Annotating Mentions Alone Enables Efficient Domain Adaptation for Coreference Resolution

Nupoor Gandhi, Anjalie Field, Emma Strubell
Carnegie Mellon University
{nmgandhi, anjalief, estrubel}@cs.cmu.edu

Abstract

Although recent neural models for coreference resolution have led to substantial improvements on benchmark datasets, transferring these models to new target domains containing out-of-vocabulary spans and requiring differing annotation schemes remains challenging. Typical approaches involve continued training on annotated target-domain data, but obtaining annotations is costly and time-consuming. We show that annotating mentions alone is nearly twice as fast as annotating full coreference chains. Accordingly, we propose a method for efficiently adapting coreference models, which includes a high-precision mention detection objective and requires annotating only mentions in the target domain. Extensive evaluation across three English coreference datasets: CoNLL-2012 (news/conversation), i2b2/VA (medical notes), and previously unstudied child welfare notes, reveals that our approach facilitates annotation-efficient transfer and results in a 7-14% improvement in average F1 without increasing annotator time[1].

1 Introduction

Neural coreference models have made substantial strides in performance on standard benchmark datasets such as the CoNLL-2012 shared task, where average F1 has improved by 20% since 2016 (Durrett and Klein, 2013; Dobrovolskii, 2021; Kirstain et al., 2021). Modern coreference architectures typically consist of an encoder, mention detector, and antecedent linker. All of these components are optimized *end-to-end*, using only an antecedent linking objective, so expensive coreference chain annotations are necessary for training (Aralikatte and Søgaard, 2020; Li et al., 2020a).

These results have encouraged interest in deploying models in domains like medicine and child protective services, where a small number of practitioners need to quickly obtain information from large volumes of text (Uzuner et al., 2012; Saxena et al., 2020). However, successes over curated data sets have not fully translated to text containing technical vocabulary, frequent typos, or inconsistent syntax. Coreference models struggle to produce meaningful representations for new domain-specific spans and may require many examples to adapt (Uppunda et al., 2021; Lu and Ng, 2020; Zhu et al., 2021).

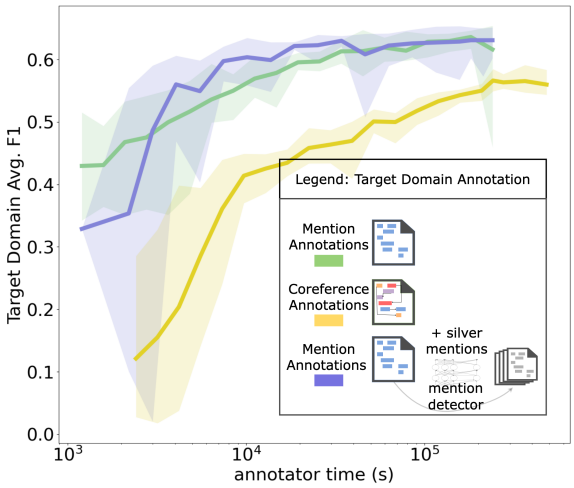

Figure 1: Model coreference performance (avg F1) as a function of continued training on limited target domain data requiring varying amounts of annotator time. The source domain is news/conversation (OntoNotes) and the target domain is medical notes (i2b2/VA). Using our method to adapt coreference models using only mentions in the target domain, we achieve strong coreference performance with less annotator time.

Further, coreference models trained on standard benchmarks are not robust to differences in annotation schemes for new domains (Bamman et al., 2020). For example, OntoNotes does not annotate *singleton* mentions, those that do not corefer with any other mention. A system trained on OntoNotes would implicitly learn to detect only entities that appear more than once, even though singleton retrieval is often desired in other domains (Zeldes, 2022). Also, practitioners may only be interested in retrieving a subset of domain-specific entities.

Continued training on target domain data is an

[1]Code is available at https://github.com/nupoorgandhi/data-eff-coref

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10543–10558

effective approach (Xia and Van Durme, 2021), but it requires costly and time-consuming coreference chain annotations in the new domain (Sachan et al., 2015). Annotating data in high-stakes domains like medicine and child protective services is particularly difficult, where privacy needs to be preserved, and domain experts have limited time.

Our work demonstrates that annotating only mentions is more efficient than annotating full coreference chains for adapting coreference models to new domains with a limited annotation budget. First, through timed experiments using the i2b2/VA medical notes corpus (Uzuner et al., 2012), we show that most documents can be annotated for mention detection twice as fast as for coreference resolution (§3). Then, we propose how to train a coreference model with mention annotations by introducing an auxiliary mention detection objective to boost mention precision (§4).

With this auxiliary objective, we observe that fewer antecedent candidates yields stronger linker performance. Continuity with previous feature-based approaches (Moosavi and Strube, 2016a; Recasens et al., 2013; Wu and Gardner, 2021) suggests this relationship between high-precision mention detection and strong coreference performance in low-resource settings extends beyond the architecture we focus on (Lee et al., 2018).

We evaluate our methods using English text data from three domains: OntoNotes (Pradhan et al., 2012), i2b2/VA medical notes (Uzuner et al., 2012), a new (unreleased) corpus of child welfare notes obtained from a county-level Department of Human Services (DHS). We experiment with standard benchmarks for reproducibility, but we focus primarily on real-world settings where there is interest in deploying NLP systems and limited capacity for in-domain annotations (Uzuner et al., 2012; Saxena et al., 2020). For a fixed amount of annotator time, our method consistently out-performs continued training with target domain coreference annotations when transferring both within or across annotation styles and vocabulary.

Our primary contributions include: Timing experiments showing the efficiency of mention annotations (§3), and methodology to easily integrate mention annotations (§4) into a common coreference architecture (Lee et al., 2018). Furthermore, to the best of our knowledge, this is the first work to examine coreference resolution in child protective settings. With empirical results demonstrating 7-14% improvements in F1 across 3 domains, we find that our approach for adaptation using mention annotations alone is an efficient approach for practical, real-world datasets.

2 Background and Task Definition

2.1 Neural Coreference Models

We focus our examination on the popular and successful neural approach to coreference introduced in Lee et al. (2017). This model includes three components: an encoder to produce span representations, a mention detector that outputs mention scores for candidate mentions, and a linker that outputs candidate antecedent scores for a given mention. For a document of length T, there are $\frac{T(T-1)}{2}$ possible mentions (sets of contiguous words).

For the set of candidate mentions, the system assigns a pairwise score between each mention and each candidate antecedent. The set of candidate antecedents is all previous candidate mentions in the document and a dummy antecedent (representing the case where there is no antecedent). For a pair of spans i, j, the pairwise score is composed of mention scores $s_m(i), s_m(j)$ denoting the likelihood that spans i and j are mentions and an antecedent score $s_a(i, j)$ representing the likelihood that span j is the antecedent of span i.

$$s(i, j) = s_m(i) + s_m(j) + s_a(i, j)$$

This architecture results in model complexity of $O(T^4)$, so it is necessary to prune the set of mentions. Lee et al. (2018) introduce coarse-to-fine (c2f) pruning: of T possible spans, c2f prunes the set down to M spans based on span mention scores $s_m(i)$. Then for each span i, we consider antecedent j based on the sum of their mention scores $s_m(i), s_m(j)$ and a coarse but efficient pairwise scoring function as defined in Lee et al. (2018).

2.2 Domain Adaptation Task Setup

In this work we investigate the following pragmatic domain adaptation setting: Given a text corpus annotated for coreference from source domain S, an un-annotated corpus from target domain T, and a limited annotation budget, our goal is to maximize coreference F1 performance in the target domain under the given annotation budget. We define this budget as the amount of annotation time.

The most straightforward approach to this task is to annotate documents with full coreference chains in the target domain until the annotation budget is

exhausted. Given an existing coreference model trained on the source domain, we can continue training on the annotated subset of the target domain. With a budget large enough to annotate at least 100 documents, this has been shown to work well for some domains (Xia and Van Durme, 2021).

2.3 Effect of In-Domain Training on Mention Detection and Antecedent Linking

Given that out-of-domain vocabulary is a common aspect of domain shift in coreference models (Uppunda et al., 2021; Lu and Ng, 2020), we hypothesize that mention detection transfer plays an important role in overall coreference transfer across domains. To test this hypothesis, we conduct a preliminary experiment, examining how freezing the antecedent linker affects overall performance in the continued training domain-adaptation setting described above. We train a c2f model with a SpanBERT encoder (Joshi et al., 2020) on OntoNotes, a standard coreference benchmark, and evaluate performance over the i2b2/VA corpus, a domain-specific coreference data set consisting of medical notes (see §5.2 for details). We additionally use the training set of i2b2/VA for continued in-domain training, and we isolate the impact of mention detection by training with and without freezing the antecedent linker.

Results are given in Table 1. Continued training of just the encoder and mention detector results in a large improvement of 17 points over the source domain baseline, whereas unfreezing the antecedent linker does not further significantly improve performance. This result implies that mention detection can be disproportionately responsible for performance improvements from continued training. If adapting only the encoder and mention detection portions of the model yields strong performance gains, this suggests that mention-only annotations, as opposed to full coreference annotations, may be sufficient for adapting coreference models to new domains.

Model	Recall	Precision	F1
SpanBERT + c2f	31.94	50.75	39.10
+ tune Enc, MD only	60.40	56.21	56.42
+ tune Enc, AL, MD	60.51	57.33	56.71

Table 1: When conducting continued training of a c2f model on target domain i2b2/VA, tuning the antecedent linker (AL) does not result in a significant improvement over just tuning the mention detector (MD) and encoder (Enc). All differences between tuned models and SpanBERT + c2f were statistically significant ($p < .05$)

3 Timed Annotation Experiments

In §2 we established that adapting just the mention detection component of a coreference model to a new domain can be as effective as adapting both mention detection and antecedent linking. In this section we demonstrate that annotating mentions is approximately twice as fast as annotating full coreference chains. While coreference has been established as a time-consuming task to annotate for domain experts (Aralikatte and Søgaard, 2020; Li et al., 2020a), no prior work measures the relative speed of mention versus full coreference annotation. Our results suggest, assuming a fixed annotation budget, coreference models capable of adapting to a new domain using only mention annotations can leverage a corpus of approximately twice as many annotated documents compared to models that require full coreference annotations.

We recruited 7 in-house annotators with a background in NLP to annotate two tasks for the i2b2/VA dataset. For the first mention-only annotation task, annotators were asked to highlight spans corresponding to mentions defined in the i2b2/VA annotation guidelines. For the second full coreference task, annotators were asked to both highlight spans and additionally draw links between mention pairs if coreferent. All annotators used INCEpTION (Klie et al., 2018) and underwent a 45 minute training session to learn and practice using the interface before beginning timed experiments.[2]

In order to measure the effect of document length, we sampled short (~200 words), medium (~500), and long (~800) documents. Each annotator annotated four documents for coreference resolution and four documents for mention identification (one short, one medium, and two long, as most i2b2/VA documents are long). Each document was annotated by one annotator for coreference, and one for mention detection. This annotation configuration maximizes the number of documents annotated (as opposed to the number of annotators per document), which is necessary due to the high variance in style and technical jargon in the medical corpus. In total 28 documents were annotated.

Table 3 reports the average time taken to annotate each document. On average it takes 1.85X more time to annotate coreference than mention detection, and the disparity is more pronounced (2X) for longer documents. In Table 6 (Appendix A)

[2] Annotators were compensated $15/hr and applied for and received permission to access the protected i2b2/VA data.

Average Task Annotation Time (s)			
Document Partition	Coreference	Mention	Speed-up
short (~200 words)	287.3	186.1	1.54
medium (~500 words)	582.5	408.8	1.42
long (~800 words)	1306.1	649.5	2.01
all	881.2	475.9	1.85

Table 2: Timed experiments of mention annotation as compared to full coreference annotations. Mention annotation 2X faster over longer documents.

we additionally report inter-annotator agreement. Agreement is slightly higher for mention detection, albeit differences in agreement for the two tasks are not significant due to the small size of the experiment, agreement is higher for mention detection.

Although results may vary for different interfaces, we show empirically that mention annotation is faster than coreference annotation.

4 Model

Given the evidence that a large benefit of continued training for domain adaptation is concentrated in the mention detector component of the coreference system (§2.3), and that mention annotations are much faster than coreference annotations (§3), in this section, we introduce methodology for training a neural coreference model with mention annotations. Our approach includes two core components focused on mention detection: modification to mention pruning (§4.2) and auxiliary mention detection training (§4.3). We also incorporate an auxiliary masking objective (§4.4) targeting the encoder.

4.1 Baseline

In our baseline model architecture (Lee et al., 2018), model components are trained using a coreference loss, where $Y(i)$ is the cluster containing span i predicted by the system, and GOLD(i) is the GOLD cluster containing span i:

$$\textbf{CL} = \log \prod_{i=1}^{N} \sum_{\hat{y} \in \mathcal{Y}(i) \cap \text{GOLD}(i)} P(\hat{y})$$

Of the set of N candidate spans, for each span i we want to maximize the likelihood that the correct antecedent set $\mathcal{Y}(i) \cap$ GOLD(i) is linked with the current span. The distribution over all possible antecedents for a given span i is defined using the scoring function s described in §2:

$$P(y) = \frac{e^{s(i,y)}}{\sum_{y' \in Y} e^{s(i,y')}}$$

4.2 Mention Pruning Modification

As described in §2, c2f pruning reduces the space of possible spans; however, there is still high recall in the candidate mentions. For example, our SpanBERT c2f model trained and evaluated over OntoNotes achieves 95% recall and 23% precision for mention detection. In state-of-the-art coreference systems, high recall with c2f pruning works well and makes it possible for the antecedent linker to correctly identify antecedents. Aggressive pruning can drop gold mentions.

Here, we hypothesize that in domain adaptation settings with a fixed number of in-domain data points for continued training, high-recall in mention detection is not effective. More specifically, it is evident that the benefits of high recall mention tagging are only accessible to highly discerning antecedent linkers. Wu and Gardner (2021) show that antecedent linking is harder to learn than mention identification, so given a fixed number of in-domain examples for continued training, the performance improvement from mention detection would surpass that of the antecedent linker. In this case, it would be more helpful to the flailing antecedent linker if the mention detector were precise.

Based on this hypothesis, we propose *high-precision c2f pruning* to enable adaptation using mention annotations alone. We impose a threshold q on the mention score $s_m(i)$ so that only the highest scoring mentions are preserved.

4.3 Auxiliary Mention Detection Task

We further introduce an additional cross-entropy loss to train only the parameters of the mention detector, where x_i denotes the span representation for the i'th span produced by the encoder:

$$\textbf{MD} = -\sum_{i=1}^{N} g(x_i) \log\left(s_m(x_i)\right) + \left(1 - g(x_i)\right) \log\left(1 - s_m(x_i)\right)$$

The loss is intended to maximize the likelihood of correctly identifying mentions where the indicator function $g(x_i) = 1$ iff x_i is a GOLD mention. The distribution over the set of mention candidates is defined using the mention score s_m. The mention detector is learned using a feed-forward neural network that takes the span representation produced by the encoder as input. The mention identification loss requires only mention labels to optimize.

4.4 Auxiliary Masking Task

We additionally use a masked language modeling objective (**MLM**) as described in Devlin et al. (2019). We randomly sample 15% of the WordPiece tokens to mask and predict the original token using cross-entropy loss. This auxiliary objective is intended to train the encoder to produce better span representations. Since continued training with an **MLM** objective is common for domain adaptation Gururangan et al. (2020), we also include it to verify that optimizing the **MD** loss is not implicitly capturing the value of the **MLM** loss.

5 Experiments

We evaluate our model on transferring between data domains and annotation styles. To facilitate reproducibility and for comparison with prior work, we conduct experiments on two existing public data sets. We additionally report results on a new (unreleased) data set, which reflects a direct practical application of our task setup and approach.

5.1 Datasets

OntoNotes (ON) (English) is a large widely-used dataset (Pradhan et al., 2012) with standard train-dev-test splits. Unlike the following datasets we use, the annotation style excludes singleton clusters. OntoNotes is partitioned into genres: newswire (nw), Sinorama magazine articles (mz), broadcast news (bn), broadcast conversations (bc), web data (wb), telephone calls (tc), the New Testament (pt).

i2b2/VA Shared-Task (i2b2) Our first target corpus is a medical notes dataset, released as a part of the i2b2/VA Shared-Task and Workshop in 2011 (Uzuner et al., 2012). Adapting coreference resolution systems to clinical text would allow for the use of electronic health records in clinical decision support or general clinical research for example (Wang et al., 2018). The dataset contains 251 train documents, 51 of which we have randomly selected for development and 173 test documents. The average length of these documents is 962.6 tokens with average coreference chain containing 4.48 spans. The annotation schema of the i2b2 data set differs from OntoNotes, in that annotators mark singletons and only mentions specific to the medical domain (PROBLEM, TEST, TREATMENT, and PERSON).

Child Welfare Case Notes (CN) Our second target domain is a new data set of contact notes from a county-level Department of Human Services (DHS).[3] These notes, written by caseworkers and service providers, log contact with families involved in child protective services. Because of the extremely sensitive nature of this data, this dataset has not been publicly released. However, we report results in this setting, as it reflects a direct, real-word application of coreference resolution and this work. Despite interest in using NLP to help practitioners manage information across thousands of notes (Saxena et al., 2020), notes also contain domain-specific terminology and acronyms, and no prior work has annotated coreference data in this setting. While experienced researchers or practitioners can annotate a small subset, collecting a large in-domain data set is not feasible, given the need to preserve families' privacy and for annotators to have domain expertise.

Out of an initial data set of 3.19 million contact notes, we annotated a sample of 200 notes using the same annotation scheme as i2b2, based on conversations with DHS employees about what information would be useful for them to obtain from notes. We adapt the set of entity types defined in the i2b2 annotation scheme to the child protective setting by modifying the definitions (Appendix A, Table 8). To estimate agreement, 20 notes were annotated by both annotators, achieving a Krippendorf's referential alpha of 70.5 and Krippendorf's mention detection alpha of 61.5 (Appendix A, Table 7).

On average, documents are 320 words with 13.5 coreference chains with average length of 4.7. We also replicated the timed annotation experiments described in §3 over a sample of 10 case notes, similarly finding that it takes 1.95X more time to annotate coreference than mention detection. We created train/dev/test splits of 100/10/90 documents, allocating a small dev set following Xia and Van Durme (2021).

We experiment with different source and target domain configurations to capture common challenges with adapting coreference systems (Table 3). We also select these configurations to account for the influence of singletons on performance metrics.

5.2 Experimental Setup

Baseline: c2f ($\mathbf{CL}_S, \mathbf{CL}_T$) For our baseline, we assume access to coreference annotations in target domain. We use pre-trained SpanBERT for our encoder. In each experiment, we train on the source

[3] Upon the request of the department, we do not report the name of the county in order to preserve anonymity.

Source S	Target T	OOV Rate	Anno. Style Match
i2b2	CN	32.3%	✓
ON	i2b2	20.8%	
ON Genre$_i$	ON Genre$_j$	(8.1%, 47.9%)	✓

Table 3: Summary of source-target configurations in our experiments. We experiment with transfer between domains with common or differing annotation style, where annotation style can dictate whether or not there are singletons annotated or domain-specific mentions to annotate for example.

domain with coreference annotations optimizing only the coreference loss $\mathbf{CL}_S$. Then, we continue training with $\mathbf{CL}_T$ on target domain examples.

We additionally experiment with an alternative baseline (high-prec. c2f $\mathbf{CL}_S, \mathbf{CL}_T, \mathbf{MD}_T$) in which coreference annotations are reused to optimize our **MD** over the target domain. This allows for full utilization the target domain annotations.

Proposed: high-prec. c2f ($\mathbf{CL}_S, \mathbf{MD}_T, \mathbf{MLM}_T$) We use the same model architecture and pre-trained encoder as the baseline, but also incorporate the joint training objective **CL + MD**. We optimize **CL** with coreference examples from the source domain ($\mathbf{CL}_S$), and **MD** with examples from the target domain ($\mathbf{MD}_T$). We report results only with $\mathbf{MD}_T$ paired with high-prec. c2f pruning (i.e. threshold $q = .5$ imposed on the mention score s_m) as described in §4. Without the threshold, $\mathbf{MD}_T$ has almost no effect on overall coreference performance, likely because the space of candidate antecedents for any given mention does not shrink.

Our model uses only mentions without target domain coreference links, while our baseline uses coreference annotations. Accordingly, we compare results for settings where there is (1) an equivalent number of annotated documents and (2) an equivalent amount of annotator time spent, estimated based on the timed annotation experiments in §3.

For each transfer setting, we assume the source domain has coreference examples allowing us to optimize $\mathbf{CL}_S$. In the target domain, however, we are interested in a few different settings: (1) 100% of annotation budget is spent on coreference, (2) 100% of annotation budget is spent on mentions, (3) the annotation budget is split between mention detection and coreference. In the first and third settings we can optimize any subset of $\{\mathbf{CL}_T, \mathbf{MD}_T, \mathbf{MLM}_T\}$ over the target domain, whereas $\mathbf{CL}_T$ cannot be optimized for the second.

We train the model with several different samples of the data, where samples are selected using a random seed. We select the number of random seeds based on the subsample size (Appendix B).

5.3 Augmented Silver Mentions

To further reduce annotation burden, we augment the set of annotated mentions over the target domain. We train a mention detector over a subset of gold annotated target-domain. Then, we use it to tag silver mentions over the remaining unlabeled documents, and use these silver mention labels in computing $\mathbf{MD}_T$.

5.4 Coreference Evaluation Configuration

In addition to the most common coreference metrics MUC, B^3, $CEAF_{\phi_4}$, we average across link-based metric LEA in our score. We also evaluate each model with and without singletons, since including singletons in the system output can artificially inflate coreference metrics (Kübler and Zhekova, 2011). When evaluating with singletons, we keep singletons (if they exist) in both the system and GOLD clusters. When evaluating without singletons, we drop singletons from both.

6 Results and Analysis

Table 4 reports results when transfering models trained on ON to i2b2 and models trained on i2b2 to CN with singletons included (for completeness Appendix A, Table 5 reports results without singletons). For both i2b2→CN and ON→i2b2, our model performs better with mention annotations than the continued training baseline with half the coreference annotations (e.g. equivalent annotator time, since the average length of i2b2 documents is 963 words; and timed experiments in CN suggested mention annotations are ~2X faster than coreference, §5.1). Combining $\mathbf{MLM}_T$ with $\mathbf{MD}_T$ results in our best performing model, but introducing $\mathbf{MD}_T$ with high-precision c2f pruning is enough to surpass the baseline. The results suggest in-domain mention annotation are more efficient for adaptation than coreference annotations.

6.1 Transfer Across Annotation Styles

ON and i2b2 have different annotation styles (§5.2), allowing us to examine how effectively mention-only annotations facilitate transfer not just across domains, but also across annotation styles. Transferring ON→i2b2 (Table 4), average F-1 improves by 6 points (0.57 to 0.63), when comparing the baseline model with 50% coreference annotations with our model (i.e. equivalent annotator time).

Model (Lee et al. (2018) + SpanBERT)	Target Anno.		ON→i2b2					i2b2→CN				
	$\mathbf{CL}_T$	$\mathbf{MD}_T$	LEA	MUC	B^3	$CEAF_\phi$	Avg.	LEA	MUC	B^3	$CEAF_\phi$	Avg.
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)	0%	0%	0.47	0.61	0.33	0.21	0.41	0.46	0.68	0.41	0.15	0.43
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)†	25%	0%	0.65	0.75	0.44	0.29	0.53	0.49	0.70	0.42	0.16	0.44
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{MD}_T$) + Silver	0%	50%	0.49*	0.63*	0.74*	0.61*	**0.63***	0.42*	0.70*	0.47*	0.22*	**0.45***
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)†	50%	0%	0.70	0.79	0.46	0.32	0.57	0.47	0.69	0.42	0.16	0.43
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{CL}_T, \mathbf{MD}_T$)†	50%	0%	0.69	0.79	0.45	0.29	0.56	0.52	0.72	0.47	0.21	0.48
+ c2f ($\mathbf{CL}_S, \mathbf{MD}_T$)	0%	100%	0.42*	0.56*	0.43	0.32	0.43	0.54*	0.77	0.47*	0.21*	0.49*
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{MD}_T$)	0%	100%	0.50*	0.63*	0.74*	0.65	0.63*	0.50	0.77*	0.52	0.35*	0.53
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{MD}_T, \mathbf{MLM}_T$)	0%	100%	0.50*	0.63*	0.77*	0.68*	**0.64***	0.57*	0.76*	0.58	0.38	**0.57***
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)	100%	0%	0.71	0.80	0.48	0.33	0.58	0.77	0.86	0.63	0.29	0.64

Table 4: We report F1 for different models with singletons included in system output, varying the type and amount of target domain annotations. Each shade of gray represents a fixed amount of annotator time (e.g. 50% Coreference and 100% Mention annotations takes an equivalent amount of time to produce). With a limited annotation budget, for both the ON→i2b2 and i2b2→CN experiments, mention annotations are a more efficient use of time, yielding performance gains over the baseline with equivalent annotator time (i.e. indicated with †). *denotes statistical significance with p-value < .05

In Figure 2 (top), we experiment with varying the amount of training data and annotator time in this setting. With more mentions, our model performance steadily improves, flattening out slightly after 1000 mentions. The baseline model continues to improve with more coreference examples. Where there is scarce training data (100-1000 mentions), mention annotations are more effective than coreference ones. This effect persists when we evaluate without singletons (Figure 5).

The baseline likely only identifies mentions that fit into the source domain style (e.g. PEOPLE). Because the baseline model assigns no positive weight in the coreference loss for identifying singletons, in i2b2, entities that often appear as singletons are missed opportunities to improve the baseline mention detector. With enough examples and more entities appearing in the target domain as non-singleton, however, the penalty of these missed examples is smaller, causing the baseline model performance to approach that of our model.

6.2 Silver Mentions Improve Performance

From Figure 2, approximately 250 gold mentions are necessary for sufficient mention detection performance for silver mentions to be useful to our model. For fewer mentions, the mention detector is likely producing silver mention annotations that are too noisy. The benefit of access to additional data starts to dwindle around 3000 mentions.

6.3 Fixed Annotation Style Transfer

We additionally compare effects when transferring between domains, but keeping the annotation style the same. When we transfer from i2b2 to CN, for equivalent annotator time, our model $\mathbf{MD}_T + \mathbf{MLM}_T$ improves over baseline $\mathbf{CL}_T$ by 14 points (.43 to .57) in Table 4. (When singletons are dropped, this effect persists — average F1 improves by 10 points, Appendix A, Table 5). When we vary the number of mentions (Figure 2), the marginal benefit of CN mention annotations deteriorates for $> 10^4$, but not as rapidly as when we transfer between annotation style in the ON→i2b2 case. While mentions in CN share the same roles as those in i2b2, some types of mentions, (e.g. PROBLEM), are more difficult to identify. Unlike settings where we transfer between annotation styles, when annotation style remains fixed, the performance improvement from our model increases with more target domain data. This suggests that adapting the mention detector is especially useful when transferring within an annotation style.

Given coreference annotations, we find that reusing the annotations to optimize $\mathbf{MD}_T$ with high-prec. c2f pruning boosts performance slightly when transferring within an annotation style. This is evident in the i2b2→CN case regardless of whether singletons are included in the output.

Figure 3 reports results for the genre-to-genre experiments within ON. For equivalent annotator time our model achieves large performance improvements across most genres. Since our model results in significant improvements in low-resource settings when there are no singletons in the system or gold clusters, it is clear that performance gains are not dependent solely on singletons in the system output. Figure 4 shows varying the number of mentions and annotator time in settings where our model performed worse ($bn \rightarrow nw$) and better ($bn \rightarrow pt$) than the baseline. Regardless of transfer setting or whether singletons are excluded from the system output, our model out-performs the baseline with few mentions.

6.4 Impact of Singletons

Under the with-singleton evaluation scheme, in the ON→i2b2 case, the baseline trained with strictly

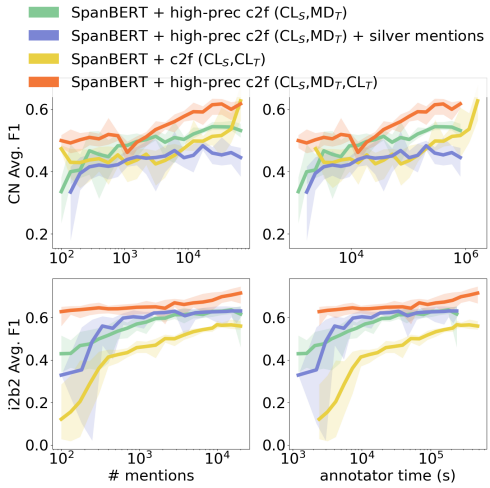

Figure 2: Each subplot shows coreference performance (singletons included) with varied amounts of annotated target domain data wrt the number of mentions (left) and the amount of annotator time (right). Note that for (**CL**$_S$, **MD**$_T$, **CL**$_T$), we vary only the amount of coreference annotations – the model accesses 100% of mention annotations. For ON→i2b2 (bottom), our model (**CL**$_S$, **MD**$_T$) has the largest improvement over the baseline (**CL**$_S$, **CL**$_T$) with limited annotations/time. For the i2b2→CN (top), however, the disparity increases with more annotations.

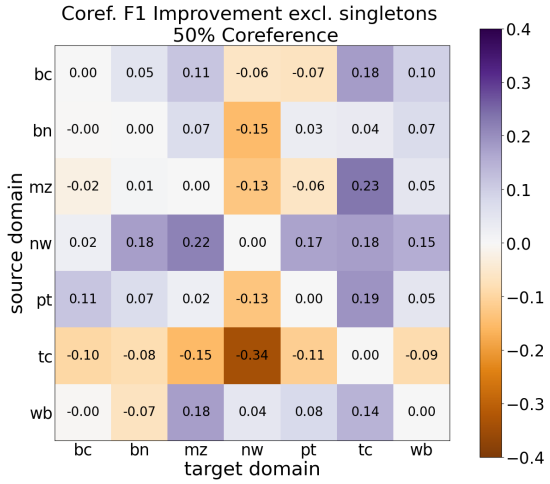

Figure 3: Heatmap represents performance improvements from our model where singletons are excluded. Our model SpanBERT + high-prec c2f (**CL**$_S$, **MD**$_T$) accesses 100% mention annotations from the target domain, and the baseline SpanBERT + c2f (**CL**$_S$, **CL**$_T$) accesses 50% of coreference examples. Annotating mentions for an equivalent amount of time is much more efficient for most ON genres.

more data performs worse than our model (Table 4, 0.58 vs. 0.64). Kübler and Zhekova (2011) describe how including singletons in system output causes artificial inflation of coreference metrics based on the observation that scores are higher with singletons included in the system output. Without high-precision c2f pruning with **MD**$_T$, the baseline drops singletons. So, the gap in Figure 2 between the baseline and our model at 10^4 mentions could be attributed to artificial inflation. In the without-

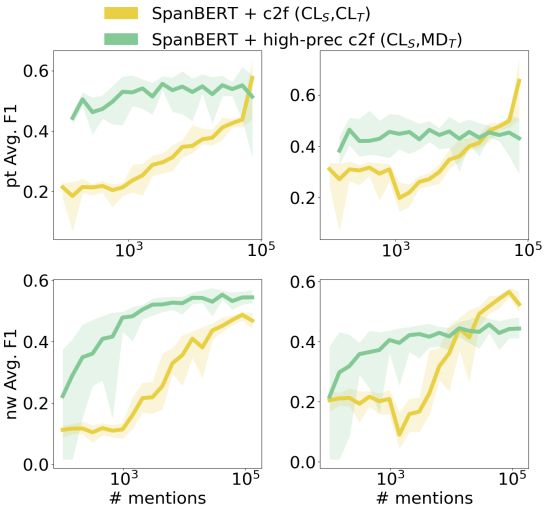

Figure 4: Each subplot shows coreference performance with varied amounts of annotated target data. We report performance with singletons included in system output (left) and singletons excluded from system output (right) for two different genre-to-genre experiments: $bn \rightarrow pt$ (top) and $bn \rightarrow nw$ (bottom). Regardless of whether singletons are included, annotating mentions is more efficient for all low-resource settings.

singleton evaluation scheme (Figure 4, bottom) the artificial inflation gap between our model and the baseline disappears with enough target examples, better reflecting our intuition that more data should yield better performance. But with fewer examples, our model still out-performs the baseline in the without-singleton evaluation scheme.

In practical applications, such as identifying support for families involved in child protective services, retrieving singletons is often desired. Further, excluding singletons in the system output incentivizes high-recall mention detection, since the model is not penalized for a large space of candidate mentions in which valid mentions make up a small fraction. A larger space of possible antecedents requires more coreference examples to adapt antecedent linkers to new domains.

7 Related Work

Previous work has used data-augmentation and rule-based approaches to adapt coreference models to new annotation schemes with some success (Toshniwal et al., 2021; Zeldes and Zhang, 2016; Paun et al., 2022). In many cases, adapting to new annotation schemes is not enough – performance degradation persists for out-of-domain data even under the same annotation scheme (Zhu et al., 2021), and encoders (SpanBERT) can struggle to represent domain specific concepts well, resulting in poor mention recall (Timmapathini et al., 2021).

Investigation of the popular Lee et al. (2017) architecture has found that coreference systems generally rely more on mentions than context (Lu and Ng, 2020), so they are especially susceptible to small perturbations. Relatedly, Wu and Gardner (2021) find that mention detection precision has a strong positive impact on overall coreference performance, which is consistent with findings on pre-neural systems (Moosavi and Strube, 2016b; Recasens et al., 2013) and motivates our work.

Despite challenges associated with limiting source domain annotation schema, with enough annotated data, coreference models can adapt to new domains. Xia and Van Durme (2021) show that continued training is effective with at least 100 target documents annotated for coreference. However, it is unclear how costly it would be to annotate so many documents: while Xia and Van Durme (2021) focus on the best way to use annotated coreference target examples, we focus on the most efficient way to spend an annotation budget.

A related line of work uses active learning to select target examples and promote efficient use of annotator time (Zhao and Ng, 2014; Li et al., 2020b; Yuan et al., 2022; Miller et al., 2012). However, since these annotations require link information, there is a persistent trade-off in active learning between reading and labeling (Yuan et al., 2022). Since our method does not require link annotations for adaptation, our annotation strategy circumvents the choice between redundant labeling or reading.

8 Limitations

Annotation speed for mention detection and coreference is dependent on many variables like annotation interface, domain expertise of annotators, annotation style, document length distribution. So, while our finding that coreference resolution is approximately 2X slower to annotate than mention detection held for two domains (i2b2, CN), there are many other variables that we do not experiment with.

We also experiment with transfer between domains with varying semantic similarity and annotation style similarity. But, our notion of annotation style is narrowly focused on types of mentions that are annotated (i.e. singletons, domain application-specific mentions). However, since our method is focused on mention detection, our findings may not hold for transfer to annotation styles with different notions of coreference linking (i.e. split-antecedent anaphoric reference (Yu et al., 2021)).

We also focus on one common coreference architecture Lee et al. (2018) with encoder SpanBERT. However, there have been more recent architectures surpassing the performance of Lee et al. (2018) over benchmark ON (Dobrovolskii, 2021; Kirstain et al., 2021). Our key finding that transferring the mention detector component can still be adopted.

9 Ethical Concerns

We develop a corpus of child welfare notes annotated for coreference. All research in this domain was conducted with IRB approval and in accordance with a data-sharing agreement with DHS. Throughout this study, the data was stored on a secure disk-encrypted server and access was restricted to trained members of the research team. Thus, all annotations of this data were conducted by two authors of this work.

While this work is in collaboration with the DHS, we do not view the developed coreference system as imminently deployable. Prior to considering deploying, at a minimum a fairness audit on how our methods would reduce or exacerbate any inequity would be required. Deployment should also involve external oversight and engagement with stakeholders, including affected families.

10 Conclusion

Through timing experiments, new model training procedures, and detailed evaluation, we demonstrate that mention annotations are a more efficient use of annotator time than coreference annotations for adapting coreference models to new domains. Our work has the potential to expand the practical usability of coreference resolution systems and highlights the value of model architectures with components that can be optimized in isolation.

Acknowledgements

Thanks to Yulia Tsvetkov, Alex Chouldechova, Amanda Coston, David Steier, and the anonymous Department of Human Services for valuable feedback on this work. This work is supported by the Block Center for Technology and Innovation, and A.F. is supported by a Google PhD Fellowship.

References

Rahul Aralikatte and Anders Søgaard. 2020. Model-based annotation of coreference. In *Proceedings of*

the 12th Language Resources and Evaluation Conference, pages 74–79, Marseille, France. European Language Resources Association.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Sandra Kübler and Desislava Zhekova. 2011. Singletons and coreference resolution evaluation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 261–267, Hissar, Bulgaria. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Maolin Li, Hiroya Takamura, and Sophia Ananiadou. 2020a. A neural model for aggregating coreference annotation in crowdsourcing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5760–5773, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Wei Zhao. 2020b. Active testing: An unbiased evaluation method for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 204–211, Online. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2020. Conundrums in entity coreference resolution: Making sense of the state of the art. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online. Association for Computational Linguistics.

Timothy Miller, Dmitriy Dligach, and Guergana Savova. 2012. Active learning for coreference resolution. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 73–81, Montréal, Canada. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016a. Search space pruning: A simple solution for better coreference resolvers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1005–1011, San Diego, California. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016b. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Silviu Paun, Juntao Yu, Nafise Sadat Moosavi, and Massimo Poesio. 2022. Scoring Coreference Chains with Split-Antecedent Anaphors.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 627–633, Atlanta, Georgia. Association for Computational Linguistics.

Mrinmaya Sachan, Eduard Hovy, and Eric P Xing. 2015. An active learning approach to coreference resolution. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Devansh Saxena, Karla Badillo-Urquiola, Pamela J Wisniewski, and Shion Guha. 2020. A human-centered review of algorithms used within the us child welfare system. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Hariprasad Timmapathini, Anmol Nayak, Sarathchandra Mandadi, Siva Sangada, Vaibhav Kesri, Karthikeyan Ponnalagu, and Vijendran Gopalan Venkoparao. 2021. Probing the spanbert architecture to interpret scientific domain adaptation challenges for coreference resolution. In *Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence*.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ankith Uppunda, Susan Cochran, Jacob Foster, Alina Arseniev-Koehler, Vickie Mays, and Kai-Wei Chang. 2021. Adapting coreference resolution for processing violent death narratives. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4553–4559, Online. Association for Computational Linguistics.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Zhaofeng Wu and Matt Gardner. 2021. Understanding mention detector-linker interaction in neural coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 150–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2021. Stay together: A system for single and split-antecedent anaphora resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4174–4184, Online. Association for Computational Linguistics.

Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. 2022. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland. Association for Computational Linguistics.

Amir Zeldes. 2022. Opinion Piece: Can we Fix the Scope for Coreference? Problems and Solutions for Benchmarks beyond OntoNotes. *Dialogue & Discourse*, 13(1):41–62.

Amir Zeldes and Shuo Zhang. 2016. When annotation schemes change rules help: A configurable approach to coreference resolution beyond OntoNotes. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 92–101, San Diego, California. Association for Computational Linguistics.

Shanheng Zhao and Hwee Tou Ng. 2014. Domain adaptation with active learning for coreference reso-

lution. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 21–29, Gothenburg, Sweden. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. Anatomy of OntoGUM—Adapting GUM to the OntoNotes scheme to evaluate robustness of SOTA coreference algorithms. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 141–149, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Additional Results

For completeness, we additionally include results with singletons omitted from system output. Table 5 reports results for both transfer settings i2b2→CN and ON→i2b2. In Figure 5, we inspect how performance changes with more annotated data. We also report for completeness the difference in model performance using mention annotations and full coreference annotations in Figure 6 for transfer between OntoNotes genres with an equivalent amount of annotated data (unequal amount of annotator time).

For our timed annotation experiment described in §3, we report more detailed annotator agreement metrics for the two annotation tasks in Table 6. We expect that agreement scores for both tasks are low, since i2b2/VA dataset is highly technical, and annotators have no domain expertise. The increased task complexity of coreference resolution may further worsen agreement for the task relative to mention detection. We do not use this annotated data beyond timing annotation tasks.

B Reproducibility Details

Implementation Details For all models, we began first with a pretrained SpanBERT (base) encoder (Joshi et al., 2020) and randomly initialized parameters for the remaining mention detector and antecedent linking. We use 512 for maximum segment length with batch size of one document similar to Lee et al. (2018). We first train the model with a coreference objective over the source domain $\mathbf{CL}_S$, and then we train over the target domain with some subset of our objectives $\mathbf{CL}_T, \mathbf{MD}_T, \mathbf{MLM}_T$

We do not weight auxiliary objectives, taking the raw sum over losses as the overall loss. When we train one objective over both the source and target domain (i.e. $\mathbf{CL}_S, \mathbf{CL}_T$), we interleave examples from each domain. For the **CL** objective, initial experiments indicated that, for fewer than 1k target domain mentions, our baseline model performed better if we interleaved target and source examples. So, we interleave target and source examples with fewer than 1k mentions from the target domain.

For experiments where the number of mentions from the target domain varied, we randomly sampled documents until the number of mentions met our cap (truncating the last document if necessary).

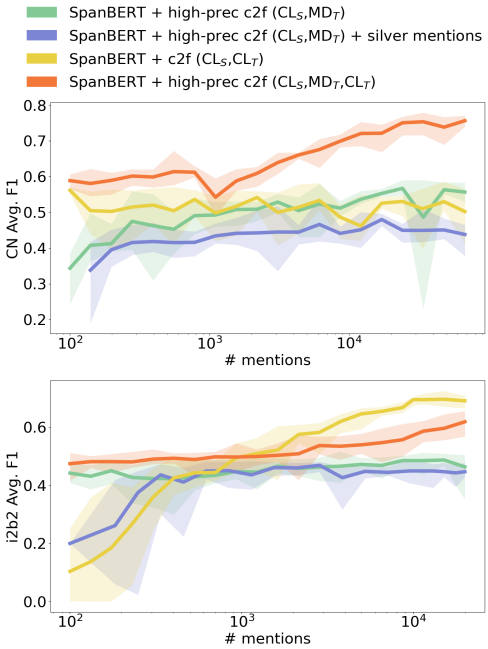

Figure 5: Each subplot shows coreference performance (singletons excluded) when trained with different amounts of annotated target domain data. We vary the amount of annotated data with respect to the number of mentions. When transferring ON→i2b2 (bottom row), our model ($\mathbf{CL}_S, \mathbf{MD}_T$) has the largest improvement over the baseline ($\mathbf{CL}_S, \mathbf{CL}_T$) with very little training data or annotator time. For the i2b2→CN (top row), however, the performance improvement increases with more annotated data.

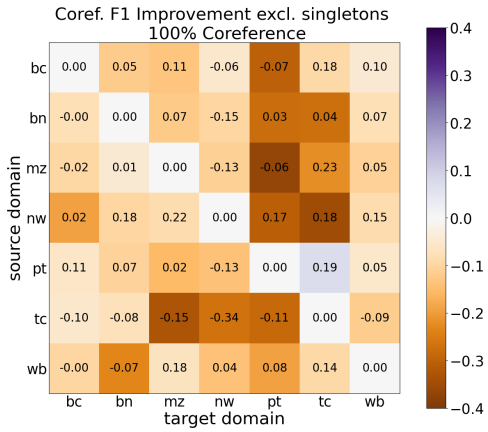

Figure 6: Heatmap represents performance improvements from our model SpanBERT + high-prec c2f ($\mathbf{CL}_S, \mathbf{MD}_T$) over the baseline SpanBERT + c2f ($\mathbf{CL}_S, \mathbf{CL}_T$) where singletons are dropped from the system output. The baseline has access to 100% of target domain coreference examples, and our model has access to 100% mention annotations.

Model (Lee et al. (2018) + SpanBERT)	Target Anno.		ON→i2b2					i2b2→CN				
	$\mathbf{CL}_T$	$\mathbf{MD}_T$	LEA	MUC	B^3	$CEAF_\phi$	Avg.	LEA	MUC	B^3	$CEAF_\phi$	Avg.
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)	0%	0%	0.47	0.61	0.49	0.24	0.45	0.46	0.68	0.49	0.38	0.50
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)†	25%	0%	0.65	0.75*	0.68*	0.50	**0.65***	0.49	0.70	0.51	0.41	**0.53**
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{MD}_T$) + Silver	0%	50%	0.49	0.63	0.50	0.15	0.44	0.42	0.70	0.44	0.23*	0.45
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)†	50%	0%	0.70	0.79	0.72	0.57	**0.70**	0.47	0.69	0.50	0.40	0.51
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{CL}_T, \mathbf{MD}_T$)†	50%	0%	0.69	0.79	0.72	0.57	0.69	0.52	0.72	0.55	0.45	0.56
+ c2f ($\mathbf{CL}_S, \mathbf{MD}_T$)	0%	100%	0.42*	0.56	0.44	0.18	0.40*	0.54	0.77*	0.56	0.45	0.58
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{MD}_T$)	0%	100%	0.50	0.63	0.53*	0.32*	0.49	0.50	0.77	0.52	0.42	0.55
+ high-prec. c2f ($\mathbf{CL}_S, \mathbf{MD}_T, \mathbf{MLM}_T$)	0%	100%	0.50	0.63	0.51	0.22	0.47	0.57	0.76*	0.60*	0.49*	**0.61***
+ c2f ($\mathbf{CL}_S, \mathbf{CL}_T$)	100%	0%	0.71	0.80	0.74	0.61	0.71	0.77	0.86	0.78	0.71	0.78

Table 5: We report F1 for different models with singletons excluded from system output, varying the type and amount of target domain annotations. Each shade of gray represents a fixed amount of annotator time (e.g. 50% Coreference and 100% Mention annotations takes an equivalent amount of time to produce). When transferring annotation styles (ON→i2b2), coreference annotations are a more efficient use of time, while when transferring within an annotation style (i2b2→CN), mention annotations are more efficient, consistent with results where singletons are included in the system output. Baselines are indicated with † and *denotes statistical significance with p-value < .05

Timed Annotation Experiment Mention Detection Agreement		
Agreement Metric	Non-expert Annotators	Domain-expert Annotators
Krippendorf's alpha	0.405	-
Average Precision	0.702	-
Average Recall	0.437	-
Average F1	0.527	-
IAA	0.691	0.97

Timed Annotation Experiment Coreference Agreement		
Agreement Metric	Non-expert Annotators	Domain-expert Annotators
Krippendorf's alpha	0.371	-
Average Precision	0.275	-
Average Recall	0.511	-
Average F1	0.342	-
IAA	0.368	0.73

Table 6: Annotation agreement metrics for timed experiments of mention detection and coreference resolution. Inter-Annotator Agreement (IAA) refers to a metric defined in (Uzuner et al., 2012). For coreference, precision, recall, and F1 are averaged over standard metrics defined in §B.

For a given number of mentions m, we generated models for $\min(\max(6, 15000/m), 15)$ random seeds. These bounds were selected based on preliminary experiments assessing deviation.

We use a learning rate of 2×10^{-5} for the encoder and 1×10^{-4} for all other parameters. We train on the source domain for 20 epochs and on the target domain for 20 epochs or until coreference performance over the dev set degrades for two consecutive iterations. Training time for all models ranges between 80-120 minutes, depending on size of dataset. We used V100, RTX8000, and RTX600 GPUS for training. To reproduce the results in this paper, we approximate at least 1,500 hours of GPU time. All our models contain ~134M parameters, with 110M from SpanBERT (base).

Evaluation We evaluate with coreference metrics: MUC, B^3, $CEAF_{\phi_4}$, LEA for the ON→i2b2 and i2b2→CN transfer settings and only MUC, B^3, $CEAF_{\phi_4}$ for ON genre transfer experiments, since these three are standard for OntoNotes. We report results with singletons included and excluded from system output. Our evaluation script can be found at `src/coref/metrics.py`.

CN Dataset Additional Details Table 8 lists the specific definitions for labels used by annotators in the CN dataset, as compared to the descriptions in the i2b2/VA dataset after which they were modeled. Table 7 reports measures for inter-annotator agreement for the CN dataset, compared to agreement reported for coreference annotations in OntoNotes.

CN Annotation Agreement		
Agreement Metric	Non-expert Annotators	OntoNotes
MUC	72.0	68.4
$CEAF_{\phi_4}$	40.5	64.4
$CEAF_m$	63.4	48.0
B^3	57.8	75.0
Krippendorf's MD alpha	60.5	61.9
Krippendorf's ref. alpha	70.5	-

Table 7: Annotation agreement metrics for the CN dataset computed over a random sample of 20 documents. We achieve agreement on par with OntoNotes (Pradhan et al., 2012).

	i2b2/VA definition	CN definition
TREATMENT	phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem (e.g. Revascularization, nitroglycerin drip)	phrases that describe efforts made to improve outcome for child (e.g. mobile therapy, apologized)
TEST	phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample in order to discover, rule out, or find more information about a medical problem (e.g. exploratory laproratomy, the ekg, his blood pressure)	phrases that describe steps taken to discover, rule out, or find more information about a problem (e.g. inquired why, school attendance)
PROBLEM	phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease (e.g. new ss chest pressure, rigidity, subdued)	phrases that contain observations made by CW or client about any client's body or mind that are thought to be abnormal or harmful (e.g. verbal altercation, recent breakdown, lack of connection, hungry)

Table 8: In addition to the PERSON entity type which is the same in both domains, we develop a set of types for the child welfare domain that can be aligned with those from the medical domain i2b2/VA as defined in (Uzuner et al., 2012). While the development of these types were intended to facilitate transfer from the medical domain, they are not necessarily comprehensive or sufficiently granular for the downstream tasks that coreference systems may be used for in child protective settings.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
9

☑ A2. Did you discuss any potential risks of your work?
10

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

5.1

☑ B1. Did you cite the creators of artifacts you used?
5.1

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
5.1

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
5.1

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. While the i2b2/VA medical notes dataset is anonymized, the Child Welfare Case Notes dataset that we developed is not anonymized, since it is not public released.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
5.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
5.1

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
i2b2/VA data is protected, so we are unable to provide example screenshots

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3