Event Argument Extraction and Linking (EAEL) Scoring Proposal

Change-log

July 14th: Updates for clarity & to reflect the implemented approach to aggregating on a per document level. (Note: this reelects the existing implementation of the code on github and did not change the numbers in Table 1)

Overview

The 2015 TAC KBP Event Argument Extraction and Linking Task requires participants to locate event arguments in text and link them together into event frames. System performance will be evaluated by comparing system output on each document to reference event frames (REFs).

Creating Reference Event Frames

REFs will be created in the following way:

1. All responses from all event frames from all systems will be pooled. This will be called the *argument pool*.
2. As in the 2014 task, LDC assessors will
    a. assess all responses in this pool as described in "Dimensions of Assessment" below.
    b. group the canonical argument strings from all responses into coreference clusters
3. A *linking response pool* will be formed from all responses which
    a. are CORRECT/INEXACT in AET, AER, BF, and CAS, and
    b. have a realis label of OTHER or ACTUAL[1]
4. All responses in the argument and linking response pool will be automatically grouped into equivalence classes called TRFRs[2] based on event type, event role, realis, and CAS coreference cluster, producing the *argument TRFR pool* (A) and the linking *TRFR pool* (L).[3]
5. LDC annotators will group all TRFRs in the linking TRFR pool into the reference event frames.

Scoring a Document against a REF set

Desirable Properties of a Scoring Function

We sought the following properties in a scoring function:

1. For the linking component of the score, neither over-merging nor over-splitting should be strongly preferred.

---

[2] for 'Type, Role, [Normalized] Filler, Realis"

[3] The argument and linking TRFR pools differ only in that only the argument pool contains generics.

2. The linking scoring component must allow for the same TRFR to appear in multiple event frames. This is a requirement because we are scoring at the Entity/CAS level. The same entity often participates in multiple events.
3. To ensure the evaluation encourages advancements in both event argument extraction and event linking, we wish to ensure that improvements to either considered separately never cause a decrease in the overall score. This is straightforward for linking but trickier for argument extraction. Formally, consider a partial ordering over event argument extraction outputs (ignoring linking) where X $>_{EAE}$ Y iff X - Y consists entirely of true positives (TP) and Y-X consists entirely of false positives (FP). The partial ordering $>_{EAEL}$ over event argument linkings induced by the EAEL scoring function should be consistent with $>_{EAE}$.

Event Argument Extraction Sub-score (unnormalized)

For the event argument extraction sub-score we use the linear function $TP_{EAE} - \beta FP_{EAE}$ for some parameter β, where TRFR true and false positives are against the *argument TRFR pool* and use the definitions of the 2014 KBP EA Standard metric. Intuitively, this corresponds to a model where the user of an EAEL system derives utility 1 from a TP_{EAE} and loses utility $\beta$from an FP_{EAE}. Note that this differs from the F-measure-based score used in 2014. We will continue to report the F-based metric as an independent diagnostic measure of extraction performance (ignoring linking)

Linking Sub-score (unnormalized)

There are a number of clustering metrics available, including CEAF, B^3, BLANC, etc. Many of them can be straightforwardly applied to event frames subject to the modification that TRFRs may appear in multiple frames.

We propose to use the following variant of B^3[4]:

1. Let $L\ \ (d)$ be the system-provided TRFR linking for a document d. Let $R(d)$ be the reference TRFR linking, where the *i*th event frame is a set of TRFRs denoted $R_i(d)$. Define $\widehat{L\ \ (d)}$to be $L\ \ (d)$ with all TRFRs not found in $R\ \ (d)$ removed (that is, $L\ \ (d)$ without EAE false positives).
2. Define $v_Y(x)$ for a linking $Y$to be $(\cup_{Z\in Y s.t. x\in Z} Z) - x$ (that is, all TRFRs which are present in a common event frame with x, excluding $x$itself).
3. Define $f_{Y,Z}(x)$, the per-TRFR link F-measure, as:
    a. If $x$is not in Z, $f(x) = 0$
    b. If $x \in Z$and $v_Y(x)$ and $v_Z(x)$ are empty, then $f(x) = 1$.
    c. Otherwise, let $p_{Y,Z}(x)$, the precision, be $\frac{|v_Z(x)\cap v_Y(x)|}{|v_Z(x)|}$. Let $r_{Y,Z}(x)$, the recall, be $\frac{|v_Z(x)\cap v_Y(x)|}{|v_Y(x)|}$. $f_{Y,Z}(x) = \frac{2p_{Y,Z}(x)r_{Y,Z}(x)}{p_{Y,Z}(x)+r_{Y,Z}(x)}$
4. Let $U_X(d)$ be the union of all event frames in X. We define $S_{EAL}(d,R,L)$ as $\sum_{x\in U_R(d)}\ \ f_{\widehat{L,R}}(x)$. Intuitively, it is the sum of the link F scores for each TRFR present in the gold standard.

While B^3 has fallen out of favor for coreference evaluations due to its tendency to compress scores into a small range when there are many singletons, singletons are far less common in the EAEL task, so this does not appear to be a concern.[5]

Aggregating S_{EAE} and S_{EL} at a Per-Document Level

Systems which wish to compute a normalized per-document score can use $[\lambda[max(0, S_{EAE})/|A_{correct}|] + (1-\lambda)S_{EAL}/|L|]$, where $|A_{correct}|$ is the number of correct TRFRs in the argument TRFR pool and and L is the number of TRFRs in the linking TRFR pools. Note that while S_{EAE}can be negative, we clip it to 0.. Similarly, for diagnostic purposes we can compute document-level $S_{EAE}$and S_{EAL} scores by dividing the raw scores by the appropriate normalizers.

Aggregating Scores across the Evaluation Corpus

We define the score of a corpus as a $\lambda \frac{\sum_{d \in D} \max(S_{EAE}(d), 0)}{\sum_{d \in D} |A_{correct}(d)|} + (1-\lambda) \frac{\sum_{d \in D} S_{EAL}(d)}{\sum_{d \in D} |L(d))|}$ where D is the set of documents.

Reasons for not Selecting Alternatives

Not Selecting F for Argument Extraction Subscore

The linking subscore is biased towards recall— only responses that are found can generate True Positive links. When we attempt to use F1 in combination with the linking subscore, we see that systems with very high recall but low precision are the top performing systems, suggesting that systems trying to achieve high performance in the evaluation would be encouraged to pursue recall while ignoring precision. This is illustrated with several examples in Table 1[6]. Each row represents a system: The first two rows are the rank1 and rank5 system from the 2014 evaluation. The following three systems are three possible operating points- a system with improved precision and recall over the 2014 rank1 system, and systems that have highly imbalanced recall and precision (at F1 close to the 2014 rank5 system). The first three columns show precision, recall and F1 for each system. The next three columns show the performance in terms of the proposed metric for each system assuming different levels of link accuracy.[7] The final three columns show performance when F1 is substituted for our proposed event argument extraction subscore. In all cases, the system with a precision of 10 and a recall of 75 is the top performing system. This ranking does not intuitively map to expected utility.

	P	R	F1	Proposed Link=0.6	Proposed Link=0.7	Proposed Link=0.8	Using F Link=0.6	Using F Link=0.7	Using F Link=0.8
2014_Rank1	43	24	30.8	15.2	16.4	17.6	22.6	23.8	25.0
2014_Rank5	19	17	17.9	4.5	5.4	6.2	14.1	14.9	15.8

[5] In ACE annotation, event frame sizes of two and three are most common and are twice as likely as singletons.

[6] As proposed in the official ranking score, we use $\beta = 1/4, \lambda = 1/2$ for this example.

[7] We estimate that for a small set of challenging documents (i.e. documents that describe several events of the same event type) a baseline link strategy of "link all arguments that participate in an event of the same type" achieves a link accuracy of 0.6, thus link accuracy in the range of 0.6-0.8 seem likely in the evaluation.

Improved	53	34	41.4	23.4	25.1	26.8	30.9	32.6	34.3
Ignore_Rec	75	10	17.6	7.6	8.1	8.6	11.8	12.3	12.8
Ignore_Prec	10	75	17.6	-24.4	-20.6	-16.9	31.3	35.1	38.8

Table 1: Example Systems with Scores[8]

Not Selecting CEAF for Linking Subscore

CEAF has two disadvantages for this task. As we explain them, we will make reference to the following key and example clusterings: **key**: { {a, b} {c, d} {e, f, g} {h, i, j} {k, l, m, n} {o} }, **example 1:** { {k,l} {m,n} }, **example 2:** { {a, b} {c, d} {x} {y} {z} }, **example 3**: { {a, b, x} {c, d, y} {z} }.

CEAF's first disadvantage is that it can harshly penalize over-splitting clusters due to its constraint that each key cluster can only align to at most one system cluster. In example 1, the correct links in one cluster would be entirely ignored. This is a particular problem for this task because judging the proper granularity of events can be challenging even for humans.

The second disadvantage is insensitivity to spurious links. Although example 3 should be penalized for the spurious inclusion of x and y in the two clusters, CEAF gives the same score to both examples 2 and examples 3 because in both instances the number of aligned elements, key elements, and system elements are the same

Official Ranking Score

For the official ranking score, we will use $\beta = 1/4, \lambda = 1/2$ to weigh argument extraction and linking performance roughly equally[9] and to encourage high recall while maintaining reasonable precision. Because the choice of these parameters is somewhat arbitrary and has a significant impact on the evaluation, we are open to input from participants about what they should be. We will also do an analysis of the sensitivity of the final ranking to variation in the parameters.

The score used for final system ranking will be the median corpus-aggregated S_{EAEL} over 1,000 corpora bootstrap-sampled from the LDC-provided evaluation corpus. We will report for each rank the fraction of samples on which it outperforms each other rank.

Diagnostic Measures

The following diagnostic measures will be calculated but will not be used for system ranking:

- scores for newswire only and discussion forums only
- scores on the LDC-provided evaluation corpus only, without sampling
- a 2014 KBP EA-style argument extraction score (F1 over the assessment pool)
- graphs of how systems' evaluation scores would vary with changes to $\beta$and λ.
- a "macro-F" version of the score, where we compute the score on a per-document basis and take the mean.

[8] The scores in this table a presented without the clipping process described above. Actual 2015 evaluation scores will not fall below 0.

[9] This is not exact because the ranges of likely variation of the two sub-scores differ somewhat and recall affects linking scores because you can't link what you can't find.