How Good is the Model in Model-in-the-loop Event Coreference Resolution Annotation?

Shafiuddin Rehan Ahmed[1] Abhijnan Nath[2] Michael Regan [3]
Adam Pollins [1] Nikhil Krishnaswamy [2] James H. Martin [1]

[1]University of Colorado, Boulder, CO, USA

[2]Colorado State University, Fort Collins, CO, USA

[3]University of Washington, Seattle, WA, USA

The 17th Linguistics Annotation Workshop
at the Association of Computational Linguistics
ACL, July 2023

University of Colorado Boulder

Introduction

Event Coreference Resolution (ECR) is the task of identifying mentions of the same event either within or across documents. Consider the following examples:

ECR vs Non-ECR Examples

e_1: 55 year old star will *replace*$_{m_1}$ Matt Smith, who announced in June that he was leaving the sci-fi show.
e_2: Matt Smith, 26, will make his debut in 2010, *replacing*$_{m_2}$ David Tennant, who leaves at the end of this year.
e_3: Peter Capaldi *takes over*$_{m_3}$ Doctor Who ... Peter Capaldi *stepped into*$_{m_4}$ Matt Smith's soon to be vacant Doctor Who shoes.

ECR Links: (m_1, m_3), (m_1, m_4), (m_3, m_4)

Non-ECR Links: (m_2, m_1), (m_2, m_3), (m_2, m_4)

Annotating ECR Links

Traditional Methodology

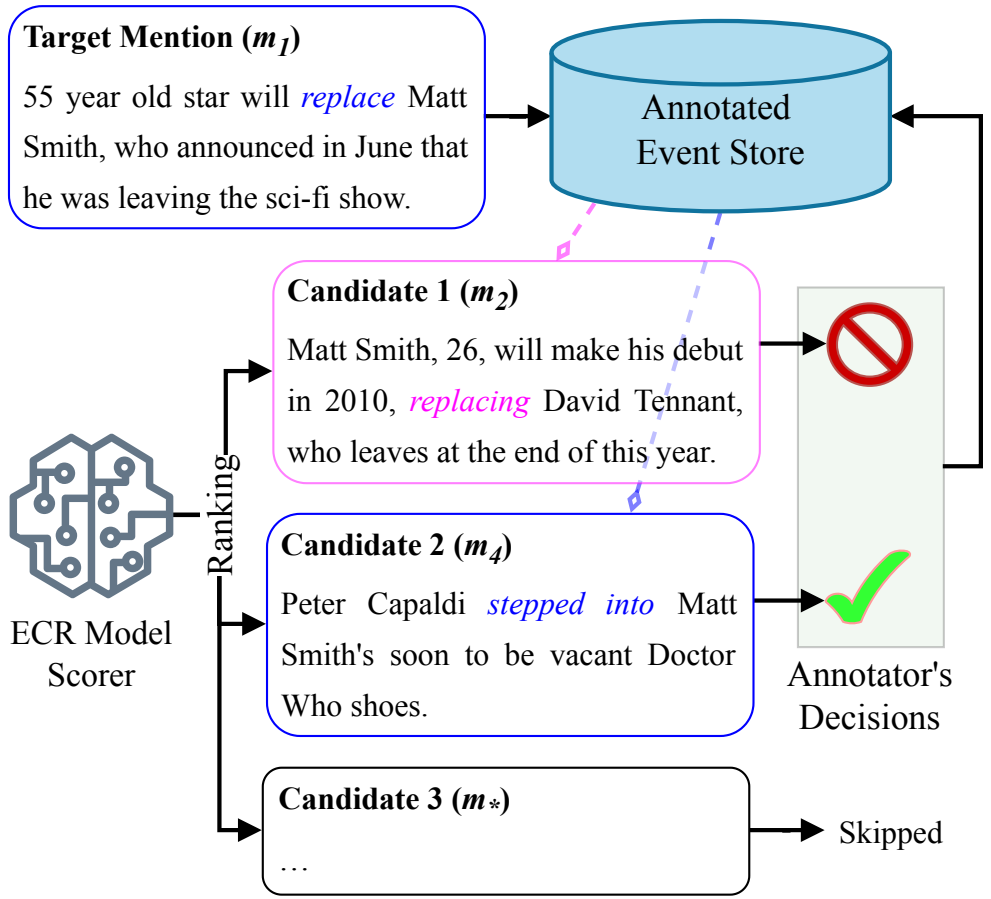
$O(n^2)$ fully manual comparison of all the mention pairs.

Challenging, time consuming and prone to errors!

(1) storage and retrieval of annotated event clusters for a new target event

(2) ML model that ranks and prunes candidate clusters

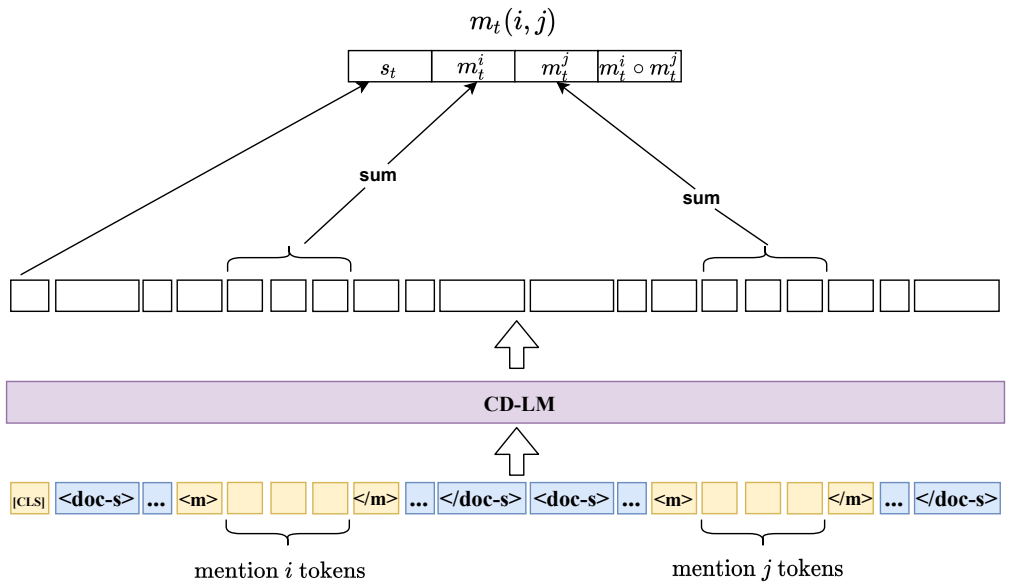
Our Model-in-the-loop Methodology



Ranking Methods: Cross-encoder (CDLM[1])

$$m_t(i,j) = \left\langle s_t, m_i^i, m_j^j, m_i^i \odot m_j^j \right\rangle$$

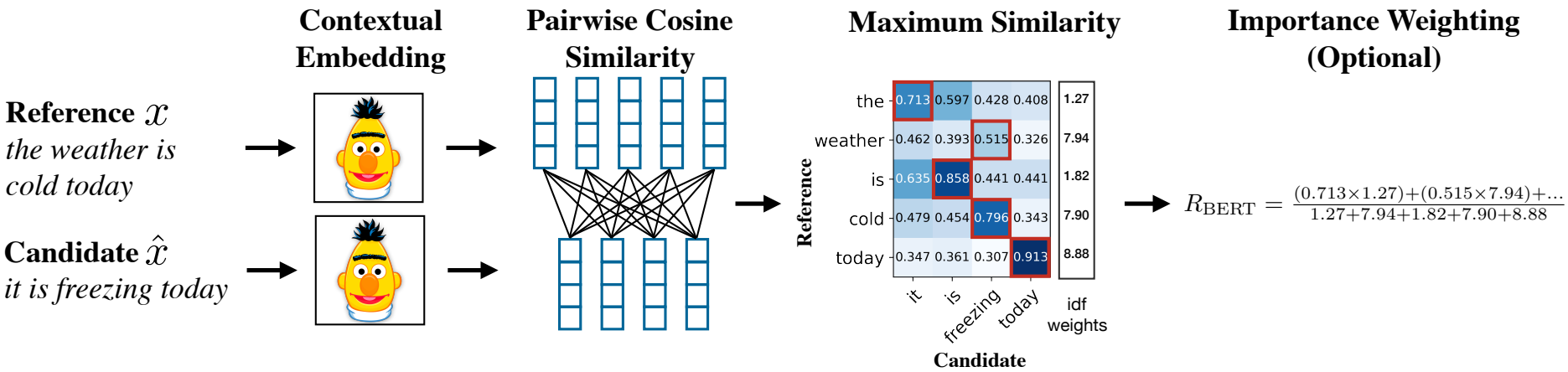
$$\texttt{CDLM}(m_i, m_j) = \texttt{mlp}(m_t(i,j))$$



- Cross-document Language Model
- Longformer-based
- **Compute Intensive** (GPU - RTX 3090 24GB)
- State of the Art in ECR performance

[1] Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-Document Language Modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*

Ranking Methods: BERTScore[2] (BERT)



- `distilbert – base` makes it **Low Compute** (CPU)
- $\text{BS} = F_{\text{BERT}}, \quad \text{S}_{\text{bert}}(m) = \langle t_m, \text{[SEP]}, S_m \rangle$
- $\text{BERT}(m_i, m_j) = \lambda\ \text{BS}(t_{m_i}, t_{m_j}) + (1-\lambda)\ \text{BS}(\text{S}_{\text{bert}}(m_i), \text{S}_{\text{bert}}(m_j))$

For a mention, m, t_m is the mention text of the event trigger and S_m is the mention's sentence.

λ is a hyper-parameter we estimate to be 0.7 for BERT.

[2] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. and Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675.*

Ranking Methods: Baselines (Lemma, Random)

Lemma Similarity (Lemma)

A weighted average of the token overlap (JS) between mentions' triggers, and the mentions' sentences

$$\text{Lemma}(m_i, m_j) = \lambda \ \text{JS}(t_{m_i}, t_{m_j}) + (1 - \lambda) \ \text{JS}(S_{m_i}, S_{m_j})$$

λ is 0.7 for Lemma as well.

No Ranking (Random)

A random baseline to serve as a reference point to compare the other methods.

Evaluation Methodology

Annotation Effort - `Comparisons`

Total `Comparisons` between Target and Candidate pairs

Annotation `Recall`

The ratio of `Comparisons` between Target and Coreferent-Candidate pairs

Estimating the metrics through **Simulation**:

Datasets (Dev and Test)

- **ECB+**: Event Coreference Bank+
- **GVC**: Gun Violence Corpus

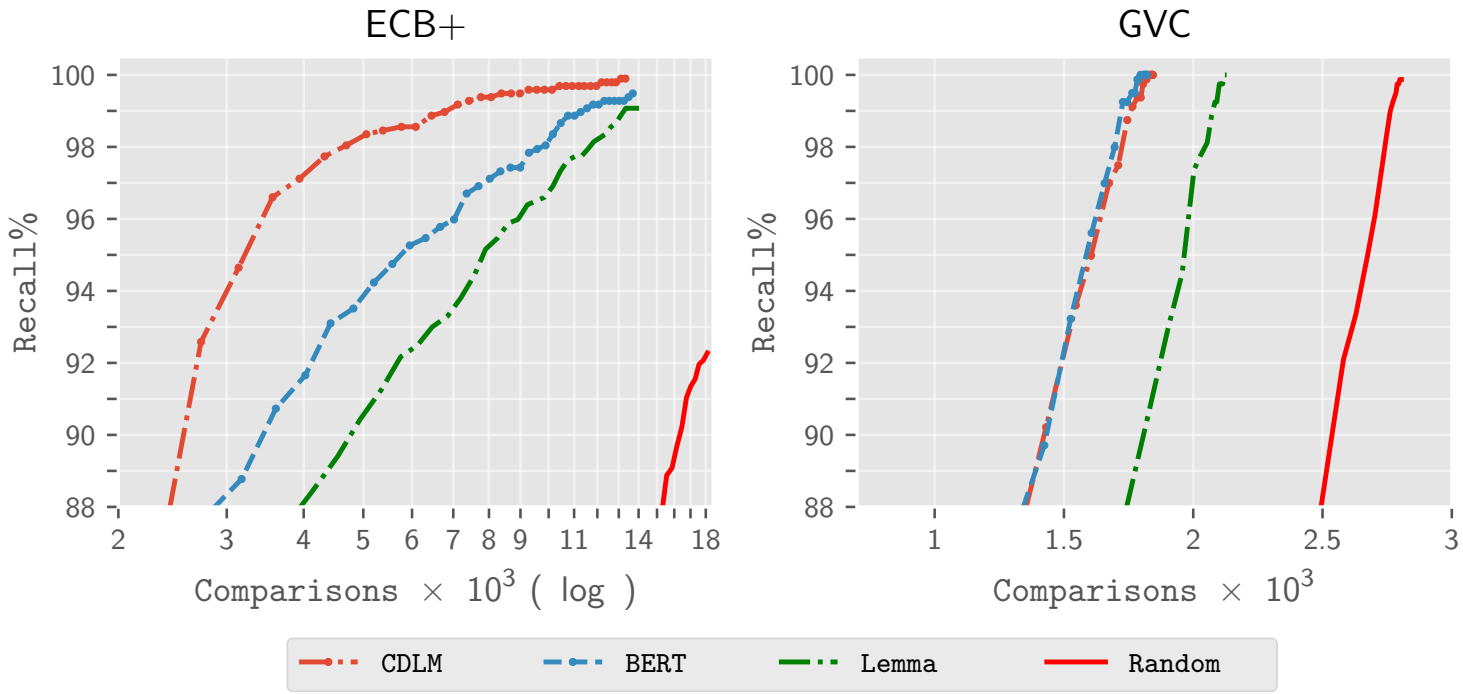
Candidates Sampling

- Same topic
- Select top-k ranked ones

Incremental Clusterting

- Using ground-truth
- Count ECR `Comparisons` and All `Comparisons`

Analysis: Recall-Annotation Effort Tradeoff



We fix `Recall` (e.g., 97%), and analyze `Comparisons` for each method.

Key Takeaways about the Ranking Methods

`CDLM` >> `BERT` > `Lemma` >>> `Random` for ECB+ (diverse dataset)

`CDLM` $\geq$ `BERT` > `Lemma` > `Random` for GVC (less diverse dataset)

`BERT` > `CDLM` in Efficiency and Generalizability!

Annotation Interface Demo

Interface

Implemented using Prodigy Annotation Tool for ease of integrating model-in-the-loop annotation methodologies

Conclusion

- We introduced a model-in-the-loop annotation method for annotating ECR links.
- We introduced a methodology to evaluate Recall-Annotation Effort Traderoff
- We compared three simulated ranking models differing in complexities using the evaluation methodology and showed their viability for this task

Paper

Github

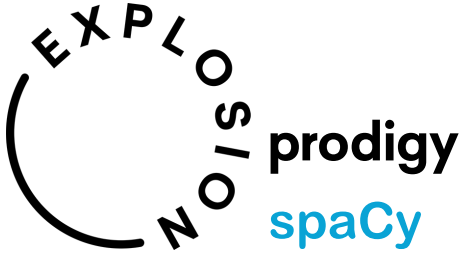
Author

Acknowledgements

The anonymous reviewers ☺!

U.S. Defense Advanced Research Projects Agency (DARPA)
Grant: FA8750-18-2-0016-AIDA – RAMFIS: Representations of vectors and Abstract Meanings For Information Synthesis.

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the U.S. government.



Thank you for watching the presentation!