Realistic Evaluation Principles for Cross-document Coreference Resolution

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi and Ido Dagan

Essence

Recent work on Cross-Document (CD) coreference resolution have used **permissive** and **inconsistent** evaluation protocols, leading to an overestimation of performance.

To allow realistic assessment, we establish a more **rigorous evaluation methodology.**

Cross-document coreference in a nutshell

*News that **Barack Obama** may **name** **Dr. Sanjay Gupta** of Emory University and CNN as **his** Surgeon General has caused a spasm of celebrity reporting…*

*CNN's management confirmed yesterday that **Dr Gupta** had been **approached** by the **Obama** team…*

***President Obama** will **name** **Dr. Regina Benjamin** as U.S. Surgeon in a Rose Garden announcement late this morning…*

***Obama nominates** new surgeon general: MacArthur "genius grant" fellow **Regina Benjamin**…*

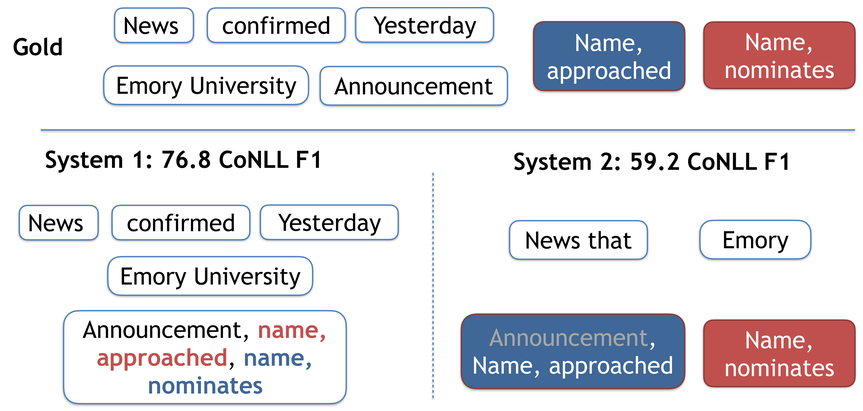
* Examples of two ***subtopics*** from ECB+, illustrating lexical **diversity** and **ambiguity.**

Unrealistic Evaluation Aspects

1. CD coreference models are evaluated **only on gold mentions**
2. Singletons artificially **inflate** results
3. Models cluster the documents into **fine-grained subtopics**

Make the task easier, hence reported results do not reflect actual performance

1. Singleton Effect



S1 is good on mention detection but worse on **coreference** links, but achieve better **coreference** results than S2 –> counterproductive for downstream tasks

2. Decoupling Coreference Evaluation

Our proposal: **decouples** the evaluation to have more faithful results

1. Mention detection — span detection tasks (if singletons are annotated)
2. Coreference link predictions without singletons, as in CoNLL-2012

	Mention Detection (Span F1)	Coreference link predictions (CoNLL F1)
S1	100	57.5
S2	66.7	86.5

3. Confronting Lexical Ambiguity

- The Inclusion of subtopics in ECB+ aims to challenge **models** with lexical ambiguity
- Recent works bypass this challenge by clustering into fine-grained **subtopics**

We propose that models evaluate at the level of the entire topic, without subtopic clustering

Results - ECB+

	CoNLL F1
Barhom et al. (2019)	79.5
Cattan et al. (2021)	**81.0**

*Table 2: **Subtopic** level, gold mentions with singletons*

	CoNLL F1
– singletons	71.1 **(-9.9)**
– topic level	62.0 **(-9.1)**
– gold mentions	48.6 **(-13.4)**

Table 3: Effect of our realistic methodologies on the results on Cattan et al. (2021)

–32.4 F1

- Singletons artificially **inflate** results
- Topic level (-9.1) —> **Ambiguity challenge** posed by ECB+ subtopics is not yet solved by current models

Conclusion

Realistic and rigorous evaluation methodology

1. ***Predicted mentions***
2. ***Decouple mention detection and coreference resolution.***
3. ***Confront lexical ambiguity challenge (topic level)***

Large room for improvement in realistic settings!

References

Cybulska, A., & Vossen, P. (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. LREC.

Barhom, Shany, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. ACL 2019.

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., & Dagan, I. (2021). Cross-document Coreference Resolution over Predicted Mentions. Findings of ACL

https://github.com/ariecattan/coref