Online Neural Coreference Resolution with Rollback

Patrick Xia and Benjamin Van Durme

human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY

Background: Example

Director	This is a terrible play! I'll see you in the morning.
Kate	I can't believe we go in, in a week.
Joey	Hey, it's gonna be all right.
Lauren	Hey! So, since we're getting off early, do you want to go paint mugs?
Joey	Y'know what, I kinda need to work on my stuff tonight.
Lauren	Oh, okay. I'll see you tomorrow. G'night.
Joey	Ah, are you okay?
Kate	Yeah, I guess.

Background: Example

Director	This is a terrible play! I'll see you in the morning.
Kate	I can't believe we go in, in a week.
Joey	Hey, it's gonna be all right.
Lauren	Hey! So, since we're getting off early, do you want to go paint mugs?
Joey	Y'know what, I kinda need to work on my stuff tonight.
Lauren	Oh, okay. I'll see **you** tomorrow. G'night.
Joey	Ah, are you okay?
Kate	Yeah, I guess.

Background: Example

Director	This is a terrible play! I'll see you in the morning.
Kate	I can't believe we go in, in a week.
Joey	Hey, it's gonna be all right.
Lauren	Hey! So, since we're getting off early, do you want to go paint mugs?
Joey	Y'know what, I kinda need to work on my stuff tonight.
Lauren	Oh, okay. I'll see you tomorrow. G'night.
Joey	Ah, are **you** okay?
Kate	Yeah, I guess.

Background: Questions

1. How well can models perform without *future context*?

2. Can we *reduce latency* of predictions in an online manner?

3. What does *recovery* from mispredictions look like?

Method: Datasets

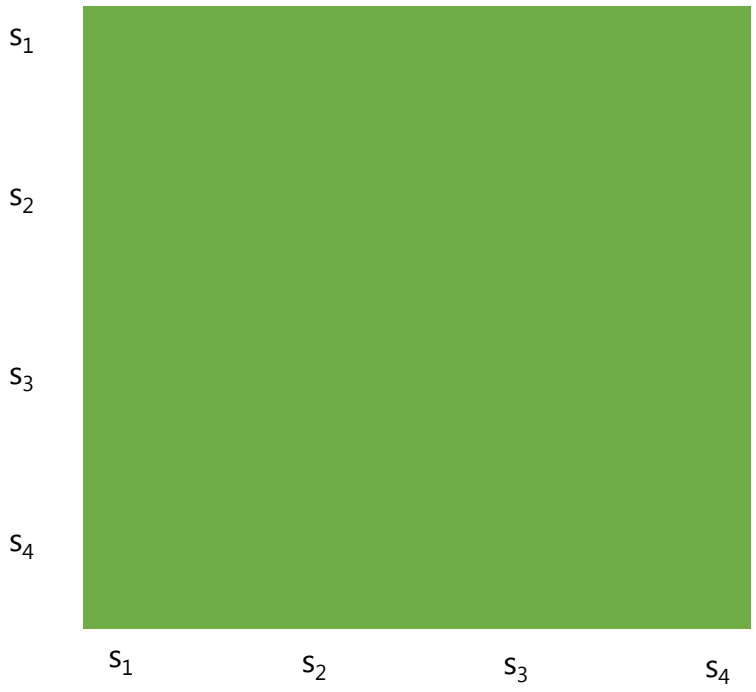
- OntoNotes (General)
  - "Conversational"
  - "Text"
- Character Identification (TV Dialogue – *Friends*)
- LitBank (Literature – Long)
- QBCoref (Trivia questions – short)

Method: Metrics

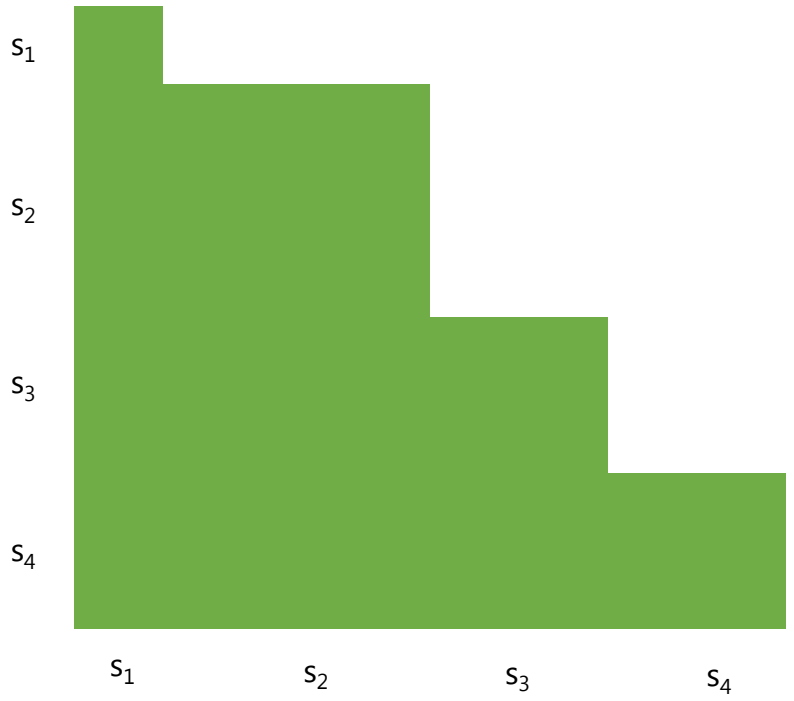
- Average CoNLL F1 (“final F1”)
- Running CoNLL F1 (“running F1”)
  - Compute MUC, B_3, $CEAF_e$ for every prefix of a *single* document
  - Macro average across corpus
- Wait Time
  - Time between end of document and end of predictions

Question 1: Masking the future

- Hide the future sentences from SpanBERT attention



Full attention



Masked at the sentence-level

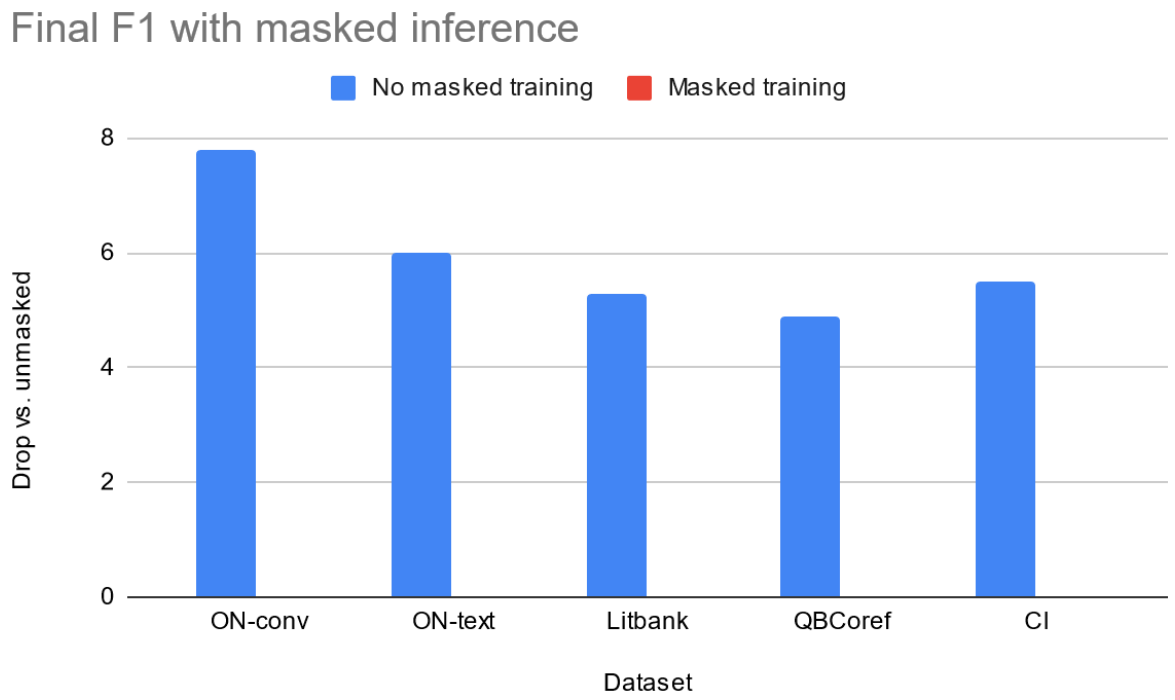
Question 1: Masking the future

- Models:
  - C2F: coarse-to-fine model (Xu and Choi 2020, 2021, Lee et al., 2018)
  - ICoref: incremental clustering model (Xia et al., 2021)
- Continued pretraining on the smaller datasets

- Train with mask?

Results 1: Masking the future

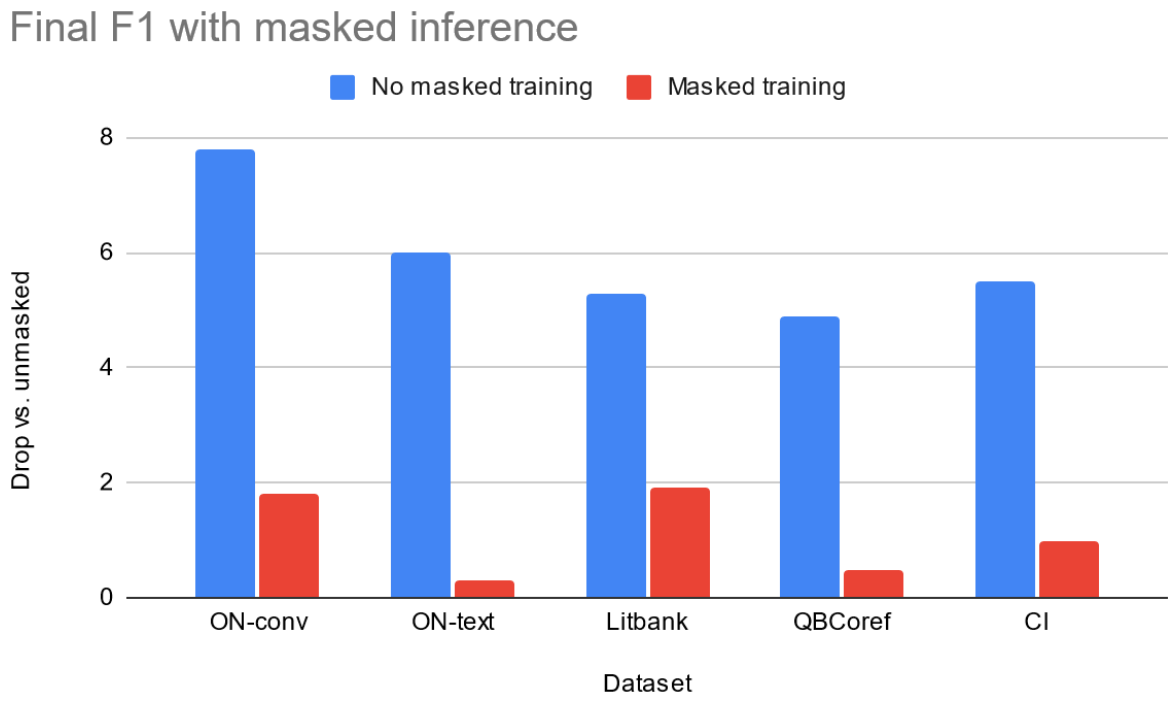
- Large drops with masked inference
- Recover by training with mask



C2F model. Similar drops with ICoref

Results 1: Masking the future

- Large drops with masked inference
- Recover by training with mask



C2F model. ICoref saw less benefit from masked training

Question 2: Online Evaluation

- Report *running F1* and *wait time*
- Models:
  - Naive online C2F: run full C2F model after every sentence
  - Online (sentence-level) ICoref: only process additional sentence

Results 2: Online Inference

LitBank: naive online C2F vs. online ICoref

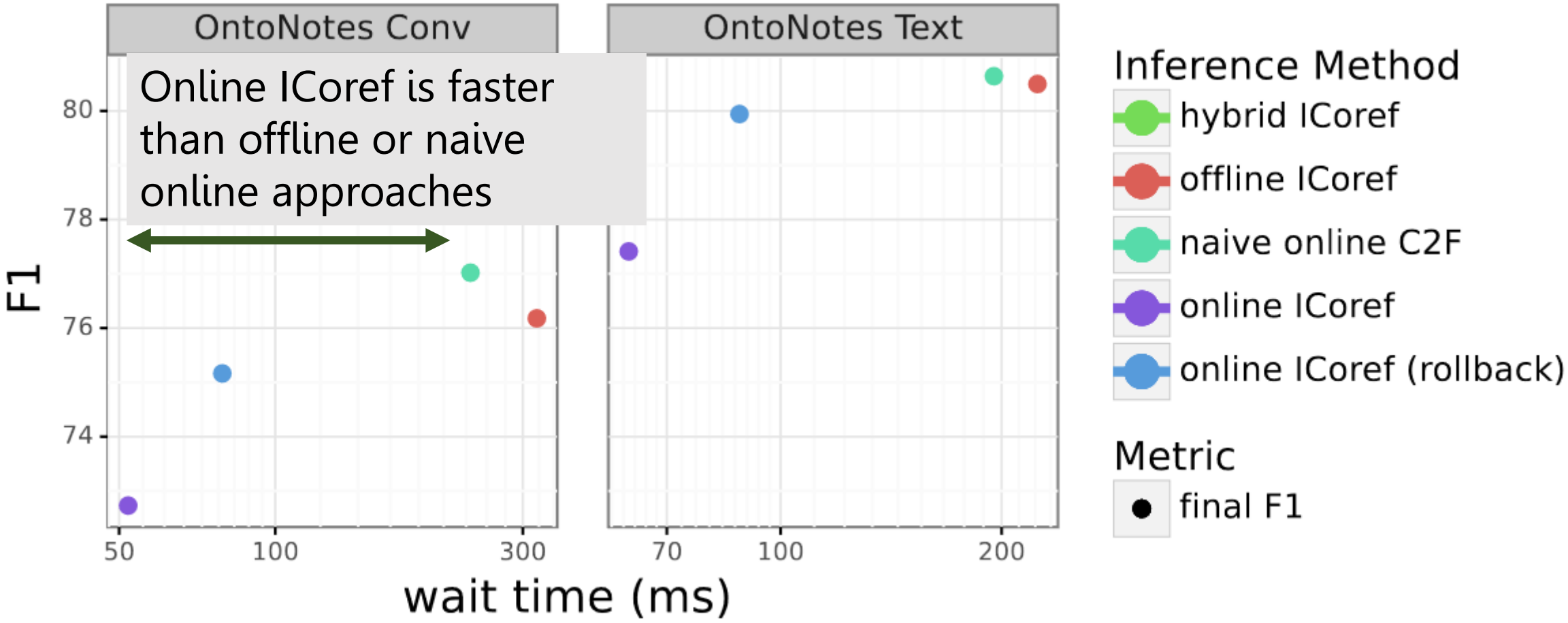
1. Substantially faster: 807ms > **74ms**
2. Worse final F1: **72.2** > 70.6
3. Worse running F1: **73.8** > 71.9

Question 3: Recovery with rollback

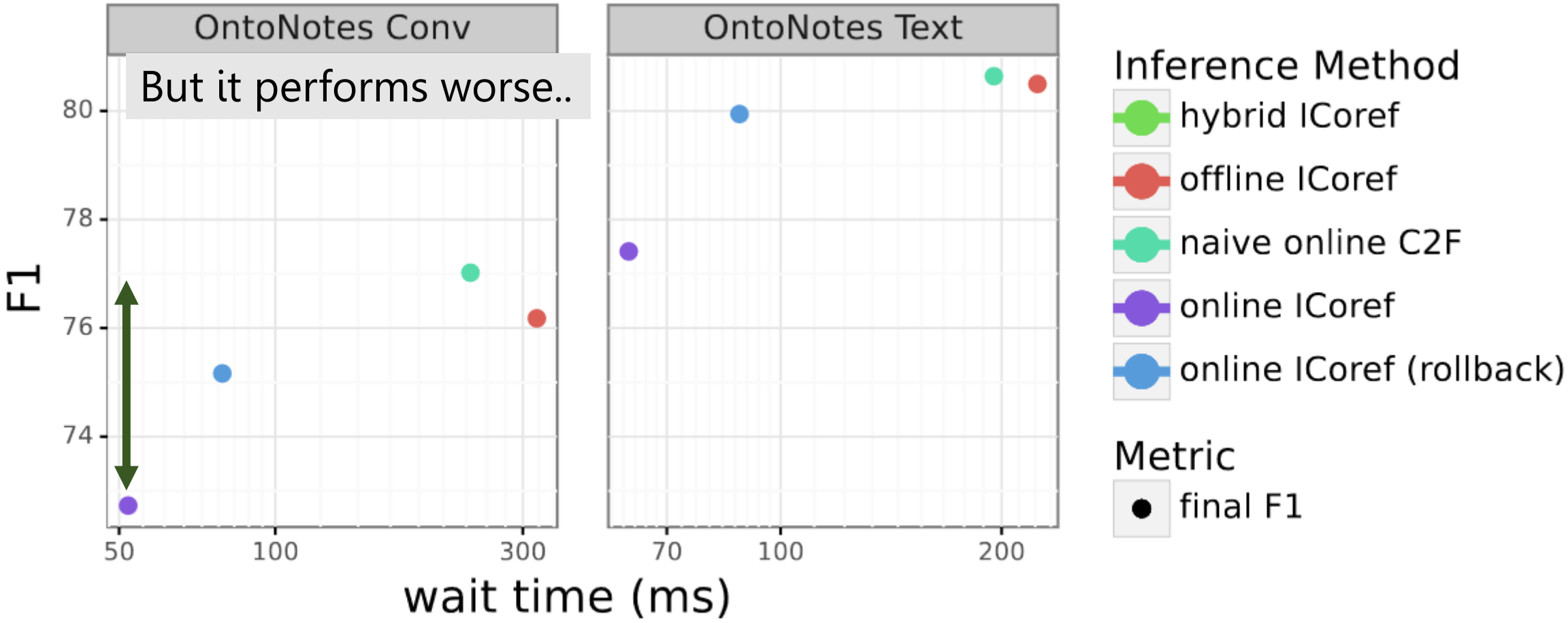
- Proposal: recovery mechanism to re-make earlier decisions
- Sentence-level Incremental Coref + Rollback:

```
for sentence t :
    if t mod r = 0:
        revert(r)  # undo predictions from last r sentences
        ICoref([s_{t-r+1}, s_{t-r+2}, …, s_t])
    else: ICoref([s_t])
```

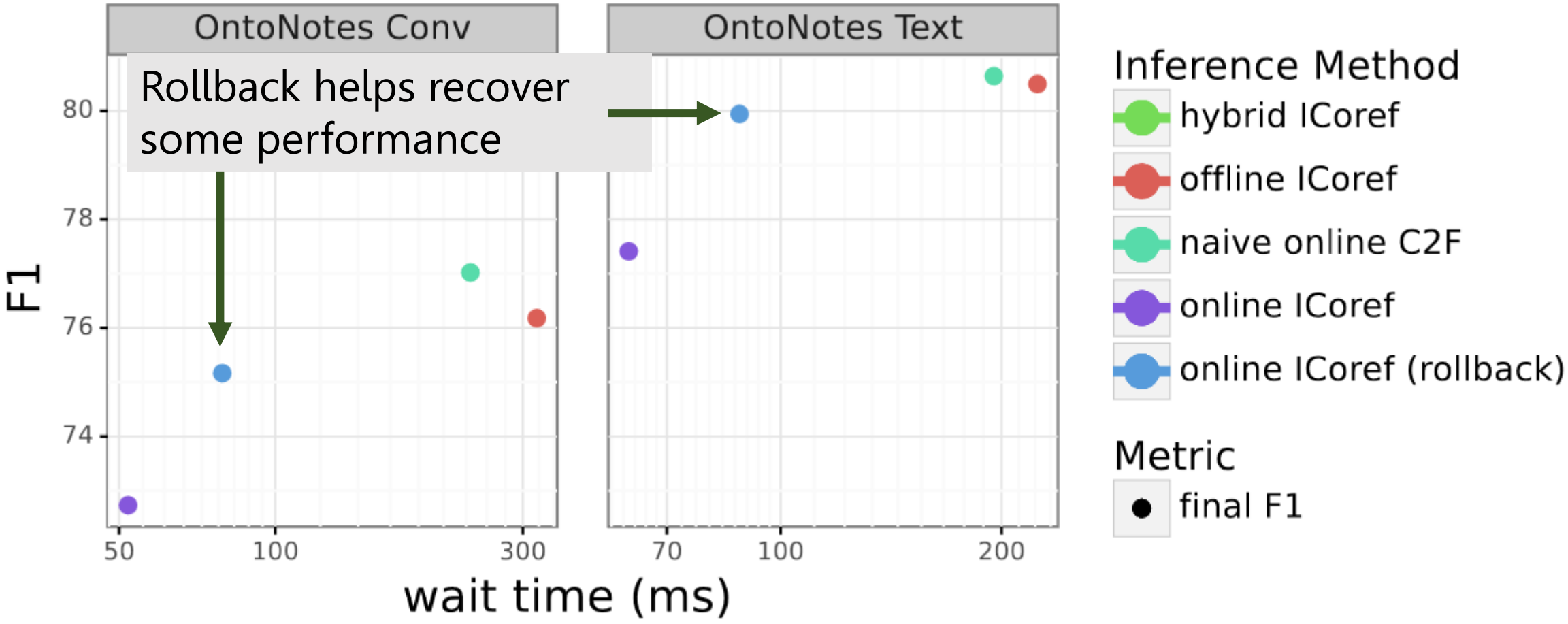
Results 3: Comparison of methods



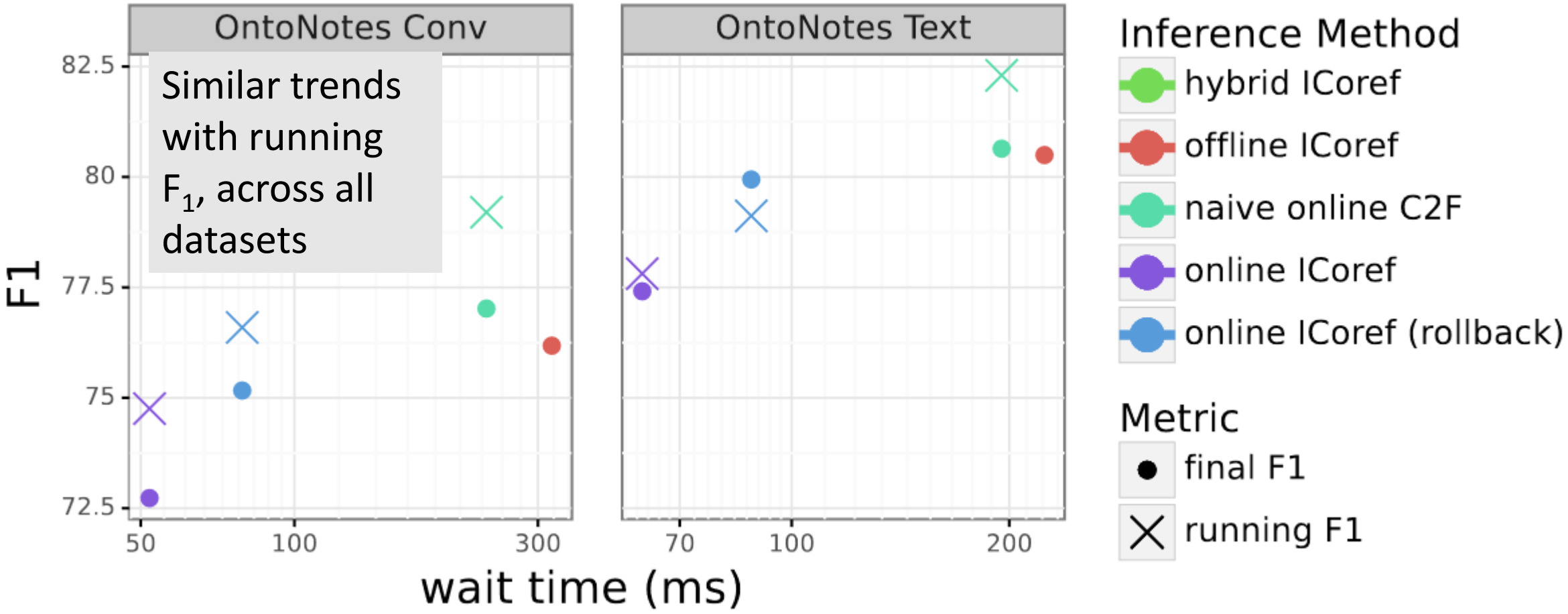
Results 3: Comparison of methods



Results 3: Comparison of methods



Results 3: Comparison of methods



Analysis: Rollback and correction analysis

- Analysis of edited links for 3 datasets: LitBank (books), QBCoref (trivia questions), and CI (*Friends* transcripts)

LitBank:

- 51.9% of edits correct mistakes, mostly to existing clusters
- 25.7% of edits introduce mistakes, related to mention detection

Analysis: Example

He looked round and lowered his voice.
"I'm carrying papers—vitally important papers.
They make all the difference to the Allies in the war.
You understand?
These papers have GOT to be saved!
They've more chance with you than with me.
Will you take them?"

Dialogue is a challenging domain

Analysis: Example

He looked round and lowered his voice.

"I'm carrying papers—vitally important papers.

They make all the difference to the Allies in the war.

You understand?

These papers have GOT to be saved!

They've more chance with you ~~than with me~~.

Will you ~~take them?"~~

The girl held out her hand

Rolling back can help, but doesn't fix everything

Conclusions

- Let’s consider the online setting for coreference resolution!!
- What’s important
  - Latency?
  - Final accuracy?
  - Running accuracy?
- Steps to address the new setting
  - Masked training
  - Sentence-level models
  - Rollback

Thank you

https://github.com/pitrack/incremental-coref/

human language technology
center of excellence

JOHNS HOPKINS
UNIVERSITY