximal slot fills, and minimal slot fills. To enter maximal slot fills, the annotators used the Tabula Rasa template tool from New Mexico State University's Computing Research Laboratory; for other aspects of the data preparation, they used a text editor (emacs).

Each of the annotators read the 482 stories in the training data and the 607 stories in the test data from five news sources to find all of the relevant natural disasters reported. After reconciling the differences in judgments, the following interannotator relevancy agreement scores were calculated in the strictest way by hand:

Annotator	Recall		Precision	
	Train	Test	Train	Test
Annotator 1	96.2	100	89.3	93.3
Annotator 2	100	89.3	89.6	96.2

The two annotators then filled templates for the relevant segments with maximal slot fills. The maximal slot fills were then reconciled and a standard set of maximals was produced. The annotators' independent fills were scored against the reconciled set and received the following scores:

Annot	Recall		Precision		F-Measure	
	Train	Test	Train	Test	Train	Test
Annot 1	83	84	83	86	83.0	84.9
Annot2	59	66	61	71	60.1	68.1

The two annotators then independently chose the minimal from the standard sets of maximals and reconciled those to form a standard set of minimal slot fills. Their independent scores against these were as follows:

Annot	Recall		Precision		F-Measure	
	Train	Test	Train	Test	Train	Test
Annot1	98	95	98	95	98.0	94.7
Annot2	81	88	80	88	80.5	88.1

These results are reassuring in that one annotator (Annot1) achieved more than 80% on all aspects of the hardest part of the task, determining the maximal fills. Eighty percent has come to be regarded by the designers of various DARPA evaluation tasks as a good indication that a task has become well enough defined to be usable in an actual evaluation. Of course, the claim to validity of the Event99 task would be much stronger if *both* annotators had exceeded the 80% threshold.

The results also distinguish the experienced annotator from the less experienced annotator for both training and test data. The less experienced annotator clearly gained experience during the exercise. Finally, the results show the relative difficulty of the stages of answer key generation. The datasets (training and test) were unequal in difficulty, but the relative difficulty of the stages still remains the same across the two datasets. The test corpus was larger and had more cases not covered in the guidelines or the previous case history from the training data and other smaller datasets. However, the annotators carried out the answer key generation in the same amount of time, that is, one week each. Better annotation tools could improve the rates of interannotator agreement (especially in terms of determining maximal and minimal slot fills) and annotator productivity.

5. STATUS AND DIRECTIONS

To date, three templette types (NATURAL_DISASTER, DEATH, BOMBING) have been carefully defined and largely "debugged" through rounds of annotation by committee members. Several other templette definitions have been drafted and are in various stages of review. SAIC is in the process of adapting and extending existing alignment and scoring software to accommodate the special features of the Event99 task. Committee membership has been extended to include representatives from both BBN and NIST in order to gear up for conducting a multi-site trial of the evaluation. In addition, negotiations are underway for support from the Linguistic Data Consortium in the areas of data collection and event-relevance assessment.

Current plans call for a trial evaluation in December 1999, followed by full scale event level evaluation in the fall of 2000.

REFERENCES

1. *Proc. of the Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann Publishers, San Francisco, CA, November 1995.
2. *Proc. of the Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA, April 29–May 1, 1998. Available at <http://www.muc.saic.com>.
3. Chinchor, N., Brown, E., Robinson, P. (1998), "HUB-4 Named Entity Task Definition (version 4.8)". Available by ftp from www.nist.gov/speech/hub4_98.
4. Fisher, W.M., Fisher, J., Martin, A., Pallett, D.S., and Przybocki, M.A., "Further Studies in Phonological Scoring," Proceedings of the Spoken Language Systems Technology Workshop, January 22-25, 1995, Austin, TX (sponsored by ARPA), Morgan Kaufmann Publishers, San Francisco, ISBN 1-55860-374-3, pp. 181-186.
5. Makhoul, J., Kubala, F., Schwartz, F., Weischedel, R., "Performance Measures for Information Extraction," these proceedings.
6. Mani, I., Firmin, T., House, D., Klein, G., Sundheim, B., and Hirschman, L. (1998), "The TIPSTER SUMMAC Text Summarization Evaluation", Proceedings of EACL'99, Bergen, Norway, June 8-12, 1999.
7. National Institute of Standards and Technology, "A Universal Transcription Format (UTF) Annotation Specification for Evaluation of Spoken Language Technology Corpora", available at ["http://www.nist.gov/speech/hub4_98/utf-1.0-v2.ps"](http://www.nist.gov/speech/hub4_98/utf-1.0-v2.ps).
8. Robinson, P., Brown, E., Burger, J., Chinchor, N., Douthat, A., Ferro, L., Hirschman, L., "Overview: Information Extraction from Broadcast News," these proceedings.
9. Sankoff D. and Kruskal J. (eds.), "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison," Addison-Wesley, Reading, MA, 1983.
10. C. J. van Rijsbergen, "Information Retrieval", London: Butterworth, 1979.