**Event Mention Detection Scoring**

# Overall workflow

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

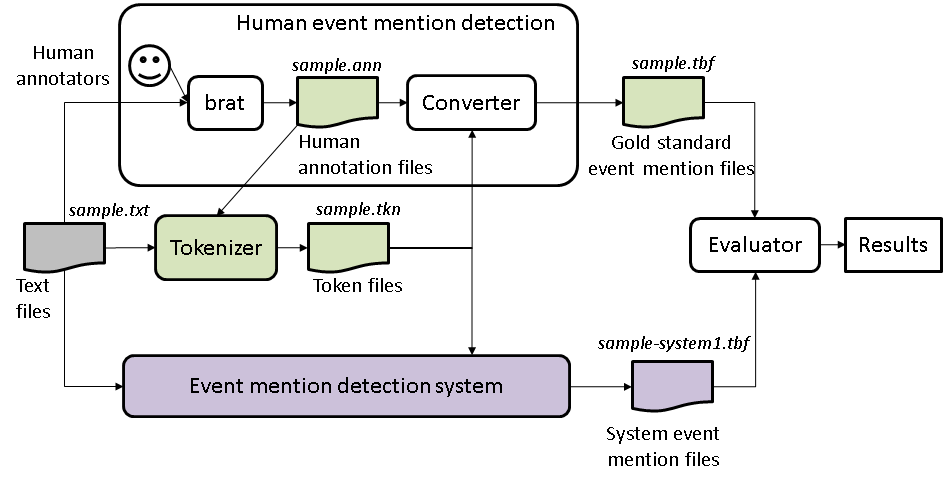


Figure 1: An overall workflow of evaluation for event mention detection.

For each text file, human annotators use the **brat** rapid annotation tool[[1]](#footnote-1) to create a gold standard annotation file. We convert the brat annotation file to our evaluation file format. We require the output of an event mention detection system to be given in the same file format as the gold standard file. The evaluator (scorer) takes the gold standard file and the output of an event mention detection system as input, and compares them to give a score for the system.

As shown in Figure 1, we carry out tokenization in our evaluation mechanism. We call it **pre-tokenization** since it is done before evaluation. The reason for pre-tokenization is two-fold. First, if we do not incorporate tokens in evaluation at all, then a system’s missing one token ends up with different scores according to the length of the token (i.e., the number of characters in the token). This is unfair from the perspective of evaluation. It is necessary to evaluate the performance of event mention detection in terms of tokens. Second, a modern English tokenizer is basically reliable enough to help human annotators correct (or rethink at least) their event mention annotation, particularly in the relatively unstable process of their creating the gold standard together. This is not only about trivial annotation errors (e.g., misselection of a span with extra whitespace in their web browser) but also about linguistically questionable event mentions. We will give a few examples to show the second point later.

The input and output of our tokenizer are defined as follows.

**Input of the tokenizer**:

1. Original text files
2. Gold standard annotation files in the brat standoff format[[2]](#footnote-2).

**Output of the tokenizer**: Token files

Each line of a token file is defined as follows: <token ID> <token string> <begin offset> <end offset>. The begin offset of the first token in the very beginning of a document starts with 0. Below is an example of the token file.

48 It 386 388

49 has 389 392

50 been 393 397

51 suggested 398 407

52 that 408 412

Our tokenizer implementation is based on the tokenizer in the Stanford CoreNLP tool[[3]](#footnote-3). It is implemented in Java, and the jar file is evm-eval-1.0.0-jar-with-dependencies.jar . The requirements of the software are as follows:

1. Java 1.7
2. The same number of text files and brat annotation files (\*.ann) with the same file base name.

You can run the tokenizer by the following command.

$ java -cp ./evm-eval-1.0.0-jar-with-dependencies.jar evmeval.TokenFileFactory

Since the following command does not have any command options, you should see the following usage.

usage: java evmeval.TokenFileFactory -a <annotation> -e <extension> [-h]

-o <output> [-s <separator>] -t <text>

-a <annotation> annotation directory

-e <extension> text file extension

-h print this message

-o <output> output directory

-s <separator> separator chars for tokenization

-t <text> text directory

As seen in the usage, the tokenizer takes a text file directory path and an annotation file directory path as input, instead of individual text files and annotation files. The tokenizer outputs the same number of output files as those input files under the specified output directory. You can run the tokenizer by specifying the options as follows, for example.

$ java -cp ./evm-eval-1.0.0-jar-with-dependencies.jar evmeval.TokenFileFactory -t ./data/LDC2014R19/ERE/ -a ./data/LDC2014R19/ERE/ -e txt -o ./output/ERE

If the tokenizer runs successfully, you might get the following message.

[WARN] Boundary mismatch found in d21dc2cb6e6435da7f9d9b0e5759e214: Token [1716,1735] [buffet/music/buying] vs. EventMentionSpan [1729,1735] [buying]

[WARN] Boundary mismatch found in d21dc2cb6e6435da7f9d9b0e5759e214: Token [1747,1754] [in/hire] vs. EventMentionSpan [1750,1754] [hire]

The message shows two boundary mismatches between a system token and an event mention span in the document d21dc2cb6e6435da7f9d9b0e5759e214. The first one says that the tokenizer has found a token buffet/music/buying whereas a human annotator annotated buying as an event mention. This example indicates that you can take advantage of the information to go over linguistically questionable event mentions as well as trivial annotation errors.

Note that an event mention span is not necessarily the whole of an event mention; the brat annotation tool can deal with discontinuous annotation, and thus an event mention span means either an event mention itself or a part of an event mention (discontinuously annotated).

**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.

# Formats

## 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

# Scoring

## Scoring for one document

We denote a gold standard mention with G, and a system mention with S, Overlap(G,S) is a token-based F1 function of G, S that returns a score between 0, 1.

(see the OVERLAP subroutine in the Pseudo-code (Appendix 1) for detail). All invisible words are already removed from G and S (See **Note 1**).

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 so that their Overlap score is the highest. However, it is possible to have multiple system mention to be mapped to one gold standard mention.

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

1. For each gold standard mention , we choose the system mention to find , denote . Let set be a set that contains all these system mentions
2. True Positive =
3. False Positive = #System Mention - | |
4. Precision = True Positive /(True Positive + False Positive)
5. Recall = True Positive / #Gold Mention
6. F1 = H (Precision, Recall), where H is the harmonic average function

To score realis status and mention type detection, we use the same mapping:

1. For each gold standard mention , we count the number of system mention that are mapped to it as ,
2. We use \_realis and \_mention to define the realis status and mention type of the mention respectively.
3. Initialize with realis\_score = 0; mention\_score = 0
4. If \_realis = realis , realis\_score = realis\_score + 1/;
5. Similarly, if \_mention = mention , mention\_score = mention\_score + 1/;
6. realis\_detection\_accuracy = realis\_score / #GoldStandardMentions
7. type\_detection\_accuracy = mention\_score / #GoldStandardMentions

**Note 1**: 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 removed from the list because they can occasionally be resolved as nominal event mentions.

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

## Summarization score

After all documents are scored, we will also report scores 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)

# Appendix 1: 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 table 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}

ELSE IF G has been mapped but S has not been mapped

THEN Find the row R that contains G, append S to its system mentions.

END IF

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

Mapping\_num:= |{S}|

Single\_score := 1/ Mapping\_num

FOR EACH LINE S in {S}

IF G\_type == S\_type

type\_correct := type\_correct + Single\_score

END IF

IF G\_realis == S\_realis

realis\_correct := realis\_correct + Single\_score

END IF

END FOR

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

# Appendix 2: 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:

1. On row1, Select Gold, E1 to map to System, E4, we also record the overlap score 1.
2. On row2, Select Gold, E2 to map to System, E3, we also record the overlap score 1.
3. On row3, Select Gold, E4 to map to System, E1, we record the overlap score 2/5
4. On row4, Select Gold, E4 to map to System, E2, we can see that Gold E4 has already been mapped to a mention System E1, we do not record overlap score, but we record the system mention here so we know that E4 is mapped to 2 system mention

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]) , (E2,[19]) | 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 type and realis status detection score:

For each row in the mapping table, we check whether the system mention have the same realis status and mention type with the gold mention.

G\_E1 – S\_E4 and G\_E2 – S\_E3 are both one to one mapping, so N will be 1. Both mention types and realis status are correct, so that we have type\_score = 2, realis\_score = 2.

G\_E4 is mapped to 2 mentions {S\_E1, S\_E2}, so N = 2. Both mention types and realis status are correct, so that we have type\_score = ½ + ½ = 1, realis\_score = ½ + ½ = 1.

The sum of type score is 2 + 1 = 3, and the total realis score 2+1 = 3. We than can calculate the accuracy:

Type\_detection\_accuracy := 3 / #GoldStandardMentions = 0.75

Realis\_detection\_accuracy:= 3 / #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.



1. http://brat.nlplab.org/ [↑](#footnote-ref-1)
2. http://brat.nlplab.org/standoff.html [↑](#footnote-ref-2)
3. http://nlp.stanford.edu/software/corenlp.shtml [↑](#footnote-ref-3)