BBN’s 2017 KBP EAL Submission

Jay DeYoung, Yee Seng Chan, Chinnu Pittapally, Hannah Provenza
Raytheon BBN Technologies
Cambridge, MA
{jay.deyoung,yeeseng.chan,chinnu.pittapally,hannah.provenza}@raytheon.com

Ryan Gabbard, Marjorie Freedman
USC ISI
Waltham, MA
{gabbard,mrf}@isi.edu

Abstract

We participated in the Event Argument Extraction and Linking (EAEL) task in TAC KBP 2017. We trained separate event anchor, event argument, and realis models using logistic regression on the ACE, TAC, and internal event mention data, employed a simple form of joint inference, applied a variety of document-level inference rules, and applied a sieve-based event linking system to find events at the document level. Our scores were the highest among all systems.

1 Introduction

In this paper, we describe the event system we developed for the multi-lingual Event Argument Extraction and Linking (EAEL) task in TAC KBP 2017. Except where explicitly specified, we implemented all features and capabilites for English, Spanish, and Chinese. In the next section, we first describe SERIF’s text graphs, an enriched dependency structure which our extraction models rely on heavily. In Section 3, Section 4, and Section 5, we describe our anchor, argument, and realis models. In Section 6, we describe our document-level inference component, in Section 7 we discuss our in document event argument coreference, and in Section 8 we enumerate our resources. In Section 9, we analyze our results and then conclude in Section 10.

Role	Description
SUBJ	logical subject of a verb
OBJ	logical object of a verb
IOBJ	logical indirect object of a verb
LOC	locative modifier to a verb (*he went* ***home*** *on Tuesday)*
TEMP	temporal modifier (*he went home on* ***Tuesday****)*
POSS	possession modifier, *his son*
PMOD	the proper noun modifier to a nominal (*US president)*
AMOD	adjectival modifier to a noun
{prep}	uses the preposition word as the role

Table 1: SERIF’s text graph edges.

2 Text Graphs

SERIF’s text graphs (TG) are enriched dependency structures which can be automatically built from parse trees. They differ from other dependency representations (de Marneffe et al., 2006) by doing more extensive normalization and incorporating long-distance dependencies. Table 1 describes the edge labels in text graphs. [1]

3 Anchor Identification

The SERIF anchor extraction system is a supervised logistic regression model which marks words as ‘an-

This work was funded by DARPA/AFRL Contract FA8750-13-C-0008.

The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views of policies of the Department of Defense or the U.S. Government.

[1] While these capabilities exist for Chinese and Spanish, they are far less developed. Spanish, in particular, suffers from the lack of a parser that handles obligatory pronoun-dropping, so it has many more incorrect propositions.

Category	Feature
Cluster	Brown cluster bit strings of a (at bit lengths 8, 12, 16, 20); Brown cluster bit string of subject and objects words)of a (if any; bit length 16)
WordNet	Hypernym synsets of a (English only)
Words	a; word before a; word after a
TG role	TG *subj* and *obj* words of a; other words that are connected to a in the TG; if a is copular, its TG *subj*, *obj* words
Topic of document	topic; topic & a; topic & brown clusters of a
Movement-related	The preposition, adverb, particle word after a; TG role of any GPE, FAC, or LOC connected to a; whether there is a locative modifier (i.e. LOC TG role) to a.
Noun-compound	If a is part of a noun compound (*army raids*), indicate the relative position of a (front, back, or middle) in the compound.
Event word classes	Event word classes associated with a. (English only)

Table 2: Event anchor a features.

choring' events (or not). We list the anchor features in Table 2. We describe some of the features below:

Topic To determine the topic of a document, we use a simple strategy of matching words in the document against short predefined word lists corresponding to seven topics (PERSONNEL, ELECTION, FINANCE, JUSTICE, VIOLENCE, MIDDLE EAST VIOLENCE, SPORTS). The same set of topics were used across all three languages, translated by a speaker proficient in casual conversation on a per-language basis.

Movement-related Two of our target event types are MOVEMENT.TRANSPORT-PERSON and MOVEMENT.TRANSPORT-ARTIFACT. To highlight the importance of prepositions, locative modifiers, and GPE, FAC, and LOC mentions for these event types, we add features that explicitly check for these.

Noun compounds Event anchors are frequently deverbal nouns and sometimes occur as part of noun compounds. We add a simple feature to capture the relative position of the anchor in such noun compounds.

Event Word Class (English only) Using word lists is a simple but useful way to define concepts (e.g. WordNet senses or synsets consist of groups of words). Hence, one of the baseline anchor features is the anchor word a. However, such a feature is inevitably sparse when collected from a limited amount of training data. To augment this, we employed a semi-automatic approach to collect word lists on concepts related to our target event types. Through leveraging WordNet (Miller, 1995), we are able to quickly gather such word lists starting from a few user supplied seed words per concept.

For each concept that we are interested in, e.g. LIFE.INJURE, we first supply a few seed words, e.g., *injure*/V, *injury*/N, *wound*/V, *hurt*/V, etc. We then automatically retrieve FrameNet frames whose lexical units overlap with these seed words and present these frames and their lexical units to the user for a binary relevant/not-relevant decision. Using WordNet, we also present derivationally related forms of the seed words and relevant lexical units to the user for selection. Finally, we leverage WordNet's hyponym hierarchy to automatically expand the set of words.

We gathered word lists for 52 concepts. These mostly overlap with the ACE event types but also contain several related concepts (e.g. arsons, bombings, deportations, embezzlements, rebellions, shootings, stabbings, etc.). We initiated the process with an average of 2.9 seed words per concept.

4 Argument Identification

The SERIF event argument identification (EA) system is a supervised logistic regression model which relies on the anchor model described in the previous section to identify event anchors. Given an anchor-mention pair (a, m), the model predicts whether any of the predefined event argument roles hold between (a, m). We list our event argument features in Table 3. They capture some of the indicators of an event role, e.g. Subject/Object, aspects of nominal relations (*mod*, but without further typing of the noun compound relation), and several of the preposition cases. In our system, event type (EVTYPE) is derived directly from the event anchor predictions. We describe some of the features below.

TG path In the baseline EA features, the text graph features capture only direct or shared connections, i.e. when m is directly connected to a via a single edge to a or when m and a directly depend on a common node (e.g. *people* ⟨*obj*⟩ *killed* ⟨*in*⟩ *raid* : ⟨obj⟩_⟨in⟩). In our full system, we add all (shortest) paths in the text graph between arguments and anchors of length up to three.

TG roles To add more contextual information, we also gathered all words (and their associated TG roles) which are directly connected to m.

Category	Feature
TG role	TG[2] & EVTYPE[3] & ENTTYPE; TG & EVTYPE & ENTTYPE & a
Candidate	hw_m & EVTYPE-R[4]; hw_m & EVTYPE-R & a
String between (a, m)	String & EVTYPE-R; stemmed string & EVTYPE-R; abbrev string & EVTYPE-R
token distance between (a, m)	distance & EVTYPE & ENTTYPE; distance & EVTYPE-R & ENTTYPE
TG path	Path from a to m, i.e. the sequence of TG edges labels connecting a to m; bag of TG edge labels in the path.
TG role	If there is a path from a to m, get the TG role directly connected to m; set of all (TG role, word) pairs for words sharing a direct TG connection with m.

Table 3: Event argument features. Abbreviations are as follows. a : anchor, m : candidate argument (mention), hw : headword, ENTTYPE:entity type of m (e.g. PER)

[2] TG here indicates the TG roles between (a,m), if any.

[3] We use both the ACE event type and subtype.

[4] To reduce sparsity, in some cases we collapsed several event subtypes. In these cases, the event type feature is denoted by EVTYPE-R

4.1 Word and Contextual Embeddings

Event arguments are often found far from their triggers, making argument-anchor relationship features sparse. However, the local context of these arguments often contain informative clues (e.g. *Acme Inc.****'s*** **creditors** were disappointed by Friday's bankruptcy filing.). We wished to learn such informative contexts which never appear in our training data based on those which do, so we employed a variant of the skip-gram word embedding model (Levy and Goldberg, 2014) over Gigaword corpora to simultaneously build dense vector representations of words and their prop-graph contexts. The corpora used for training are listed in 4.1.

Language	Release	Description
English	LDC2011T07	English, V5
Chinese	LDC2011T13	Chinese, V5
Spanish	LDC2011T12	Spanish, V3

Table 4: Word and Context Embedding Corpora

5 Realis Prediction

For the TAC KBP event argument attachment evaluation, it was necessary to label each event argument with a realis of ACTUAL, GENERIC, or OTHER (where OTHER includes negatives, future, etc.). Our system determines realis on a per-event basis and propagates it to the arguments. An event here is defined as an anchor and its arguments as provided by the anchor and argument identification models. When additional arguments are created later by inference rules, realis is propagated to them as well. We compute realis by first computing two separate scores: p_{ASSERTED} and p_{SPECIFIC}. We assign probabilities to each of ACTUAL, GENERIC, and OTHER as follows:

- $p_{\text{ACTUAL}} = p_{\text{ASSERTED}} * p_{\text{SPECIFIC}}$
- $p_{\text{GENERIC}} = (1.0 - p_{\text{SPECIFIC}})$
- $p_{\text{OTHER}} = p_{\text{SPECIFIC}} * (1.0 - p_{\text{ASSERTED}})$

p_{SPECIFIC} is always 0 or 1, derived by rule from modality information which is itself added by rule to BBN's text graphs. p_{ASSERTED} comes from an event-level classifier trained on the genericity annotations in the ACE corpus. The features used are listed in Table 5.

Category	Features
Of the whole event	type, number of arguments
Of the anchor	POS; word itself; suffixes; properties of TG node; POS and word for preceding and following tokens; preceding and following POS trigrams
Of each argument	**entity type**, **entity level** (name, descriptor, pronoun), headword, determiners, **part-of-speech**, **if all person arguments are plural**
Of the anchor's sentence	Contains a specific date, a number, or an infinitive; **is a question** ; **token length**; **sentence contains certain function words**
Of the document	**genre**
Conjunctions	Anchor POS & infinitive in sentence, Anchor POS and argument determiner

Table 5: Features of Genericity Classifier.

6 Document-level inference

For our TAC KBP event argument attachment system, we first generate all events predicted with probability greater than 10% by our anchor model and all event arguments predicted with greater than 10% probability for those events. The event mention predictions are passed on to a document-level inference component which produces our final output.

6.1 Inference Rules employed in BBN's submission

The document-level inference component applies the rules below.

Copy Violent Event Existence This rule copies violent events to other violent events under certain conditions. It copies a LIFE.DIE or LIFE.INJURE event to a CONFLICT.ATTACK event with the same arguments if an AGENT is present in the source event.

Aggressive Role Search For each event type, we define a set of roles which the system should attempt to find aggressively and other roles that it should attempt to find aggressively if certain conditions are satisfied. At runtime, for every event, the system checks if there are any missing roles which should be searched for aggressively. If so, it searches as follows:

- First, does an argument of the desired role exist in another event which did not meet the scoring threshold? If so, use it.
- Second, does an argument of the desired role exist in any event within the preceding four or following two sentences? If so, use it.
- Third, search the output of our model which predicts event argument status independently for all mentions in a document. If any matching argument exists, take the nearest by sentence distance. In discussion forum documents, all searches are limited to the post of the original event. Only the first search strategy is used for non-ACTUAL events.

Location Inference If any mention x filling a PLACE argument role is known to be part of another entity y by our ACE relation system, we add y as a filler for the PLACE role as well.

Geonames If any mention x filling a PLACE argument role is known to be part of another entity y by a gazeteer, we add y as a filler for the PLACE role as well.

Delete non-GPE Places The system deletes all PLACE arguments which are not GPEs for all but a few event types.

Delete Actual *You* This system deletes all arguments where the base filler is some variant of the second-person pronoun and the realis is ACTUAL.

Delete Missing This rule deletes all PERSONNEL.-END-POSITION events which lack a POSITION.

Movement-Transport Copy Copies all MOVEMENT.TRANSPORT-ARTIFACT events without an ARTIFACT to MOVEMENT.TRANSPORT-PERSON and all MOVEMENT.TRANSPORT-PERSONs without a PERSON to MOVEMENT.-TRANSPORT-ARTIFACT.

Copy Contact For some anchors, it is difficult to tell what sort of ‘Contact‘ event should be predicted. For selected anchors, we copy all ‘Contact‘ events of one sub-type to another sub-type.

Generate with altered realis Whenever a non-best realis assignment has positive expected value given the scoring metric, add a copy of the argument with that realis.

6.2 Final Scoring

The final score for a tuple is the geometric mean of the following sub-scores:

- The anchor model score
- The event argument attachment model score
- A coreference score which is always 1.0, except when the base filler is a non-relative pronoun, in which case it is 0.75.
- The realis model score

For our primary submission, all tuples with a score over 0.50 were kept, using a threshold tuned from the 2016 dry run data. Responses deleted by inference rules are not considered in the final scoring phase. Inference rules will supply values for the sub-scores above for newly added responses in ways which vary from one rule to another, and they will occasionally alter the sub-scores for existing responses.

7 Linking Sieve

After finding event arguments, all submissions linked them into event-frames using a sieve-based approach, applying deterministic linking rules in order from highest to lowest precision. The rules were developed based on the newswire portion of the 2015 EAL training data (LDC2015E41).

Our linking rules are:

1. Arguments which share an event anchor and are within the same sentence are linked.

2. In each sentence, moving from left to right, merge event frames of the same type unless the ontology constraints below would be violated or certain discourse connectives are observed.

3. Moving through the document from earlier to later sentences, merge events of the same type unless they violate the ontology constraints below.

The ontology constraints are:

- Certain roles must be filled with only a single entity within an event frame (e.g. ORG in BUSINESS.DECLARE-BANKRUPTCY).

- Voluntary anchors may not be combined with involuntary anchors (e.g. *give* and *steal*).

8 Training Data

The following training data was used for all document-level English tasks:

- ACE English Events data

- LDC2015R26_TAC_KBP_2015_Event_Nugget_and_Event_Corefence_Linking

- LDC2015E29_DEFT_Rich_ERE_English_Training_Annotation_V1

- LDC2015E68_DEFT_Rich_ERE_English_Training_Annotation_R2_V2

- LDC2015E78_DEFT_Rich_ERE_Chinese_and_English_Parallel_Annotation_V2

- LDC2016E31_DEFT_Rich_ERE_English_Training_Annotation_R3

- LDC2016E73_TAC_KBP_2016_Eval_Core_Set_Rich_ERE_Annotation_with_Augmented_Event_Argument_V2

- Targeted Training data, described in 8.1

For Chinese we used:

- ACE Chinese Events data

- LDC2015E105_DEFT_Rich_ERE_Chinese_Training_Annotation

- LDC2015E112_DEFT_Rich_ERE_Chinese_Training_Annotation_R2

- LDC2015E78_DEFT_Rich_ERE_Chinese_and_English_Parallel_Annotation_V2

- LDC2016E73_TAC_KBP_2016_Eval_Core_Set_Rich_ERE_Annotation_with_Augmented_Event_Argument_v2

- Targeted Training data, described in 8.1

For Spanish we used:

- LDC2015E107_V2

- LDC2016E34_R2

- LDC2016E73_TAC_KBP_2016_Eval_Core_Set_Rich_ERE_Annotation_with_Augmented_Event_Argument_v2

Additionally, targeted training data described in the following subsection was used for all English and Chinese document-level tasks except realis.

8.1 Targeted Training

Community resources such as ACE and Rich ERE have to serve many purposes, some of which require full-document annotation. However, achieving broad coverage training data for events with full document annotation is challenging: in ACE 2005, 10 of 33 event types occur less than 25 times. Even when an event is common, each anchor may occur only one or two times, making it difficult for a classifier to learn them.

We explored the creation of focused training data specifically for the EA task. We prioritize creating a data set the system can easily learn from (even though it would make a terrible test set) and finding examples of things we care about over the natural distribution of instances.

We perform sentence-selected rather than full-document annotation. Annotators are provided with an simple search interface and are allowed to use their own judgment to locate useful training examples. They are explicitly encouraged to skip ”confusing” sentences the system may have difficulty learning from. Annotators are also allowed to mark more than one anchor for an event.

For English, our annotators annotated about 5,800 positive and 6,400 negative training sentences. Each event type had from two to eight hours of annotation. For Chinese, our annotator annotated about 200 positive and 100 negative training sentences. Each event type had one to two hours of annotation. The resulting annotation is far denser than ACE and has negative examples which are particularly useful because they are expected to be closer to the classifier's decision margin (e.g. involving alternative senses of a potential anchor).

9 Discussion

9.1 Performance

We report argument scores in Table 6 for F-measure and Table 7 for the official linear score. Linking scores are given by a bootstrap interval reported in Table 8. We provide a brief breakdown of component measures and limiting factors in 9.2.

Lang/Realis	Precision	Recall	F
eng/with	33.36	17.30	22.79
eng/no	44.51	25.01	32.03
cmn/with	39.48	17.74	24.48
cmn/no	46.27	21.70	29.54
spa/with	22.53	5.22	8.47
spa/no	28.38	7.12	11.39

Table 6: Document-level argument scores.

Lang/Realis	ArgScore 5%	50%	95%
eng/with	8.72	9.63	10.62
eng/no	16.51	17.49	18.49
cmn/with	10.36	11.81	13.00
cmn/no	14.37	15.75	16.93
spa/with	2.01	2.47	2.92
spa/no	3.22	3.82	4.37

Table 7: Document-level argument scores. Arg X% indicates the official argument score at the Xth percentile of bootstrap samples

Lang/Realis	Link 5%	Link 50%	Link 95%
eng/with	6.96	7.81	8.78
eng/no	8.58	9.50	10.62
cmn/with	4.98	5.78	6.60
cmn/no	6.06	6.92	7.78
spa/with	1.29	1.70	2.19
spa/no	1.55	2.05	2.60

Table 8: Document-level linking scores, percentage intervals determined by bootstrap sampling.

9.2 Limitations of Event Argument Performance

We produce internal measures of several types of scores that contribute to overall argument performance: anchor detection in Table 9, event mention identification at the sentence level in Table 10, and NER in Table 11.

From the two event detection measures, we find when we miss a anchor of a particular event type, we typically find a different anchor of the same type in the sentence: this more relaxed metric gives a boost of 4.5F in English and 6.5F in Chinese. The limitation of the precise trigger model is partially mitigated by the discovery of the second anchor.

We additionally consider our event performance on gold standard NER and information extraction, with results in Table 11. We show that our event system performance is penalized by the base NER system, which was trained targeting ACE and thus suffers a domain mismatch compared to the TAC 2017 data.

10 Conclusion

In this paper, we describe our systems which participated in the Event Argument Extraction and Linking task of TAC KBP 2017. We note the particular limitations of our current system and identify potential areas of improvement.

Acknowledgments

References

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*.

Omar Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL 2014*.

Lang/Genre	Precision	Recall	F
eng/nw	50.36	46.20	48.19
eng/df	37.20	31.19	33.93
eng/both	43.79	38.37	40.90
cmn/nw	61.15	42.04	49.83
cmn/df	52.59	30.67	38.75
cmn/both	57.02	36.09	44.20
spa/nw	33.59	24.44	28.30
spa/df	43.44	19.02	26.46
spa/both	37.73	21.48	27.37

Table 9: Identification of the correct event type at the anchor level, broken down by genre, on the LDC2017E55 dataset. Note that this is different from the event nuggets measure because it identifies the type and anchor combination on a sentence level, counted once per sentence, not on an argument level.

Lang/Genre	Precision	Recall	F	Fimprov
eng/nw	55.04	52.35	53.66	+5.5
eng/df	39.48	35.92	37.62	+3.7
eng/both	47.12	43.81	45.41	+4.5
cmn/nw	68.26	50.35	57.95	+8.1
cmn/df	58.13	37.18	45.35	+6.6
cmn/both	63.48	43.66	51.74	+7.5
spa/nw	20.26	15.39	17.49	-10.8
spa/df	45.69	21.71	29.43	+ 3.0
spa/both	31.36	18.88	23.57	-3.8

Table 10: Identification of the correct event type at the sentence level, broken down by genre. An astute reader will note that Spanish sentence level event type detection scores are lower than anchor identification in Table 9. This is an artifact of how our internal pipeline segments sentences differently given gold standard information than without it; for Spanish these segmentations are inconsistent. Without gold standard segmentation, this metric should only be used to evaluate whether or not the system generally finds a anchor of the correct type while missing the gold standard anchor.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November.

Lang/Realis	Precision	Recall	F	+Fimprov
eng/with	37.25	17.52	23.84	+1.0
eng/no	52.58	26.76	35.47	+3.4
cmn/with	46.26	22.74	30.49	+6.0
cmn/no	56.28	28.89	38.18	+8.6
spa/with	25.65	7.78	11.94	+3.5
spa/no	37.30	12.22	18.41	+7.0

Table 11: System results on gold standard information extraction (NER and values). This shows large improvements over those in Table 6, particularly in the noRealis case. This is shared across languages, although much smaller for the English system.