

Moving on from OntoNotes: Coreference Resolution Model Transfer

Patrick Xia and Benjamin Van Durme

Motivation

Coref: cluster spans of text that refer to the same entity

Poor coref transfer across:

- Annotation standards – include singleton clusters?
- Domain – which entity types are annotated?
- Language – is there any cross-lingual transfer?

And **Jo** shook **the blue army sock** till **the needles** rattled like **castanets**, and **her ball** bounded across **the room**.

All mentions

And **Jo** shook the blue army sock till the needles rattled like castanets, and **her** ball bounded across **the room**.

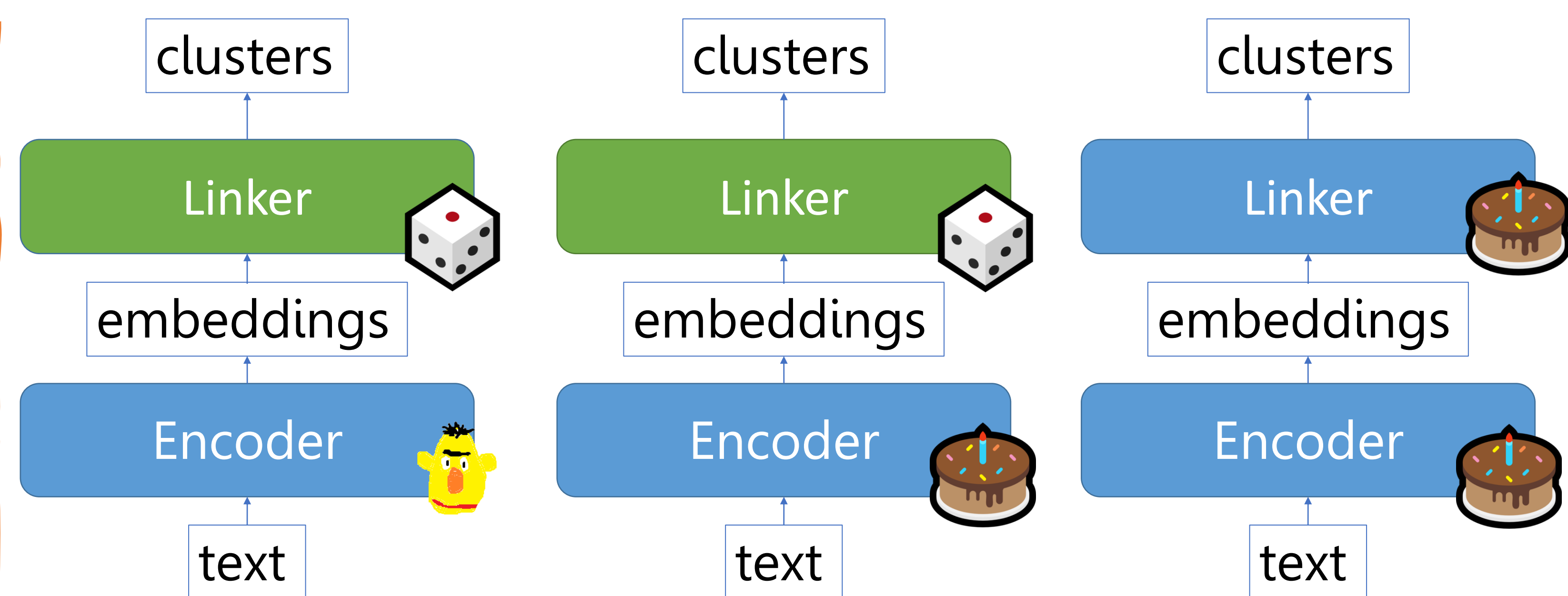
Literature

(1) In general, The term **"employer"** means with respect to **any calendar year**, **any person** who

Legal

Methods

Three model initialization methods:



1. Pretrained encoder only

2. Finetuned encoder on coref

3. Fully trained on source

Source datasets: OntoNotes, PreCo

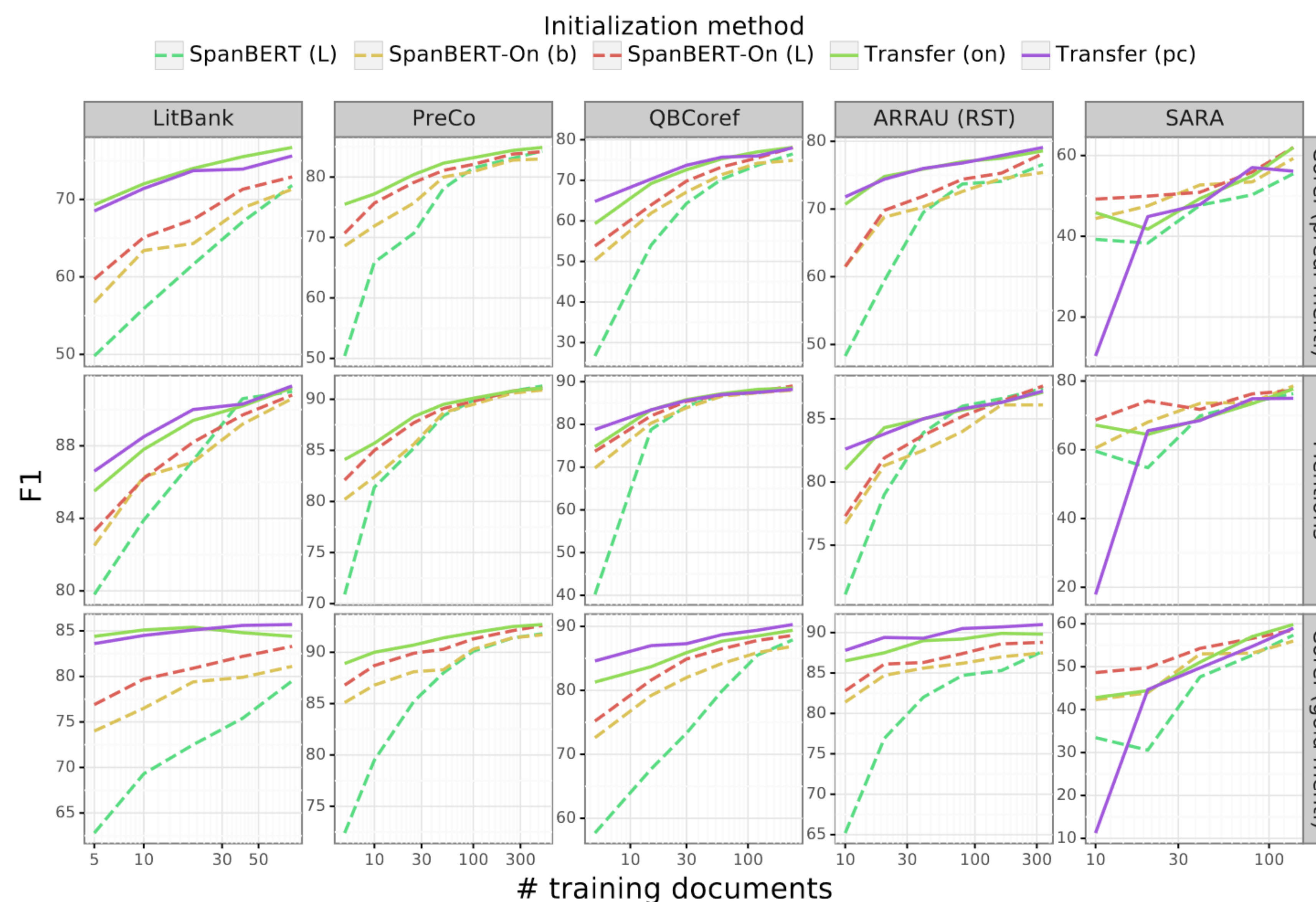
Target datasets:

English: PreCo, LitBank (books), ARRAU (news), SARA (legal), QBCoref (quiz)

Other Languages: OntoNotes (zh, ar), SemEval (ca, es, it, nl)

Encoders: SpanBERT, XLM-R

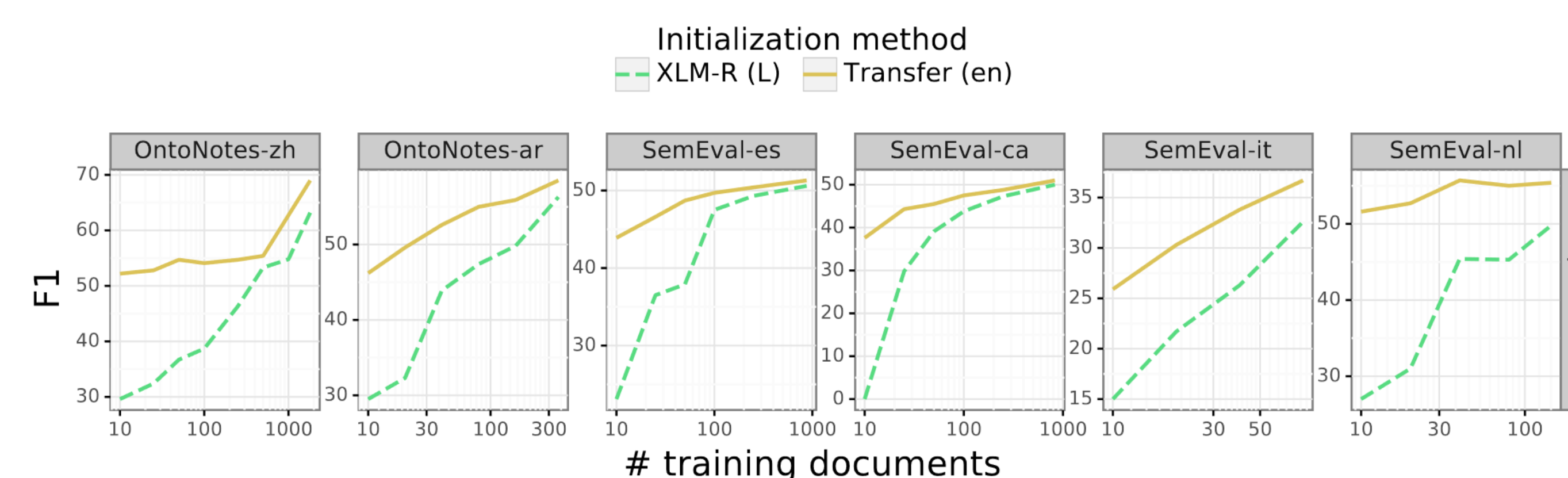
Results



- Transfer models usually outperform randomly initialized models
- PreCo is as effective as OntoNotes
- PreCo is better with gold mention boundaries
- Continued training of small (publicly available) encoders is effective with low # training docs

Additional Findings

1. Continued training is effective cross-lingually
New baseline numbers on several datasets



2. Allocate few docs for model selection

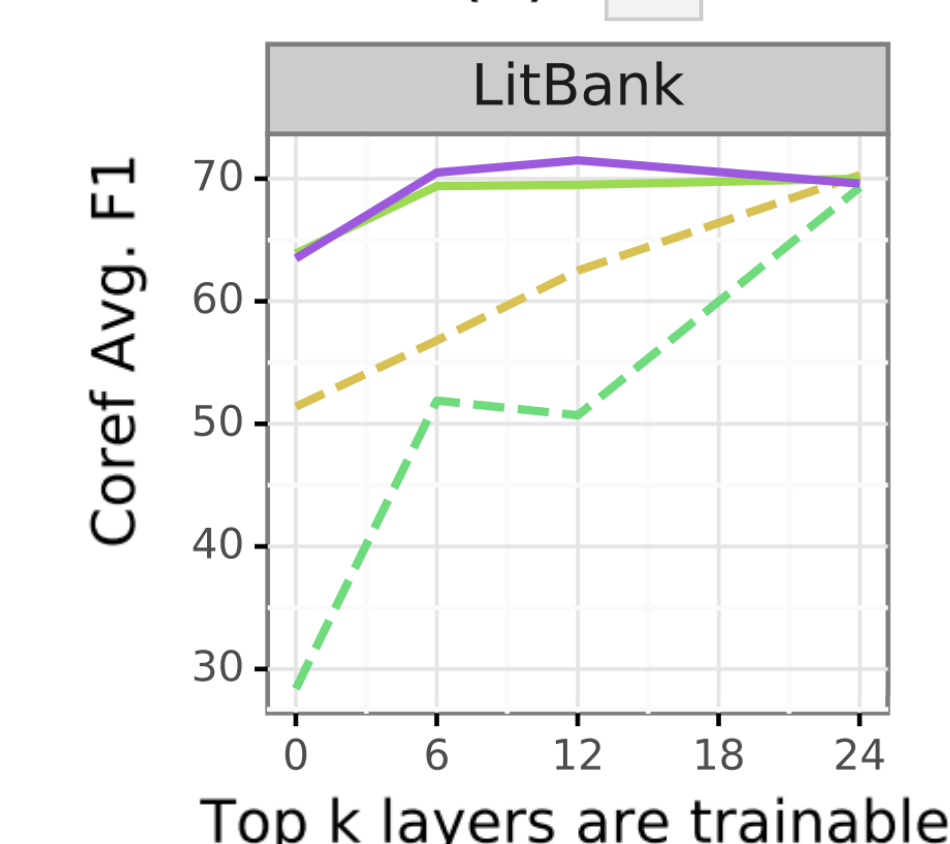
84.6 vs. 84.9
F1 with 5 vs. 500 dev docs

# dev documents	Transfer (on)
500	72.2 75.4 78.0 79.2 80.3 81.6 83.1 83.7 84.3 84.9
250	72.3 75.3 78.0 79.0 80.1 81.9 83.1 83.7 84.3 84.9
150	72.3 75.2 78.0 79.1 80.1 82.1 83.2 83.8 84.2 84.9
100	72.3 75.1 77.8 79.0 80.0 82.0 83.1 83.7 84.2 84.9
75	72.3 75.1 77.6 78.9 80.0 81.9 83.0 83.6 84.2 84.9
50	72.2 75.0 77.6 78.7 79.8 81.8 82.8 83.5 84.1 84.8
25	72.2 74.9 77.2 78.4 79.7 81.6 82.7 83.4 84.0 84.6
15	72.2 74.6 77.2 78.4 79.5 81.4 82.4 83.3 84.0 84.6
10	72.1 74.7 76.8 78.3 79.4 81.5 82.5 83.3 83.9 84.6
5	71.9 73.2 75.4 76.6 78.5 80.5 82.3 82.5 83.6 84.1

3. There is still catastrophic forgetting:

- Larger drops across annotation guideline changes
- Smaller drops across domain or language

4. Finetune only top encoder layers with continued training



Code and models available at:

<https://nlp.jhu.edu/coref-transfer/>