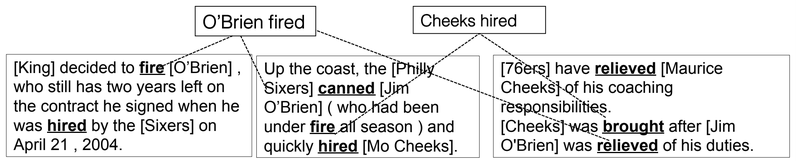
Revisiting the Evaluation for Cross Document Event Coreference

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, Dan Roth
University of Illinois at Urbana-Champaign, IL

Cross Document Event Coreference (CDEC)



Event mention is defined by a predicate (**fire**) and its arguments ([King] and [O'Brien])
Two event mentions are coreferent if they describe the same event.

The CDEC Task

Input: A document collection, describing several events.
Output: The system outputs an assignment of a cluster-id to each event mention, such that event mentions belonging to different documents but sharing the same cluster-id are coreferent.

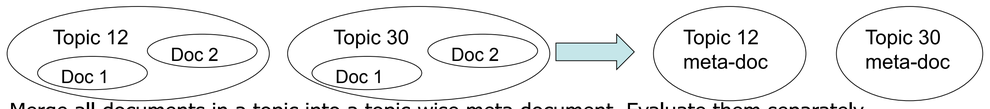
Aim: identifying coreferent event mentions across documents.
Value: information extraction and aggregation, topic tracking, multi-document summarization, relation extraction, paraphrasing …

This Work: We revisit the evaluation paradigm for CDEC and reveal their limitations. We propose **two new evaluations** to address these limitations.

Current Evaluation Setups

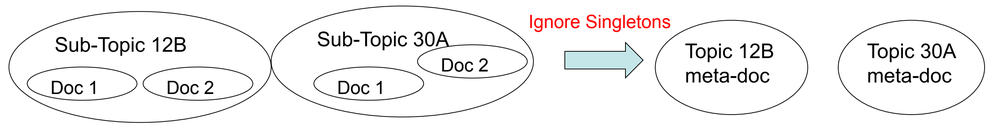
All evaluation setups convert the cross document CDEC problem to a WDEC problem.

◆**Bejan and Harabagiu (B&H)**



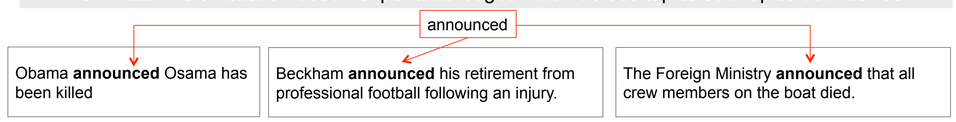
Merge all documents in a topic into a topic-wise meta document. Evaluate them separately.

◆**Yang et al. (YCF)**



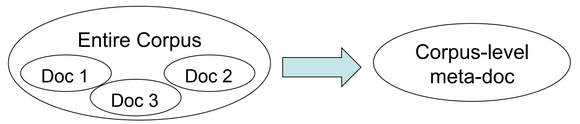
- Merge all documents in a topic into a topic-wise meta document. Evaluate them separately.

PITFALL: the evaluation does not penalize edges made across topics/sub-topics boundaries

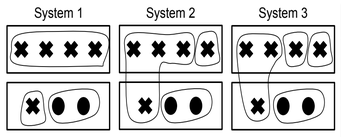


Proposed Evaluations

Simple-CDEC

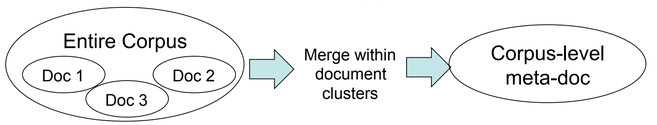


PITFALL: cannot isolate cross-doc links from within-doc links



	MUC	B3	CEAF	BLANC
Sys. 1	88.8	87.0	75.5	80.5
Sys. 2	88.8	87.0	75.5	80.5
Sys. 3	74.9	74.9	58.3	63.7

Pure-CDEC

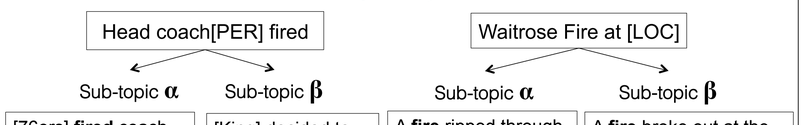


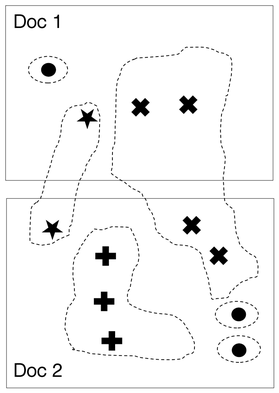
Pure-CDEC isolates cross document decisions -- a clearer approach to evaluate cross document performance than Simple-CDEC.

The Evaluation Corpus

ECB+ Corpus: sole corpus with cross document event coreference annotations.

Corpus Layout





● singleton events.
✚ within-document events.
✖ cross-document events.

	Train	Dev	Test
Docs	462	73	447
Topics	20	3	20
Sub-topics	40	6	40
Event mentions	5443	608	8951
Singletons	**1866**	**167**	**5572**
WD event	**23**	**0**	**89**
CD event	**3554**	**441**	**3290**
WD chain	2502	316	2138
CD chain	691	47	479

Experimental Results

Evaluating Lemma Baseline under Different Evaluation Setups

Evaluation		MUC P	MUC R	MUC F	B3 P	B3 R	B3 F	CEAF P	CEAF R	CEAF F	BLANC PW	BLANC PWN	BLANC F	Avg.	Avg.
With Singleton	Simple-CDEC	30.5	75.7	43.4	28.4	81.9	42.2	65.6	18.6	29.0	11.1	98.4	80.5	38.2	42.3
	B&H	37.4	75.3	50.0	49.4	81.6	61.5	39.9	70.8	51.0	30.5	93.4	80.5	54.2	56.1
	ST	40.9	75.9	53.2	58.8	82.7	68.7	71.3	45.7	55.7	37.0	95.4	66.2	59.2	61.0
Without Singleton	IS	79.5	76.1	77.8	50.7	54.0	52.3	39.9	46.7	43.1	31.0	98.4	64.7	57.7	59.5
	YCF (=IS+ST)	94.5	75.8	84.2	92.0	53.6	67.8	36.2	75.2	48.9	53.3	98.7	76.0	67.0	69.2

Existing evaluations are too lenient and there is a meaningful gap across setups.

Evaluating Different Baselines under Simple-CDEC and Pure-CDEC

We consider 3 other baselines in this comparison

- **Lemma-WD** − version of Lemma which only links mentions within the same document
- **Lemma-δ** − version of Lemma which links mentions across documents with TFIDF similarity > δ
- **SAC** − supervised greedy agglomerative clustering using binary SVM.

		MUC P	MUC R	MUC F	B3 P	B3 R	B3 F	CEAF P	CEAF R	CEAF F	BLANC PW	BLANC PWN	BLANC F	Avg.	Avg.
SIMPLE	Lemma-WD	38.0	20.4	26.5	88.7	68.4	77.2	67.5	80.8	73.6	53.0	98.5	51.9	59.1	57.3
	Lemma	30.5	75.7	43.4	28.4	81.9	42.2	65.6	18.6	29.0	11.1	98.4	54.8	38.2	42.3
	Lemma-δ	40.9	72.5	52.3	59.0	81.1	68.3	73.6	45.5	56.2	32.8	98.5	65.6	58.9	60.6
	SAC	44.2	52.9	48.2	75.2	76.0	75.6	70.5	62.6	66.3	28.6	98.5	63.5	63.4	63.4
PURE	Lemma-WD	0.0	0.0	0.0	90.1	65.6	75.9	67.0	80.2	73.0	0.0	76.2	38.1	49.6	46.8
	Lemma	18.7	62.7	28.8	27.2	74.8	39.9	65.7	18.6	29.0	72.0	76.2	41.7	32.6	34.9
	Lemma-δ	29.4	42.4	34.7	68.6	71.6	70.1	70.6	56.8	62.9	26.5	76.2	51.3	53.0	52.6
	SAC	30.5	42.0	35.3	70.6	74.5	72.5	71.2	63.2	67.0	25.3	79.0	52.2	58.3	56.7

Previous work claimed Lemma is a strong baseline for CDEC. However, Lemma-δ is a better baseline.

Annotation Quality of ECB+

Error analysis also revealed many annotation errors in ECB+.

Missing Singleton-to-Singleton Link
Two mentions of an event were marked as singletons.

Aceh was hit by the massive Boxing Day **earthquake** and **tsunami** in 2004 …

A massive **quake** struck off Aceh in 2004 , sparking a **tsunami** that …

Missing Singleton-to-Cluster Link
A mention of an event was left out of the true event cluster.

co-refers to another mention

LaRue , who was found guilty of participating in the **Watergate coverup** …

… then a significant participant in the **Watergate scandal** .

Identifying all such errors manually will be tedious ($O(n^2)$ cases).

Semi-automatic Annotation Error Identification

To semi-automatically identify such errors, we proceeded as follows

- Assign every instance (mention pair) a unique ***anchor feature.***
- Train a linear classifier on this data and examine its weights.
- **High weight of anchor feature for instance A →**
  - **either instance A is genuinely hard or**
  - **it is a corpus error.**
- Ask annotator to examine these cases only (<< n^2).

We found over 300 event mentions which had above errors.

Detected Errors at http://cogcomp.cs.illinois.edu/page/publication_view/801

References

- Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In ACL.
- Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. TACL.
- Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In LREC.
- Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and easy entity evaluation. In ACL.

Acknowledgements

Thanks to Bishan Yang, Piek Vossen and Joel Nothman for answering questions and sharing system outputs. This work was supported by Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.