Adapting Coreference Resolution Models through Active Learning

ACL 2022

Michelle Yuan[1] Patrick Xia[2] Chandler May[2]
Benjamin Van Durme[2] Jordan Boyd-Graber[1]

[1]Department of Computer Science
University of Maryland

[2]Department of Computer Science
Johns Hopkins University

DEPARTMENT OF
COMPUTER SCIENCE

JOHNS HOPKINS
UNIVERSITY

Coreference Resolution (CR)

The task of discovering spans of text that refer to the same entity

Source

Traders said municipals were underpinned by influences, including the climb in Treasury issue prices. Also, municipal bonds lured buying because the stock market remains wobbly, traders contended. Mainly though, it was a favorable outlook for yesterday's new supply that propped up municipals , some traders said. Among the new issues was Massachusetts's $230 million of general obligation bonds . The bonds were won by a Goldman Sachs & Co. group with a true interest cost of 7.17%.

Cluster: { municipals , municipal bonds , municipals , general obligation bonds , The bonds }

Coreference Resolution (CR)

The task of discovering spans of text that refer to the same entity

Source

Traders said municipals were underpinned by influences, including the climb in Treasury issue prices. Also, municipal bonds lured buying because the stock market remains wobbly, traders contended. Mainly though, it was a favorable outlook for yesterday's new supply that propped up municipals , some traders said. Among the new issues was Massachusetts's $230 million of general obligation bonds . The bonds were won by a Goldman Sachs & Co. group with a true interest cost of 7.17%.

Cluster: { municipals , municipal bonds , municipals , general obligation bonds , The bonds }

Neural, end-to-end models (Lee et al., 2018; Joshi et al., 2020) are SOTA for OntoNotes 5.0

Problem: Adapting CR Models

Models trained on OntoNotes may not immediately generalize to new domains

Source

Traders said municipals were underpinned by influences, including the climb in Treasury issue prices. Also, municipal bonds lured buying because the stock market remains wobbly, traders contended. Mainly though, it was a favorable outlook for yesterday's new supply that propped up municipals , some traders said. Among the new issues was Massachusetts's $230 million of general obligation bonds . The bonds were won by a Goldman Sachs & Co. group with a true interest cost of 7.17%.

Cluster: { municipals , municipal bonds , municipals , general obligation bonds , The bonds }

Target

A molecule is a group of atoms held together by chemical bonds . Imagine you and your friends standing in a circle holding hands. Each person stands for one atom, your hands represent the bonds , and the entire circle represents a molecule.

Problem: Adapting CR Models

UNIVERSITY OF MARYLAND

Models trained on OntoNotes may not immediately generalize to new domains

Source

Traders said municipals were underpinned by influences, including the climb in Treasury issue prices. Also, municipal bonds lured buying because the stock market remains wobbly, traders contended. Mainly though, it was a favorable outlook for yesterday's new supply that propped up municipals , some traders said. Among the new issues was Massachusetts's $230 million of general obligation bonds . The bonds were won by a Goldman Sachs & Co. group with a true interest cost of 7.17%.

Cluster: { municipals , municipal bonds , municipals , general obligation bonds , The bonds }

Target

A molecule is a group of atoms held together by chemical bonds . Imagine you and your friends standing in a circle holding hands. Each person stands for one atom, your hands represent the bonds , and the entire circle represents a molecule.

Impedes immediate application for scenarios like distinguishing entities in legal documents

Problem: Adapting CR Models

- (Xia and Van Durme, 2021) show the benefits of *continued training* where a model trained on OntoNotes is further trained on the target dataset

Problem: Adapting CR Models

- (Xia and Van Durme, 2021) show the benefits of *continued training* where a model trained on OntoNotes is further trained on the target dataset
- However, they assume labeled data already exist in the target domain

Problem: Adapting CR Models

- (Xia and Van Durme, 2021) show the benefits of *continued training* where a model trained on OntoNotes is further trained on the target dataset
- However, they assume labeled data already exist in the target domain
- How can we adapt CR models without requiring large amounts of newly annotated data?

Method: Active Learning

Find particular spans of text for users to label

Target

(1)

A molecule is a group of atoms held together by chemical bonds . Imagine you and your friends standing in a circle holding hands. Each person stands for one atom, your hands represent **the bonds** , and the entire circle represents a molecule.

Active query: **the bonds** *Answer:* chemical bonds

Queries: your **the bonds** a molecule

(2)

Dr. Raizman thinks the animals help decrease patients' stress by increasing the release of oxytocin, which is the hormone that bonds mothers to babies. Studies have shown patients who spend time with dogs have higher levels of dopamine.

Active query: **bonds** *Answer:* query is not entity

Queries: oxytocin **bonds** patients dogs

Method: Active Learning

Find particular spans of text for users to label

Target

(1)

A molecule is a group of atoms held together by chemical bonds . Imagine you and your friends standing in a circle holding hands. Each person stands for one atom, your hands represent the bonds , and the entire circle represents a molecule.

Active query: the bonds *Answer:* chemical bonds

Queries: your the bonds a molecule

(2)

Dr. Raizman thinks the animals help decrease patients' stress by increasing the release of oxytocin, which is the hormone that bonds mothers to babies. Studies have shown patients who spend time with dogs have higher levels of dopamine.

Active query: bonds *Answer:* query is not entity

Queries: oxytocin bonds patients dogs

The goal is to adapt the model to the target domain by continue training it on these newly labeled spans

Method: Active Learning

We explore active learning for adapting CR models by:

1. Sampling spans according to different sources of model uncertainty

Method: Active Learning

We explore active learning for adapting CR models by:

1. Sampling spans according to different sources of model uncertainty

2. Understanding the trade-off between reading and labeling costs

Uncertainty Sampling

- A common active learning strategy is to sample data that the model is most uncertain about

Uncertainty Sampling

- A common active learning strategy is to sample data that the model is most uncertain about
- Maximum entropy sampling (Lewis and Gale, 1994) chooses data with highest predictive entropy

Uncertainty Sampling

- A common active learning strategy is to sample data that the model is most uncertain about
- Maximum entropy sampling (Lewis and Gale, 1994) chooses data with highest predictive entropy
- For CR, these strategies are not as straightforward because the model has different scores for mention detection and clustering

Decomposing Decisions in CR

1. **Mention detection:**
   - Given span x, let X be the random variable encoding whether x is an entity mention (1) or not (0)
   - The CR model first scores the likelihood of X being 1

Decomposing Decisions in CR

1. **Mention detection:**
   - Given span x, let X be the random variable encoding whether x is an entity mention (1) or not (0)
   - The CR model first scores the likelihood of X being 1
2. **Mention clustering:**
   - Given span x, let C be the random variable associated with the entity cluster of x
   - For spans that are likely entity mentions, the CR model determines $P(C = c \mid X = 1)$ for each observed entity cluster c

Maximum Entropy Sampling for CR

UNIVERSITY OF MARYLAND

ment-ent: Mention detection entropy

$$\begin{aligned} \mathrm{H}_{\mathrm{MENT}}(x) &= \mathrm{H}(X) \\ &= -\sum_{i=0}^{1} P(X=i)\log P(X=i). \end{aligned} \tag{1}$$

Samples spans that challenge mention detection (e.g. class-ambiguous words like “park”).

Maximum Entropy Sampling for CR

clust-ent: Mention clustering entropy

$$\mathrm{H}_{\mathrm{CLUST}}(x) = \mathrm{H}(C \mid X = 1) \tag{2}$$
$$= -\sum_{c \in \mathcal{C}} P(C = c \mid X = 1) \log$$
$$P(C = c \mid X = 1).$$

Entropy computation does not explicitly address uncertainty in mention detection.

Maximum Entropy Sampling for CR

cond-ent: Conditional entropy

$$\begin{aligned}
\mathrm{H}_{\mathrm{COND}}(x) &= \mathrm{H}(C \,|\, X) \\
&= \sum_{i=0}^{1} P(X=i)\mathrm{H}(C \,|\, X=i) \\
&= P(X=1)\mathrm{H}(C \,|\, X=1) \\
&= P(X=1)\mathrm{H}_{\mathrm{CLUST}}(x).
\end{aligned} \tag{3}$$

Samples words like pronouns because they are obviously entity mentions but difficult to cluster.

Maximum Entropy Sampling for CR

joint-ent: Joint entropy

$$\begin{aligned} \mathrm{H}_{\mathrm{JOINT}}(x) &= \mathrm{H}(X, C) = \mathrm{H}(X) + \mathrm{H}(C \mid X) \\ &= \mathrm{H}_{\mathrm{MENT}}(x) + \mathrm{H}_{\mathrm{COND}}(x). \end{aligned} \tag{4}$$

Samples spans that are difficult to detect as entity mentions *and* too confusing to cluster.

Experiments

Baselines:

1. **random:** Sample from all spans in the document
2. **random-ment:** Like other uncertainty strategies, sample only from the pool of spans that are likely entity mentions

Experiments

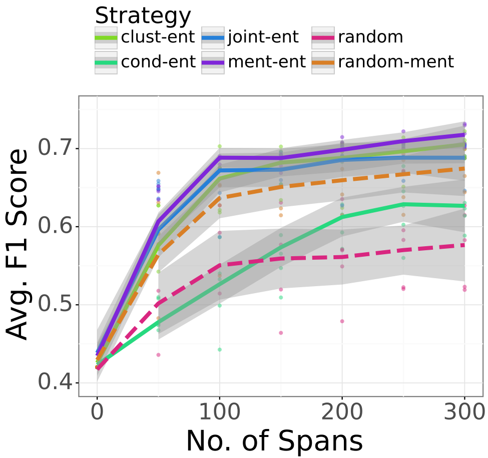
Baselines:

1. **random:** Sample from all spans in the document
2. **random-ment:** Like other uncertainty strategies, sample only from the pool of spans that are likely entity mentions

Datasets:

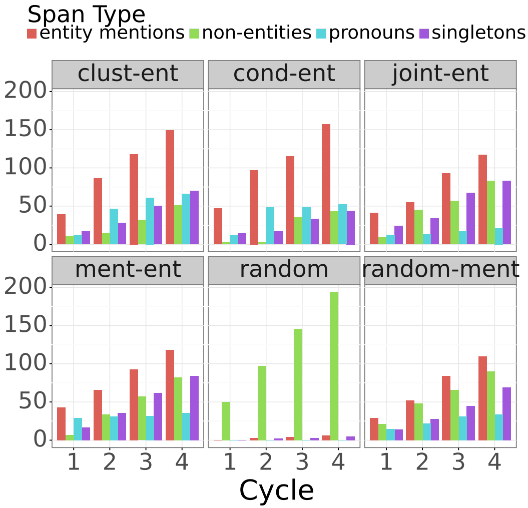
1. **OntoNotes 5.0:** Most common dataset for training and evaluating CR that contains news articles and telephone conversations (Pradhan et al., 2013). Only non-singletons are annotated.
2. **PreCo:** Large corpus of grad-school reading comprehension texts with annotated singletons (Chen et al., 2018).

Evaluating Sampling Strategies



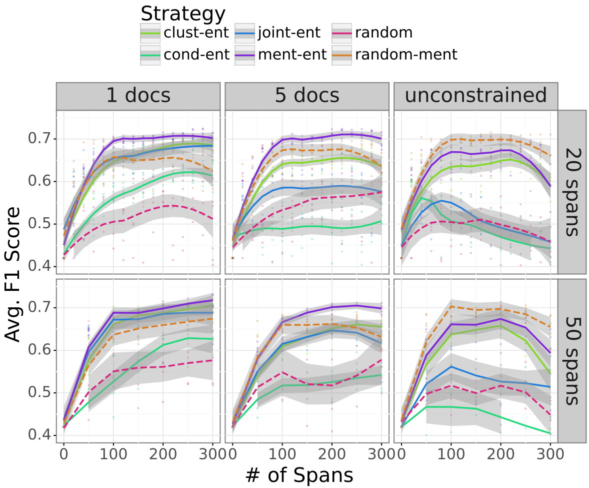
Test Avg. F1 on PreCo of models trained with each strategy where we simulate fifty spans from one document being labeled on each cycle

Distribution of Sampled Span Types



Cumulative counts of entities, non-entities, pronouns, and singletons sampled for each strategy over first four cycles of the PreCo simulation

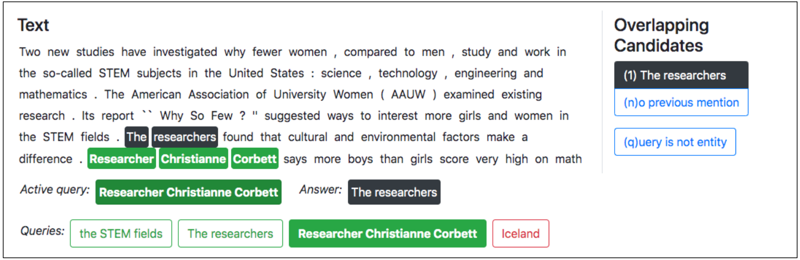
Reading and Labeling Trade-off



- Test Avg F1 on PreCo of models trained with each strategy
- Each column varies in the maximum number of docs. read per cycle
- Each column varies in the number of labeled spans per cycle

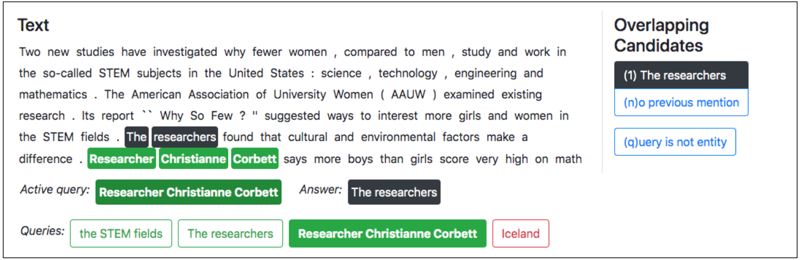
Reading and Labeling for Humans

Three users label spans sampled from PreCo



Reading and Labeling for Humans

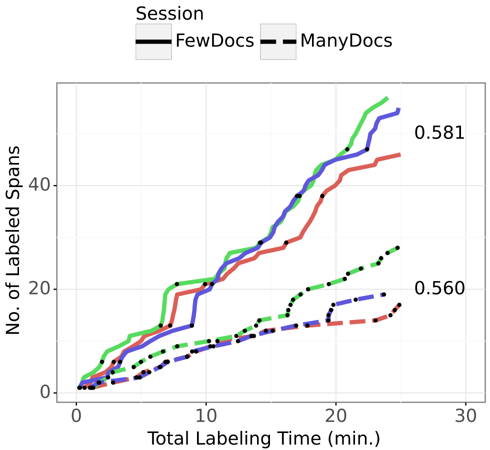
Three users label spans sampled from PreCo



Users complete two twenty-five minute sessions:

1. **FewDocs:** Read fewer docs. and label multiple spans per doc.
2. **ManyDocs:** Read more docs. and label one span per doc.

Reading and Labeling for Humans



- Each color indicates one of three users and the linetype designates the session
- Black dots mark the first span labeled in a different document
- The mean Avg F1 across users for each session is on the right

Summary

1. Neural CR models cannot immediately adapt without training on in-domain, labeled data
2. To reduce amount of annotation, we use active learning to choose particular text spans for users to label
3. We explore various aspects of active learning for CR, including sources of model uncertainty and the trade-off between reading and labeling
4. Sampling by mention detection entropy is more useful for domains like PreCo
5. In both simulations and the user study, CR improves from continued training on spans sampled from the same document rather than different contexts

Thanks

Any Questions?

myuan@cs.umd.edu

DEPARTMENT OF COMPUTER SCIENCE

References I

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. 6:557–570, 2018. doi: 10.1162/tacl_a_00039.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. 8:64–77, 2020. doi: 10.1162/tacl_a_00300.

Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. 2018. doi: 10.18653/v1/N18-2108.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. 1994.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. 2013. URL `https://www.aclweb.org/anthology/W13-3516`.

Patrick Xia and Benjamin Van Durme. Moving on from OntoNotes: Coreference resolution model transfer. 2021. doi: 10.18653/v1/2021.emnlp-main.425.