On Generalization in Coreference Resolution

Shubham Toshniwal

Patrick Xia Sam Wiseman Karen Livescu Kevin Gimpel

`github.com/shtoshni92/fast-coref`

Motivation

- Current coreference resolution models generalize poorly to out-of-domain evaluations (Moosavi and Strube 2018; Subramanian and Roth 2019; Zhu et al 2021)

- Previous work does limited out-of-domain evaluations

Motivation

- Current coreference resolution models generalize poorly to out-of-domain evaluations (Moosavi and Strube 2018; Subramanian and Roth 2019; Zhu et al 2021)
  - Develop coreference resolution models with better generalization
- Previous work does limited out-of-domain evaluations
  - Evaluate models over a bigger suite of datasets

Evaluation Suite

Datasets
OntoNotes
PreCo
LitBank
Character Identification
WikiCoref
QuizBowl
WSC
GAP

Train, Test

Evaluation Suite

Datasets	Domain
OntoNotes	News, Conversations, Web
PreCo	School books
LitBank	Literary text
Character Identification	TV show transcripts
WikiCoref	Wikipedia
QuizBowl	Quizbowl
WSC	-
GAP	Wikipedia

Train, Test

Evaluation Suite

Datasets	Domain	Full Coreference
OntoNotes	News, Conversations, Web	✓
PreCo	School books	✓
LitBank	Literary text	✓
Character Identification	TV show transcripts	✓
WikiCoref	Wikipedia	✓
QuizBowl	Quizbowl	✓
WSC	-	✗
GAP	Wikipedia	✗

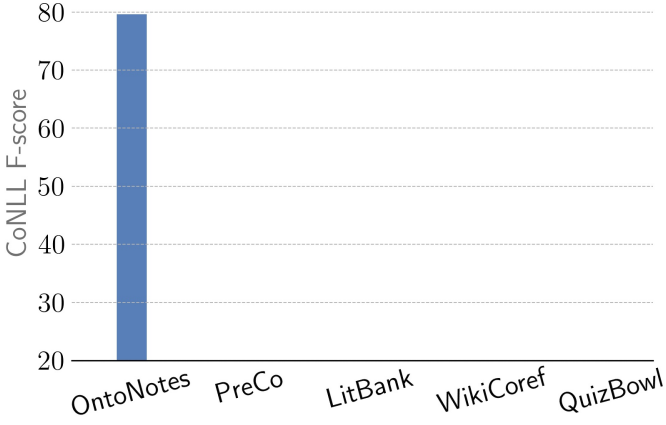
Train, Test

Evaluation Suite

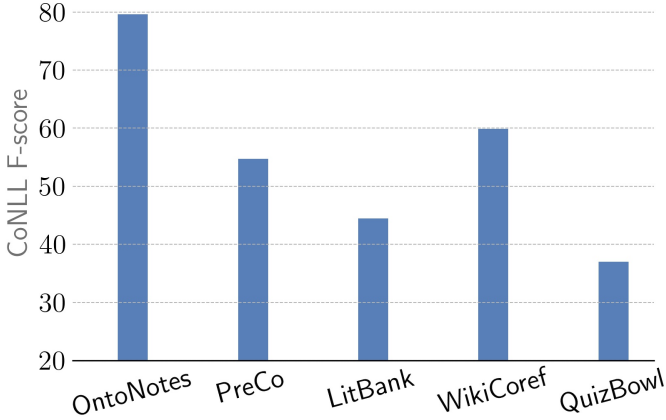
Datasets	Domain	Full Coreference	Singletons
OntoNotes	News, Conversations, Web	✓	✗
PreCo	School books	✓	✓
LitBank	Literary text	✓	✓
Character Identification	TV show transcripts	✓	✓
WikiCoref	Wikipedia	✓	✗
QuizBowl	Quizbowl	✓	✓
WSC	-	✗	N/A
GAP	Wikipedia	✗	N/A

Train, Test

Out-of-Domain Evaluation



Out-of-Domain Evaluation



79.6 for OntoNotes $\longrightarrow$ 37.0 for QuizBowl!

Challenges in Generalization: Domain Shift

News vs Literary Text

OntoNotes [The Federal Reserve]$_1$ considers interest rates ... On [Tuesday]$_2$, the Federal Reserve Open Market Committee meets for the final time this year to discuss interest rates ...

LitBank Down the [Rabbit-Hole]$_1$ [Alice]$_2$ was beginning to get very tired of sitting by [[her]$_2$ sister]$_3$ on [the bank]$_4$, and of having nothing to do: once or twice [she]$_2$ had peeped into ...

Challenges in Generalization: Annotation Differences

Singletons are not annotated in OntoNotes!

OntoNotes [The Federal Reserve]$_1$ considers interest rates ... On [Tuesday]$_2$, the Federal Reserve Open Market Committee meets for the final time this year to discuss interest rates ...

LitBank Down the [Rabbit-Hole]$_1$ [Alice]$_2$ was beginning to get very tired of sitting by [[her]$_2$ sister]$_3$ on [the bank]$_4$, and of having nothing to do: once or twice [she]$_2$ had peeped into ...

Challenges in Generalization: Annotation Differences

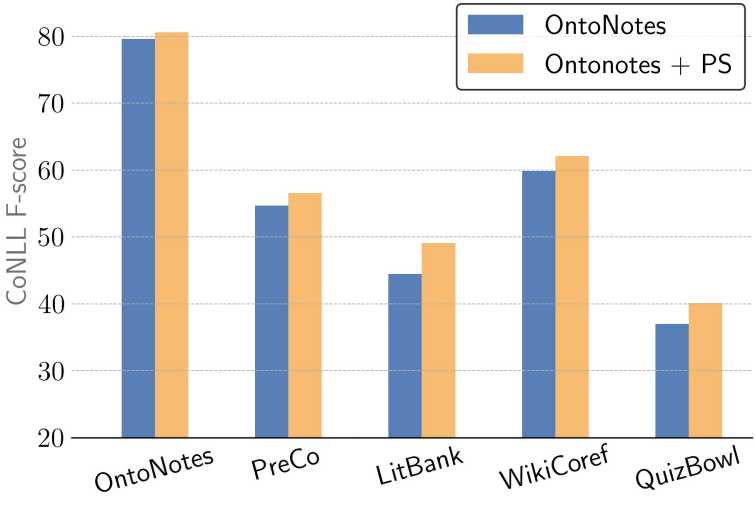
Singletons are not annotated in OntoNotes!

Add Pseudo-Singletons!

OntoNotes `[The Federal Reserve]`$_1$ `considers interest rates ... On [Tuesday]`$_2$`, the Federal Reserve Open Market Committee meets for the final time this year to discuss interest rates ...`

LitBank `Down the [Rabbit-Hole]`$_1$ `[Alice]`$_2$ `was beginning to get very tired of sitting by [[her]`$_2$ `sister]`$_3$ `on [the bank]`$_4$`, and of having nothing to do: once or twice [she]`$_2$ `had peeped into ...`

Impact of Adding Pseudo-Singletons

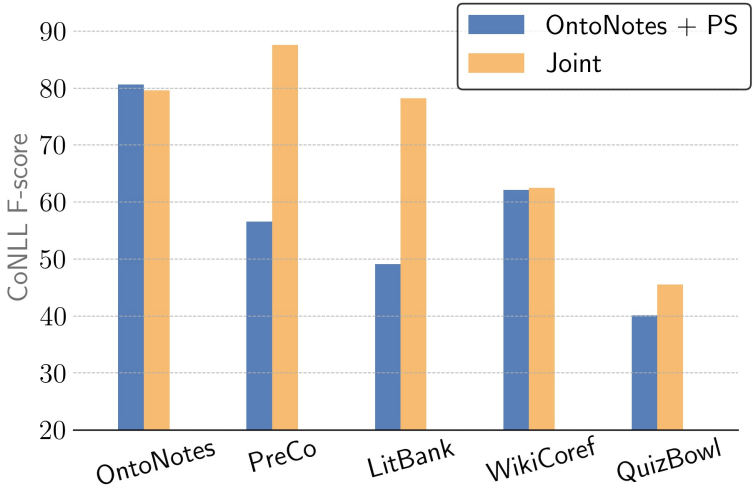


Improvements across the board!

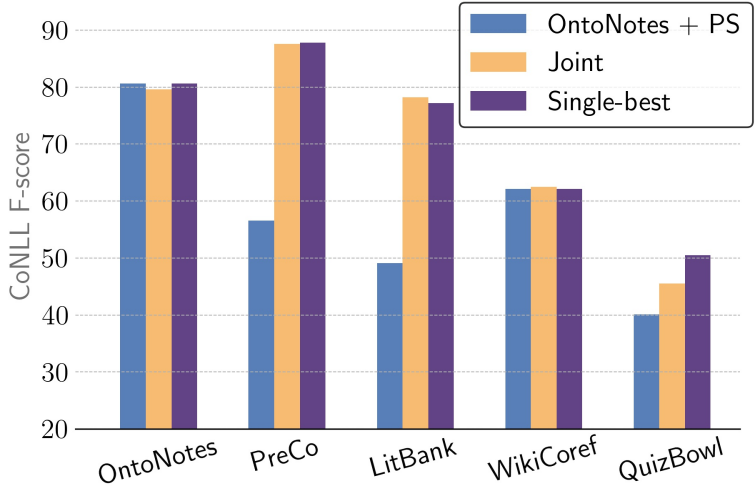
Joint Training

- Joint training with OntoNotes, PreCo, and LitBank
- Dataset-identity of examples is not used
- Baseline: Single-Best = $\max\{$OntoNotes-only, PreCo-only, LitBank-only$\}$

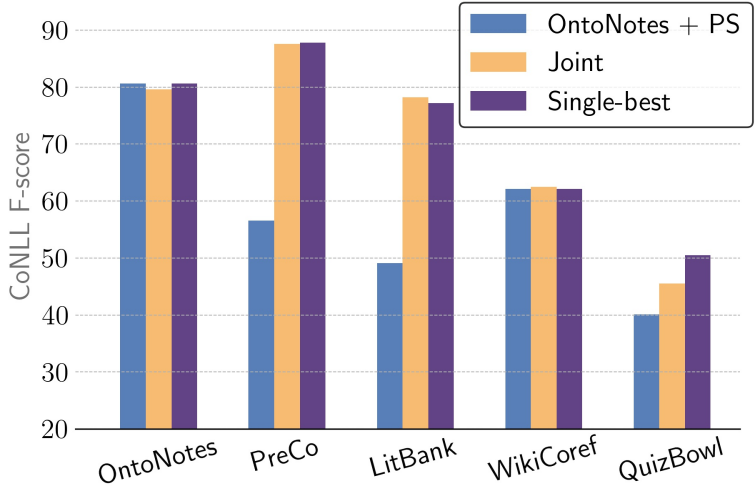
Impact of Joint Training



Impact of Joint Training

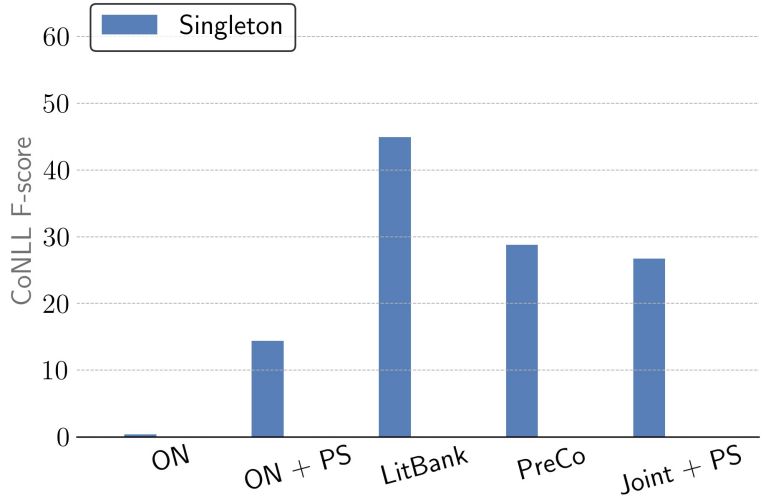


Impact of Joint Training

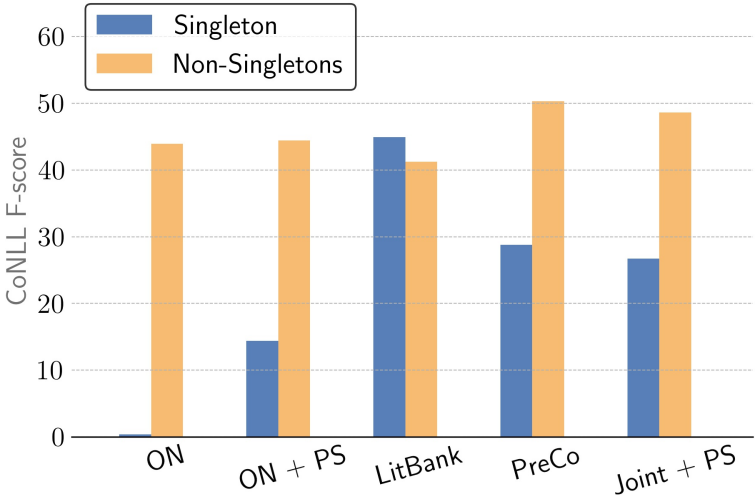


Jointly trained model performs as well as an oracle-ensemble of three models!

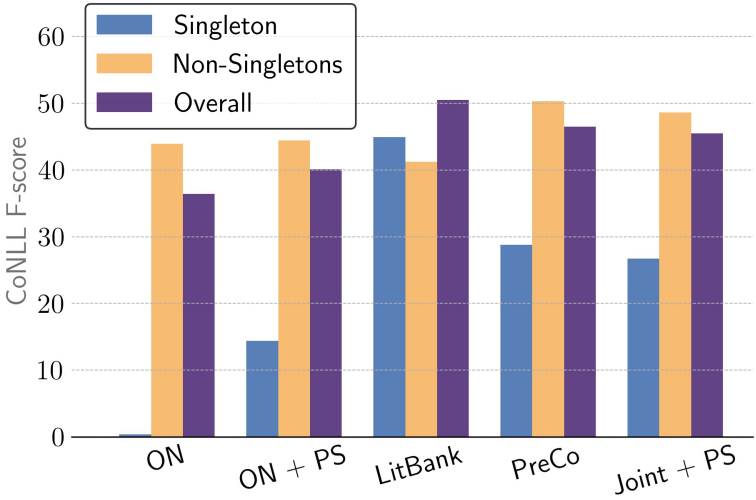
Impact of Singletons in QuizBowl



Impact of Singletons in QuizBowl



Impact of Singletons in QuizBowl



Takeaways

- Propose a evaluation suite consolidating eight-datasets for tracking generalization
- Improve generalization by: (a) data augmentation with pseudo-singletons, and (b) joint training
- State-of-the-art results for LitBank and WikiCoref
- Google Colab demo of models available on our Github repo
  `github.com/shtoshni92/fast-coref`

Backup Slides

Mention Pruning

- Prior work uses top-$\mathcal{K}$ mentions where $\mathcal{K} = 0.4 \times \mathcal{T}$
  - High Recall, Low Precision
- Dependent on document length rather than content
- Use mention score threshold to filter mentions

$$\{x \in X \mid s_m(x) \geq 0\}$$

Mention Pruning

- Problematic for OntoNotes: Doesn't annotate singletons (false negatives)

> [Rafael Nadal] is the champion ~~[Roland-Garros]~~ for an unprecedented 13th time, [his] victory over [world No.1 Novak Djokovic] elevating [him] level with ~~[Roger Federer's]~~ all-time mark of 20 major titles. [The Spaniard] delivered one of [his] finest performances against arguably [[his] toughest rival] to prevail 6-0, 6-2, 7-5.

Error Analysis: QuizBowl

Missing Entities	This author's non fiction works … another work , a plague strikes secluded valley where teenage boys have been evacuated … name this author of Nip the Buds, Shoot the Kids
Extra Entities	This poet of "(**I**) felt a Funeral in (**my**) Brain" and "I'm Nobody, Who are you?" wrote about a speaker who hears a Blue, uncertain, stumbling buzz before expiring in "(**I**) heard a fly buzz when (**I**) died". For 10 points, name this female American poet of Because (**I**) could not stop for Death.