**Event Mention Detection scoring**

**Overall workflow**

We show an overall workflow of evaluation for event mention detection in Figure 1.

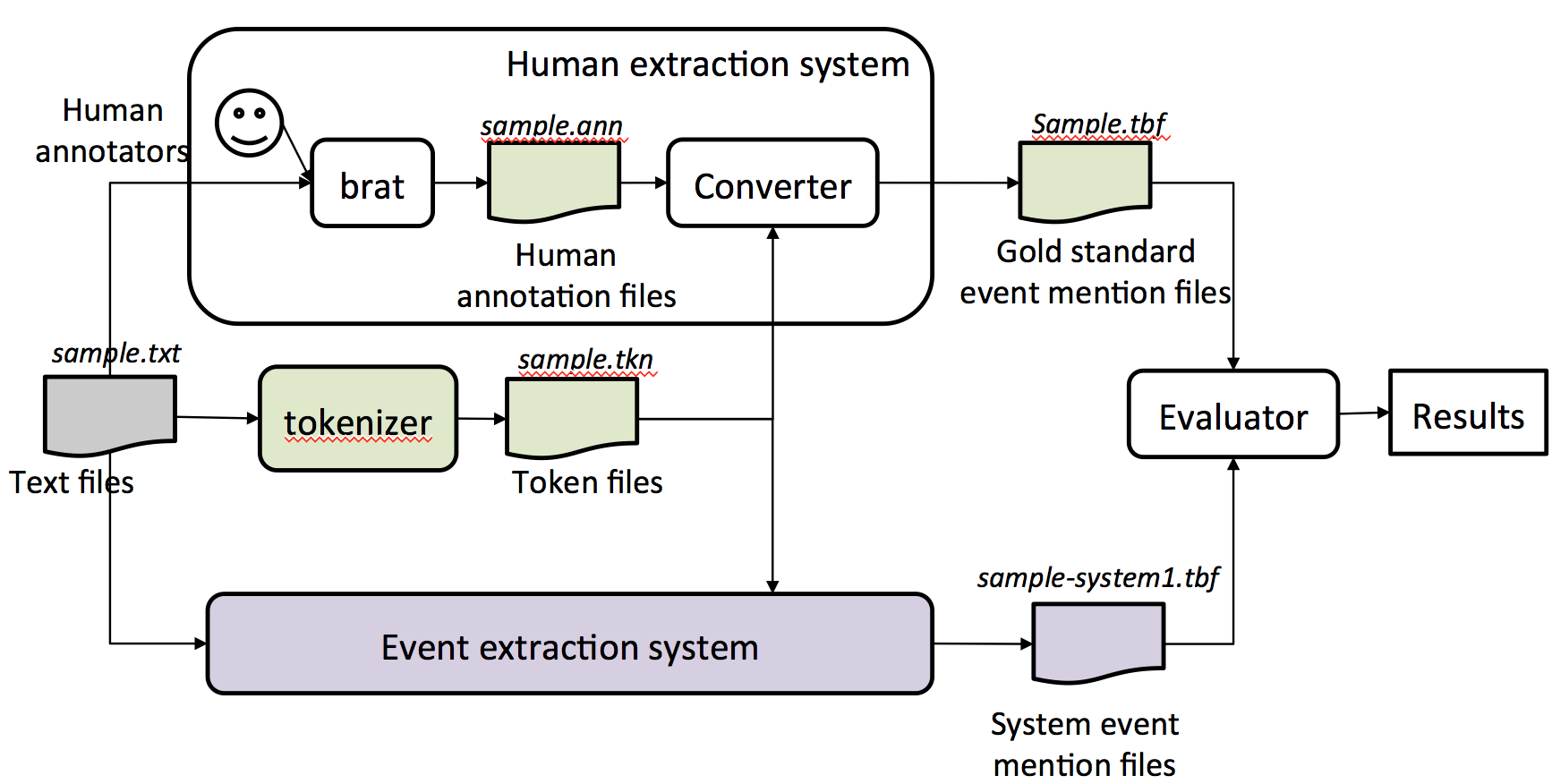


Figure 1: An overall workflow of event mention evaluation.

For each text file, human annotators use the **brat** rapid annotation tool to create a gold standard annotation file. We convert a brat annotation file to our evaluation file format. We assume event extraction systems output system event mentions in the same file format. The evaluator (scorer) then reads the output of event mention detection systems and compares them to the gold standard.

The scorer reads the output of event mention detection systems and compares them to the gold standard.

**Input of Scorer:**

1. Gold standard annotation for documents, in format (one line per mention), all annotations are contained in one file only.

2. System output annotation for documents submitted by participants, in format (one line per mention), all annotations are contained in one file only.

3. Tokenization files associated with each document, one file per document.

**Output of Scorer:**

1. System output annotation as item 2 in Input, with addition of a mention detection score, realis status detection and mention type detection score for each mention appended to each line.

2. Overall performance report for system, as described in “Scoring” section.

**System and gold standard annotation file format:**

1. All event mention annotations for all documents in the corpus are written into one single file
2. A header will indicate the start of a new document
   1. Header := #BeginOfDocument<s><doc ID>
3. A footer will indicate the end of a document
   1. Footer := #EndOfDocument
4. Different event mentions should not include the same token

For each mention line, we follow the following format,

**Definition of event mention format (one per line):**

event-mention := <system ID><TAB><doc ID><TAB><mention ID><TAB><token ID list><TAB> <mention><TAB><event-type><TAB><realis status><TAB><score1><TAB> <score2><TAB><score3>

**Explanation:**

<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 (,)

<mention> := the actual character string of the mention

<event-type> := the ACE hierarchy type

<realis status> := the REALIS label

<score1> := any score (confidence, etc.) the system wants to assign (ignored)

<score2> := score assigned in the evaluation

<score3> := additional possible score assigned by human

<TAB> := tab character

**Pseudo-code for Scoring one document:**

Let mappingScores = {}

#STEP 1 : Compute overlap scores for each pair of Gold/System Mention

FOR each system mention S := {S\_mid, S\_tokens, S\_realis, S\_type} (one per line)

Let S\_mid := mention id of S

Let S\_tokens := token IDs associated with S

Let S\_tokens := S\_tokens – {token IDs of invisible words} **#See NOTE 1**

Let S\_realis := realis status of S

Let S\_type := mention type of S

FOR each gold mention G:= {G\_mid, G\_tokens, G\_realis, G\_type}

Let G\_mid := mention id of G

Let G\_tokens := token IDs associated with G

Let G\_tokens := G\_tokens – {token IDs of invisible words}

Let G\_realis := realis status of G

Let G\_type := mention type of G

Let overlap := OVERLAP(S\_tokens, G\_tokens)

IF overlap > 0

mappingScores := mappingScores + (G, S, overlap)

END IF

END FOR

END FOR

#STEP2: After the calculation of all pairs, we can find the best mapping between

#System Mention and Gold Standard Mentions

Sort mappingScores based on overlap

Mapping = {} # create a empty mapping set to hold mappings

WHILE mappingScores != {}:

(G, S, overlap) = mappingScores.pop() #get the item with the highest overlap

#if G and S have not been mapped,

#it means there are no better overlap than this one

IF G has not been mapped and S has not been mapped

THEN Mapping := Mapping + {G,S, overlap}

END WHILE

#Append system score to the gold standard file

FOR each gold mention G:

Score := Mapping[G].overlap

append Score to the end of the line of G\_mid in Gold Standard, in position <score2>

END FOR

#STEP3.1: Compute document level errors and corrects on mention detection

TP := 0

FOR EACH System Mention S

IF S is contained in Mapping

TP := TP + Mapping[S].overlap

ELSE

FP := FP + 1

END IF

END FOR

#STEP3.2: Compute document level precision, recall for mention detection:

Precision := TP / (TP+FP)

Recall := TP / #GoldStandardMentions

F1\_Score := 2\*Precision\*Recall/(Precision+Recall)

#STEP3.3: Compute mention and realis type detection score:

num\_type\_correct := 0

num\_realis\_correct := 0

FOR EACH LINE (G,S, overlap) in Mapping

IF G\_type == S\_type

type\_correct := type\_correct +1

END IF

IF G\_realis == S\_realis

realis\_correct := realis\_correct +1

END IF

END FOR

Type\_detection\_accuracy := num\_type\_correct / #GoldStandardMentions

Realis\_detection\_accuracy:= num\_realis\_correct / #GoldStandardMentions

# Return and report the following measures for this document:

Measures for this doc = {TP, FP, num\_type\_correct, num\_realis\_correct, Precision, Recall, F1\_Score, Type\_detection\_accuracy, Realis\_detection\_accuracy } #**Note 2**

**Subroutine OVERLAP(G,S):**

IF G == S, THEN score := 1.0

IF G∧S == {}, THEN score := 0.0

ELSE

precision\_m := (|S∧G|)/|S|

recall\_m := (|S∧G|)/|G|

score := 2\*precision\_m\*recall\_m / (precision\_m + recall\_m)

RETURN score

End Subroutine

**Note 1**: Invisible words are ignored in scoring. They include: determiners {the, a, an}, pronouns {I, you, he, she, we, they, his, her, my, your, mine, yours, our, ours}, relative pronouns {who, what, where, when}, …?

Note that “it” and “that” are removed from the list because they can occasionally be resolved as nominal event mentions.

**Note 2:** After all documents are scored, we will also report a score that give a summarized performance on the whole corpus by taking the average across documents. We use the standard Micro and Macro average definition:

**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)

Type\_detection\_accuracy\_macro = # sum of all type\_detection\_ accuracy / # document

Realis\_detection\_ accuracy\_macro = # sum of all realis\_detection\_ accuracy / #document

**Micro Average Scores (Sum up the individual true positives, false positives, and false negatives of each mention and calculate the overall F-Score)**

Precision\_micro = (sum of TP on all doc )/ (sum of TP on all doc + sum of FP on all doc)

Recall\_micro = (sum of TP on all doc) / (#total number of gold standard mention in all documents)

F1\_micro = 2\* Precision\_ micro \* Recall\_ micro / (Precision\_ micro + Recall\_ micro)

Type\_detection\_accuracy\_micro = sum of num\_type\_correct / (#total number of gold standard mention in all documents)

Realis\_detection\_accuracy\_micro = sum of realis\_detection\_score / (#total number of gold standard mention in all documents)

Examples:

Rule 1: do not accept prepositions but include particles

* "[look] up a chimney" vs "[look up] a dictionary"
* "[climb] up the ladder"
* [take responsibility for]
* sing [all the way] to school
* [go] to school

Rule 2: consider the maximum extent of an event mention, but don't worry about determiners (they are invisible)

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

# Appendix: Example of scoring computation:

Sample System output:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| System Id | Doc Id | Event Mention Id | Token Id List | Mention Text | Event Type | Realis Status | System Confidence |
| sue | sample | E1 | 17 | advice | Communicate | Other | 1 |
| sue | sample | E2 | 19 | reassurance | Communicate | Other | 1 |
| sue | sample | E3 | 33 | came | Transport-Person | Actual | 1 |
| sue | sample | E4 | 52 | going | Transport-Person | Actual | 1 |

Gold annotations:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| System Id | Doc Id | Event Mention Id | Token Id List | Mention Text | Event Type | Realis Status | System Confidence |
| gold | sample | E1 | 52 | going | Transport-Person | Actual | 1 |
| gold | sample | E2 | 33 | came | Transport-Person | Actual | 1 |
| gold | sample | E3 | 87 | got | Transport-Person | Actual | 1 |
| gold | sample | E4 | 14,17,18,19 | offer advice or reassurance | Communicate | Other | 1 |

In the following tables, the “Event Type” and “Realis Status” are omitted for clarity

## STEP 1 : Compute overlap scores for each pair of Gold/System Mention

There are no invisible words, so no removal will be done

Compute the “mappingScore” table as followed:

|  |  |  |
| --- | --- | --- |
| Gold Mention | System Mention | Overlap |
| (E1, [52]) | (E4, [52]) | 1 |
| (E2, [33]) | (E3, [33]) | 1 |
| (E4, [14,17,18,19]) | (E1, [17]) | 2/5 (See #) |
| (E4, [14,17,18,19]) | (E2, [19]) | 2/5 (Same as above) |

# Example calculation of overlap:

Prec(G\_E4,S\_E1) = (|E1 ^ E4|) / |E1| = 1/1 = 1;

Recall(G\_E4,S\_E1) = (|E1 ^ E4|) / |E4| = ¼ = ¼;

Overlap(G\_E4,S\_E1) = 2 \* Prec(G\_E4,S\_E1) \* Recall(G\_E4,S\_E1) / (Prec(G\_E4,S\_E1) + Recall(G\_E4,S\_E1) ) = 2 \* 1 \* ¼ / (1 + ¼ ) = 2/5

## STEP2: After the calculation of all pairs, we can find the best mapping between System Mention and Gold Standard Mentions

Sort the “mappingScore” table based on overlap (Ties are currently break on their appearance in data):

|  |  |  |
| --- | --- | --- |
| Gold Mention | System Mention | Overlap |
| (E1, [52]) | (E4, [52]) | 1 |
| (E2, [33]) | (E3, [33]) | 1 |
| (E4, [14,17,18,19]) | (E1, [17]) | 2/5 |
| (E4, [14,17,18,19]) | (E2, [19]) | 2/5 |

We select mappings from the table above from top to bottom, and put already mapped items into a set A:

1. On row1, Select Gold, E1 to map to System, E4, A = {G\_E1, S\_E4}
2. On row2, Select Gold, E2 to map to System, E3, A = {G\_E1, S\_E4,G\_E2, S\_E3}
3. On row3, Select Gold, E4 to map to System, E1, A = {G\_E1, S\_E4,G\_E2, S\_E3, G\_E4, S\_E1}
4. On row4, We found that Gold E4 is already being mapped, we found that Gold E4 is already in set A, we will not include this mapping

We have the following mapping table (mappingScore table):

|  |  |  |
| --- | --- | --- |
| Gold Mention | System Mention | Overlap |
| (E1, [52]) | (E4, [52]) | 1 |
| (E2, [33]) | (E3, [33]) | 1 |
| (E4, [14,17,18,19]) | (E1, [17]) | 2/5 |

## STEP3.1: Compute document level errors and corrects

TP is the sum of the overlap in the mappingScore table:

TP = 1 + 1 + 2/5 = 2.4

S{E2} is not contained in the mappingScore table, so

FP = 1

## STEP3.2: Compute document level precision, recall:

Precision := TP / (TP+FP) = 2.4 / (2.4+1) = 0. 7059

Recall := TP / #GoldStandardMentions = 2.25/4 = 0.6

F1 := 2\*Precision\*Recall/ (Precision+Recall) = 2\*0. 7059\*0.6/ (0. 7059+0. 6) = 0.6487

## #STEP3.3: Compute mention and realis type detection score:

We see that the system get 3 mention type correctly (S{E1, E3, E4}) and 3 realis status correctly (S{E1, E3, E4}), we have totally 4 gold standard mentions.

Type\_detection\_accuracy := num\_type\_correct / #GoldStandardMentions = 0.75

Realis\_detection\_accuracy:= num\_realis\_correct / #GoldStandardMentions = 0.75

## Final Output:

### Output1: The score appended gold standard file will be like the following

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| System Id | Doc Id | Event Mention Id | Token Id List | Mention Text | Event Type | Realis Status | System Confidence | Sue Mention score |
| gold | sample | E1 | 52 | going | Transport-Person | Actual | 1 | 1 |
| gold | sample | E2 | 33 | came | Transport-Person | Actual | 1 | 1 |
| gold | sample | E3 | 87 | got | Transport-Person | Actual | 1 | - |
| gold | sample | E4 | 14,17,18,19 | offer advice or reassurance | Communicate | Other | 1 | 0.4 |

### Output2: Individual document performance and averaged performance

We only take one document as example, which make the micro and macro measures to be the same.

