Moving on from OntoNotes: Coreference Resolution Model Transfer

Patrick Xia and Benjamin Van Durme
Background: Coreference Resolution

Determine which spans of text refer to the same entity

**Hong Kong Wetland Park, which is currently under construction**, is also one of the designated new projects of the Hong Kong government for advancing the tourism industry.

This is a park intimately connected with nature, being built by the Hong Kong government for its people who live in a city of reinforced concrete.
Background: Coreference Resolution

Determine which spans of text refer to the same entity

**antecedent**

Hong Kong Wetland Park, which is currently under construction, is also one of the designated new projects of the Hong Kong government for advancing the tourism industry.

**mention span**

This is a park intimately connected with nature, being built by the Hong Kong government for its people who live in a city of reinforced concrete.
Coreference Resolution

Determine which spans of text refer to the same entity

**Antecedent**

Hong Kong Wetland Park, which is currently under construction, is also one of the designated new projects of the Hong Kong government for advancing the tourism industry.

**Mention Span**

This is a park intimately connected with nature, being built by the Hong Kong government for its people who live in a city of reinforced concrete.
Annotation type:
- Singletons
- Entity types

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

Only coreferring mentions (OntoNotes)
And Jo shook the blue army sock till the needles rattled like castanets, and her ball bounded across the room.

All mentions, including singletons (ARRAU)
And Jo shook the blue army sock till the needles rattled like castanets, and her ball bounded across the room.

Only certain ACE entity types (LitBank)
Background: **Dataset Differences**

- **Domain**

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

**Literature**
Background: Dataset Differences

• Domain

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

Invisible Man is Ellison’s best known work, most likely because it was the only novel he ever published during his lifetime…
Dataset Differences

• Domain

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

Invisible Man is Ellison’s best known work, most likely because it was the only novel he ever published during his lifetime...

(1) In general, The term “employer” means with respect to any calendar year, any person who -
Background: Dataset Differences

• Domain

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

Invisible Man is Ellison’s best known work, most likely because it was the only novel he ever published during his lifetime...

• Language

  • Cross-lingual transfer of coreference resolution
Background: Poor Transferability
Background: Poor Transferability
Background: Poor Transferability
Background: Poor Transferability
Background: Poor Transferability

- LitBank
- Twitter
- en->zh
Research Questions
Research Questions

Goal: reduce cost of creating a coref model on entirely new dataset
Research Questions

Goal: reduce cost of creating a coref model on entirely new dataset

1. How effective is continued training for domain adaptation?
Research Questions

Goal: reduce cost of creating a coref model on entirely new dataset

1. How effective is continued training for domain adaptation?
2. How to allocate annotated documents?
Research Questions

Goal: reduce cost of creating a coref model on entirely new dataset

1. How effective is continued training for domain adaptation?
2. How to allocate annotated documents?
3. How much do source models forget?
Research Questions

Goal: reduce cost of creating a coref model on entirely new dataset

1. How effective is continued training for domain adaptation?
2. How to allocate annotated documents?
3. How much do source models forget?
4. Which encoder layers are important?
Methods: Source Models
Methods: Source Models

Memory-efficient coreference model
Methods: Source Models

Memory-efficient coreference model
Pretrained encoders only vs. fully-trained models
Methods: **Source Models**

Memory-efficient coreference model
Pretrained encoders only vs. fully-trained models

Pretrained encoder only

VS.

Encoder

Clusters

Linker

embeddings

text

SpanBERT, Longformer, etc
Methods: **Source Models**

Memory-efficient coreference model

Pretrained encoders only vs. fully-trained models

- Pretrained encoder only
- Trained encoder only
Methods: **Source Models**

Memory-efficient coreference model
Pretrained encoders only vs. fully-trained models

- **Pretrained encoder only**
  - Linker
  - Encoder
  - Text
  - Clusters
  - Embeddings
  - SpanBERT, Longformer, etc

- **Trained encoder only**
  - Linker
  - Encoder
  - Text
  - Clusters
  - Embeddings
  - SpanBERT-coref

- **Transfer model trained on source domain**
  - Linker
  - Encoder
  - Text
  - Clusters
  - Embeddings
Methods: Datasets
Methods: **Datasets**

**Source Datasets:** OntoNotes, PreCo
Methods: Datasets

Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Methods: Datasets

Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Methods: **Datasets**

**Source Datasets:** OntoNotes, PreCo

Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Methods: Datasets

Source Datasets: OntoNotes, PreCo

Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Methods: Datasets

Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Methods: Datasets

Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Multi-lingual: OntoNotes (en, zh, ar), SemEval (ca, es, it, nl)
Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Multi-lingual: OntoNotes (en, zh, ar), SemEval (ca, es, it, nl)
Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Multi-lingual: OntoNotes (en, zh, ar), SemEval (ca, es, it, nl)
Methods: **Datasets**

**Source Datasets:** OntoNotes, PreCo

Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)

Multi-lingual: OntoNotes (en, zh, ar), SemEval (ca, es, it, nl)
Methods: Datasets

Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Multi-lingual: OntoNotes (en, zh, ar), SemEval (ca, es, it, nl)
Methods: Datasets

Source Datasets: OntoNotes, PreCo
Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)
Multi-lingual: OntoNotes (en, zh, ar), SemEval (ca, es, it, nl)
Datasets

Source Datasets: OntoNotes, PreCo

Single domains: ARRAU (news), LitBank (books), SARA (legal), QBCoref (quiz questions)

Multi-lingual: OntoNotes (en, zh, ar), SemEval (ca, es, it, nl)
Methods: Training

Use standard train/dev splits
Sample a subset of training set to simulate lower-data setting
Research Question:

How effective is continued training for domain adaptation in coref?
Research Question:

How effective is continued training for domain adaptation in coref?

Encoder

Linker

clusters

embeddings

Encoder

text

Off-the-shelf trained encoder only

🎂

🎲
Research Question:

How effective is continued training for domain adaptation in coref?

Off-the-shelf trained encoder only

VS.
Research Question:

How effective is continued training for domain adaptation in coref?

Off-the-shelf trained encoder only

VS.

Transfer model trained on source domain
RQ1: Continued training for domain adaptation
RQ1: Continued training for domain adaptation

- **Transfer models** usually outperform **randomly initialized models**
RQ1: Continued training for domain adaptation

- **Transfer models** usually outperform **randomly initialized models**
- **PreCo** is as effective as **OntoNotes**
RQ1: Continued training for domain adaptation

- Transfer models usually outperform randomly initialized models
- PreCo is as effective as OntoNotes
- PreCo is better with gold mention boundaries
Research Question:

What’s better?

- Encoder
- Linker
- Clusters
- Embeddings
- Text

Untrained large encoder
Research Question:

What’s better?

Encoder text embeddings Linker clusters

VS.

Untrained large encoder
Research Question:

What’s better?

Untrained large encoder

VS.

Off-the-shelf trained small encoder
RQ1: Pretraining and model size
RQ1: Pretraining and model size

• Compare
  • **SpanBERT (L):** large unspecialized model (떡 + 빵)
  • **SpanBERT-On (b):** small specialized model (떡 + 빵)
RQ1: Pretraining and model size

- Compare
  - **SpanBERT (L):** large unspecialized model (🍪 + 🍬)
  - **SpanBERT-On (b):** small specialized model (🍪 + 🍬)
- Continued training of small (publicly available) encoders is effective with low # training docs
Additional Findings
RQ1: Continued training also improves cross-lingual transfer
RQ1: Continued training also improves cross-lingual transfer

- **Transfer model** ( роль + роль ) outperforms **XLM-R** ( роль + 🎲 )
RQ1: Continued training also improves cross-lingual transfer

• **Transfer model** (🎂 + 🎂) outperforms **XLM-R** (🎲 + 🎲)
• Improves SOTA performance on cross-lingual coreference

![Graph showing F1 score improvement with increasing training documents](image-url)
RQ2: How many documents should be in the dev set?
RQ2: How many documents should be in the dev set?

Answer: Increasing dev set from 5 to 500 documents only gains 0.3 F1
RQ3: How much do the models forget?
RQ3: How much do the models forget?

Largest drops:
- Annotation guideline changes
RQ3: How much do the models forget?

Largest drops:
• Annotation guideline changes

Small(er) drops:
• Cross-domain
• Cross-lingual
RQ4: Do we need to train the full encoder?
RQ4: Do we need to train the full encoder?

Answer:
RQ4: Do we need to train the full encoder?

Answer:

• For transfer (🎂 + 🎂) models, top 6-12 layers is probably enough
RQ4: Do we need to train the full encoder?

Answer:

- For transfer (🎂 + 🎂) models, top 6-12 layers is probably enough
- Not always true for other models
Conclusions
Conclusions

• Continued training is effective for coreference resolution:
  • Better overall performance
  • Good initial (zero-shot) performance
  • Cheaper training of new model
Conclusions

• Continued training is effective for coreference resolution:
  • Better overall performance
  • Good initial (zero-shot) performance
  • Cheaper training of new model

• PreCo is as good as OntoNotes
  • OntoNotes requires a license
Conclusions

• Continued training is effective for coreference resolution:
  • Better overall performance
  • Good initial (zero-shot) performance
  • Cheaper training of new model
• PreCo is as good as OntoNotes
  • OntoNotes requires a license
• For coreference, use annotated documents for training
Conclusions

• Continued training is effective for coreference resolution:
  • Better overall performance
  • Good initial (zero-shot) performance
  • Cheaper training of new model
• PreCo is as good as OntoNotes
  • OntoNotes requires a license
• For coreference, use annotated documents for training
• Fresh benchmarks on a wide set of datasets across domains and languages
Questions?
Come to poster session
Or email paxia@jhu.edu
Code/pretrained models at: https://nlp.jhu.edu/coref-transfer/