Event Nugget Detection and Coreference Scoring

May 4, 2015
Language Technologies Institute
Carnegie Mellon University

1 Overall workflow

This document describes the event nugget and coreference evaluation. We show an overall workflow of evaluation in Figure 1.

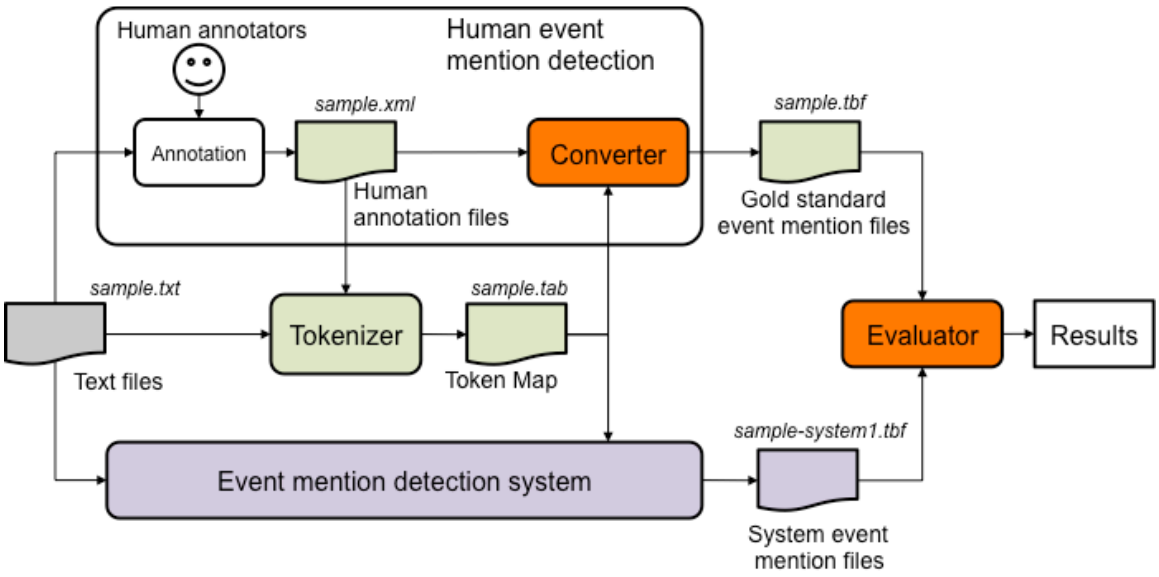

Figure 1: An overall workflow of evaluation for event nugget detection.

We first give an overview of the whole scoring process. For detailed format description please refer to the corresponding sections.

For each text file, LDC annotators provide human annotations as gold standard files. System submissions will be evaluated against certain gold standard based on the task it participates. Because existing evaluation for event nugget detection evaluation and event coreference evaluation all score systems on token level, we will conduct a post-tokenization step after the annotations. The post-tokenization step will create a token mapping file, which provide token offset and token id information.

Each participant event nugget detection system will be given two files as input: (1) the source text and (2) a token mapping table. The latter specifies token ID for every token. The ID information is used in the output of the system. Let us refer to the output of the system as a system event nugget file. We require a system event

nugget file to be given in the same file format as the gold standard file. The evaluator (scorer) takes the gold standard file, the system event nugget file, and the token mapping file as input, and compares them to give a score for the system.

2 Evaluation Process and Formats

The submission format described below will be used for both Event nugget Detection task and Event Coreference task.

Input of Scorer:

1. Gold standard annotation for documents, in Token Based Format (tbf).
2. System output annotation for documents submitted by participants, in Token Based Format (tbf)
3. Tokenization mapping files associated with each document, "tab" extension is appended to the file extension of its corresponding source filename.

Output of Scorer:

1. Overall performance report for system, as described in "Scoring" section.
2. Visualization in text form or html form if the corresponding arguments are specified. (This is not the core functionality for scoring, please refer to the README file of the scorer for details)

Evaluation Script Options:

Please follow the README and help information in the script distribution.

2.1 Formats

2.1.1 Tokenization file format

Standard token mapping files (.tab) will be provided to participants. These are tab-delimited mapping files for evaluation purpose. These files map the tokens to their offsets [1] in the source files. "tab" is appended to the file extension of its corresponding source filename. Participants should report their system output in terms of the token id provided in the token mapping files.

A mapping table contains 4 columns for each row, and the rows contain an ordered listing of the document's tokens. The columns are:

- token_id: A string of "t" followed by a token-number beginning at 0
- token_str: The literal string of a given-token
- tkn_begin: Index of the token's first character in the source file
- tkn_end: Index of the token's last character in the source file

2.1.2 The submission format

There are 3 possible tasks for evaluation this year, the submission format of these tasks are the same. The content of the submission may vary:

[1] Offsets are character offsets where the first character is 0

1. Event Nugget Evaluation: Only contain the event nugget annotations by the systems
2. Event Nugget Detection and Coreference: Contain both event nugget annotations and coreference by the systems
3. Event Nugget Coreference: Contain both event nugget annotations and coreference, **where event nugget annotations should be copied directly from provided event nugget files, coreference annotations are generated by the systems**

The following sections describe the format in details.

2.1.3 System and gold standard annotation file format:

1. All event nugget annotations (and/or coreference annotation) for all documents in the corpus are written into one single file
2. A header will indicate the start of a new document (<s> is the space character)
    a. Header := #BeginOfDocument<s><doc ID>
3. A footer will indicate the end of a document
    a. Footer := #EndOfDocument

2.1.4 Definition of event nugget format:

Event nuggets are represented as tab-separated lines. To be specific, for each mention line, we have the following columns:

- <system ID> := the name of the system
- <doc ID> := the ID of the input document
- <mention ID> := the ID of the mention, which should uniquely identify the mention within the current document
- <token ID list> := list of IDs for the token(s) of the current mention, in ascending order, separated by commas (,) . Each ID is a string of "t" followed by a token-number beginning at 0, the same as how they appear in the tokenization files
- <mention> := the actual character string of the mention
- <event-type> := the ACE hierarchy type
- <realis status> := the REALIS label

It is optional to append the following confidence columns:

- <span confidence> := a score (confidence, etc.) the system wants to assign for the mention span detection. This score will not affect the evaluation results
- <type confidence> := a score (confidence, etc.) the system wants to assign for the mention type detection. This score will not affect the evaluation results
- <realis confidence>:= a score (confidence, etc.) the system wants to assign for the mention realis detection. This score will not affect the evaluation results

2.1.5 Definition of event coreference format:

All coreference clusters annotations are **appended after all event nuggets lines of this file, before the #EndOfDocument footer**. System submissions should make sure transitive closure of coreference are resolved so that each coreference line indicate the whole cluster.

Each coreference cluster should also be represented as a tab-separated line, with the following columns:

- <relation name> := The relation name with a special indicator character (@) as prefix , in our case, it is always "@Coreference"
- <relation id> := A relation id. This is for bookkeeping purposes, which will not be read by the scorer. The relation id used in the gold standard files will be in form of "R<id>" (e.g. R3). However, system should always give a non-empty string without whitespace here. We recommend to put the cluster id as a placeholder.
- <event mentions> := A list of event mentions in this coreference cluster, separated by comma (,). In terms of coreference, the ordering of event mentions does not matter.

Example:

```
#BeginOfDocument sample
system1	sample	E2	t1069	married	Life_Marry	Actual 1 1 1
system1	sample	E4	t1096	divorced	Life_Divorce	Actual 1 1 1
system1	sample	E5	t1109	married	Life_Marry	Actual 1 1 1
system1	sample	E6	t1157	married	Life_Marry	Actual 1 1 1
@Coreference	R1	E6,E2
#EndOfDocument
```

3 Event Nugget Detection Scoring

In this section we describe the algorithm for scoring event nugget detection.

3.1 Scoring for one document

We denote a gold standard mention with **G**, and a system mention with **S**. We use $\mathbf{T_S}$ to represent the tokens of the mention. $\mathbf{Dice(T_G,T_S)}$ is the token-based dice coefficient function that returns a score between 0 and 1 (All invisible words are already removed from $\mathbf{T_G}$ and $\mathbf{T_S}$ [2]).

[2] Invisible words are ignored in scoring. They include: determiners {the, a, an}, pronouns {I, you, he, she, we, my, your, her, our}, relative pronouns {who, what, where, when}.
Note that "it" and "that" and pronouns including {his, ours, mine, yours, ours, they} are not included in the invisibles list because they can occasionally be resolved as nominal event nuggets.

3.1.1 Create a mapping

To perform scoring for a document, system mentions are mapped to gold standard mentions based on the Overlap score. A system mention is always mapped to one and only one gold standard mention with which the system mention will have the highest overlap score. However, one gold standard mention can be mapped to multiple system mentions.

This can be described using the following pseudo-code:

Input: A list L of scores $Dice(T_G, T_S)$ for all pair of G, S in the document

$M \leftarrow \varnothing$; $U \leftarrow \varnothing$
while $L \neq \varnothing$ **do**
  $G_m, S_n \leftarrow \arg\max_{(G,S)\in L} Dice(T_G, T_S)$
  if $S_n \notin U$ **and** $Dice(T_{G_m}, T_{S_n}) > 0$ **then**
    $M_{G_m} \leftarrow M_{G_m} \cup (S_n, Dice(T_{G_m}, T_{S_n}))$
    $U \leftarrow U \cup \{S_n\}$

Output: The mapping M

3.1.2 Scoring mention detection

To score mention detection, a mention-based F1 score is computed in the following way:

1. For each gold standard mention G, recall that G can be mapped to multiple system mentions, we only choose one system mention S that maximize $Dice\ (G, S)$, and denote $TP_i = \max Dice\ (G, S)$.
2. True Positive = $\sum_i TP_i$
3. Precision = True Positive / #System Mention
4. Recall = True Positive / #Gold Mention
5. F1 = H (Precision, Recall), where H is the harmonic average function

The following pseudo code shows how to calculate the True Positive:

Input: The set of gold standard $\mathcal{G}$; The mapping M indexed by G; Number of system mentions N_S

$TP \leftarrow 0$; $FP \leftarrow 0$
for $\forall G \in \mathcal{G}$ **do**
  $S_T \leftarrow \arg\max_{Dice}(S, Dice) \in M_G$
  $TP \leftarrow TP + Dice(G, S_T)$

Output: TP

We can then simply compute F1 score with the following:

$$P = \frac{TP}{N_s}\ ;\ R = \frac{TP}{N_G}; F1 = \frac{2PR}{(P+R)}$$

3.1.3 Scoring realis status and mention type detection

To score realis status and mention type detection, we augment the Span-based F-1 score slightly. The only difference is that we will choose one system mention S that have the correct attribute according the gold standard to maximize $Dice\ (G, S)$. In the following pseudo-code: we use **A** to denote the attributes of the mention. Note that we can alter **A** to be any attribute sets we are interested in. For the final comparison of the systems, **A** will be the set of mention type and realis status.

Input: The set of gold standard mentions $\mathcal{G}$; The mapping M indexed by gold standard mentions; Number of system mentions N_S; The set $\mathcal{A}$ indexing the attributes that will be evaluated for all mentions

1: $TP \leftarrow 0$

2: **for** $G \in \mathcal{G}$ **do**

3: $S_{max} \leftarrow \arg\max_{Dice}(S, Dice) \in M_G$ Subject to $\mathcal{A}_{S_{max}} = \mathcal{A}_G$

4: $TP \leftarrow TP + Dice(S_{max}, G)$

Output: TP

We can then calculate the attribute augmented F1 the same as above.

3.2 Summarization score

After all documents are scored, we also report scores that give a summary of performance over the whole corpus by taking the average across documents. We use the standard Micro and Macro average definition, as listed below:

3.2.1 Macro Average Scores (numerical average over the document scores):

Precision_macro = sum of all Precision / #document
Recall_macro = sum of all Recall / #document
F1_macro = 2* Precision_macro * Recall_macro / (Precision_macro + Recall_macro)

3.2.2 Micro Average Scores (sum of the individual true positives, false positives, and false negatives of each mention to calculate the overall F-Score)

Precision_micro = (sum of TP on all docs)/ (sum of TP on all docs + sum of FP on all docs)
Recall_micro = (sum of TP on all docs) / (total number of gold standard mention in all docs)
F1_micro = 2* Precision_ micro * Recall_ micro / (Precision_ micro + Recall_ micro)

3.2.3 Note on "invisible words":

Consider the maximum extent of an event nugget, but don't worry about determiners (they are invisible)

- [takes a shower] ==> it is okay for annotators to include "a" in their annotation; we ignore "a" for evaluation
- [make a quick decision] ==> it is okay for annotators to annotate the whole phrase; we ignore "a" and include "quick" in the evaluation

4 Event Coreference Scoring

Our evaluation script will simply call the Reference Implementation of Coreference Scoring algorithms used in CoNLL shared tasks. The script will convert both gold standard and system results to the required format of the CoNLL scorer. We will use the latest version of the scorer[3]. At the time of this writing, the scorer version is v8.01.

Systems will be ranked using the unweighted average of the following 4 metrics produced by CoNLL scorer:

1. MUC
2. BCUBED
3. CEAFE (entity based CEAF)
4. BLANC

Note that following the CoNLL evaluation, we choose to use CEAFE, and do not include CEAFM (mention based CEAF) in our final scores.

[3] http://conll.github.io/reference-coreference-scorers/