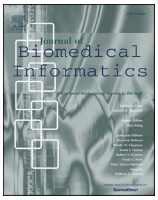
Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Fine-tuning coreference resolution for different styles of clinical narratives

Yuxiang Liao, Hantao Liu, Irena Spasić [*]

School of Computer Science and Informatics, Cardiff University, United Kingdom

ARTICLE INFO

Keywords:
Natural language processing
Coreference resolution
Transfer learning
Active learning
Ensemble algorithm

ABSTRACT

Objective: Coreference resolution (CR) is a natural language processing (NLP) task that is concerned with finding all expressions within a single document that refer to the same entity. This makes it crucial in supporting downstream NLP tasks such as summarization, question answering and information extraction. Despite great progress in CR, our experiments have highlighted a substandard performance of the existing open-source CR tools in the clinical domain. We set out to explore some practical solutions to fine-tune their performance on clinical data.

Methods: We first explored the possibility of automatically producing silver standards following the success of such an approach in other clinical NLP tasks. We designed an ensemble approach that leverages multiple models to automatically annotate co-referring mentions. Subsequently, we looked into other ways of incorporating human feedback to improve the performance of an existing neural network approach. We proposed a semi-automatic annotation process to facilitate the manual annotation process. We also compared the effectiveness of active learning relative to random sampling in an effort to further reduce the cost of manual annotation.

Results: Our experiments demonstrated that the silver standard approach was ineffective in fine-tuning the CR models. Our results indicated that active learning should also be applied with caution. The semi-automatic annotation approach combined with continued training was found to be well suited for the rapid transfer of CR models under low-resource conditions. The ensemble approach demonstrated a potential to further improve accuracy by leveraging multiple fine-tuned models.

Conclusion: Overall, we have effectively transferred a general CR model to a clinical domain. Our findings based on extensive experimentation have been summarized into practical suggestions for rapid transferring of CR models across different styles of clinical narratives.

1. Introduction

The main aim of this study is to identify a practical solution for rapidly developing coreference resolution (CR) models for a specific style of clinical narratives. We will demonstrate a proof of concept using radiology reports as a case study. This choice has been motivated by the vital role these reports play in patient care as referring clinicians use them to determine an appropriate course of action. Narrative radiology reports vary excessively in their language, length and style, which may affect their clarity and hence the referring clinicians' decision-making [1]. These issues gave rise to an idea of structured reporting, which has the potential for improving the clarity of radiology reports. Automated structuring of narrative reports can facilitate extraction, storage and retrieval of information they describe [2]. CR, which aims to explicitly link up all expressions that mention the same entity [3] (see Fig. 1 for examples), is necessary to identify sentences that belong to topically cohesive observations of a structured report.

A variety of CR models have been integrated into popular open-source natural language processing (NLP) tools. For example, Stanford CoreNLP [4] alone incorporates three methodologically different CR approaches: statistical [5], deterministic [6] and neural [7]. The most recent trend in NLP towards neural network approaches can also be observed in CR. A CR model based on SpanBERT [8] has been incorporated into spaCy [9] and AllenNLP [10]. These NLP tools are generic and as such their performance does not necessarily translate into specialized domains such as the clinical one [11]. Even within a single domain, sublanguages can vary [12], which means that when a model is trained within a domain, its performance may still vary across different types of documents.

Nonetheless, fine-tuning an existing CR model to previously unseen data represents a cost-effective strategy. Fine-tuning can be implemented by simply injecting a small subset of newly annotated data into

[*] Corresponding author.
E-mail addresses: liaoy11@cardiff.ac.uk (Y. Liao), liuh35@cardiff.ac.uk (H. Liu), spasici@cardiff.ac.uk (I. Spasić).

https://doi.org/10.1016/j.jbi.2023.104578
Received 8 August 2023; Received in revised form 22 November 2023; Accepted 12 December 2023
Available online 18 December 2023

the training set. However, manual annotation of the gold standard data has been identified as a major bottleneck in machine learning approaches to clinical NLP [13]. In an attempt to bypass this bottleneck, we explored a possibility of automatically producing silver standards following the success of such an approach in other clinical NLP tasks [14–17]. Unfortunately, our experiments demonstrated that this approach negatively impacted the performance of a fine-tuned model.

We subsequently manually annotated two gold standards. One gold standard was based on random sampling and the other on active learning, which had been shown to be effective in training a CR model [18]. To facilitate the manual annotation process, we proposed a semi-automatic annotation process. We used a pre-trained CR model to annotate the new data before passing them to human annotators for curation. Our experiments have demonstrated improved consistency and efficiency of human annotation as well as the improved performance of a fine-tuned model. In an attempt to further improve accuracy, we leveraged multiple fine-tuned models in an ensemble approach. In the remainder of the paper, we provide specific details of the suggested approaches and compare their results in order to suggest the best practices for rapid fine-tuning of existing CR models for different styles of clinical narratives.

2. Related work

In 2017, Lee et al. [19] proposed an end-to-end neural CR model, which used a long short-term memory (LSTM) encoder. The success of large language models (LLMs) in NLP has also seen widespread development of neural-based CR models. The benchmark of neural-based CR systems on the OntoNotes dataset [20] has been improved from the early 65.7 % CoNLL F1-score [7] to the current 81 % [21]. These models have been routinely employed by open-source NLP tools. For example, the most recent updates were found in the spaCy and AllenNLP, both of which were based on SpanBERT [8] and c2f-coref model [22], which achieved 79.6 % F1-score on OntoNotes.

To enhance a CR model to handle unseen data, Toshniwal et al. [23] proposed joint training to improve the generalization ability of a CR model. Their joint model achieved 70 % F1-score on average on three known datasets (used for joint training) and five unknown datasets (used only for testing) while achieving 79.6 % on OntoNotes. However, on the five general domain datasets that did not participate in the training, the F1 score improved modestly from 59.0 % to 64.1 %. On the other hand, Zhang et al. [17] designed an interesting joint training pipeline for syntactic analysis (SA) tasks, achieving high-quality predictions on unannotated clinical data. They utilized a general SA model to annotate clinical notes. The silver-standard data were then merged with the original training data. Subsequently, they re-trained the model and improved the labelled attachment score [24] from 76.0 % to 82.8 %. Notably, the test set on which they evaluated the model was annotated by the model itself, which might not correctly represent the model's performance due to its bias.

To evaluate LLMs on domain-specific CR, Lu and Poesio [3] conducted comparative experiments on biomedical coreference corpora (CRAFT-CR) using a general-domain SpanBERT and BERT variants pre-trained on biomedical data. Their results showed that the general SpanBERT model (F1 = 47.8 %) is superior to the biomedical variants of BERT (with F1-scores ranging from 27.4 % to 45.3 %). As a long-sequence transformer model, Longformer [25] is a better alternative to SpanBERT for CR and has been shown to generalize better on this task [23].

Yuan et al. [18] explored the practical adaptation of active learning in general domain CR. They pointed out that different uncertainty-based query strategies may be suitable for different datasets, but in any case, perform better than a random query strategy. Moreover, due to the time consumption in understanding the context, they encouraged annotators to do comprehensive document annotation instead of trying to rapidly go through many documents with incomplete annotation.

3. Methods

The methods described here are designed to answer the following research questions. First, are pre-trained models appropriate for domain-specific CR? Second, can existing CR models be fine-tuned for a specific domain? If so, what would be the most effective data annotation strategy? Specifically, is manual annotation more effective than semi-automatic annotation where training data are annotated automatically by a pre-trained model and then curated manually? When manually annotating data, is active learning more effective than random sampling of training data? When automatically annotating training data, does an ensemble approach improve the accuracy of annotation? When adding new datasets, which training strategy is most effective? To answer these questions, we first need to select a pre-trained CR model and the source of data to annotate in order to fine-tune the given model.

3.1. Pre-trained model

Many of the recent well-performing CR models were built upon various LLMs and achieved very similar performances, including c2f-coref [22] + SpanBERT [8] (F1 = 79.6 %), Longdoc [23] + Longformer [25] (F1 = 79.6 %) and wl-coref [21] + RoBERTa [26] (F1 = 80.0 %). Furthermore, Lu and Poesio [3] discovered that SpanBERT outperformed other biomedical variants of BERT. The above evidence suggests that these recent LLM-based models may also be eligible for the

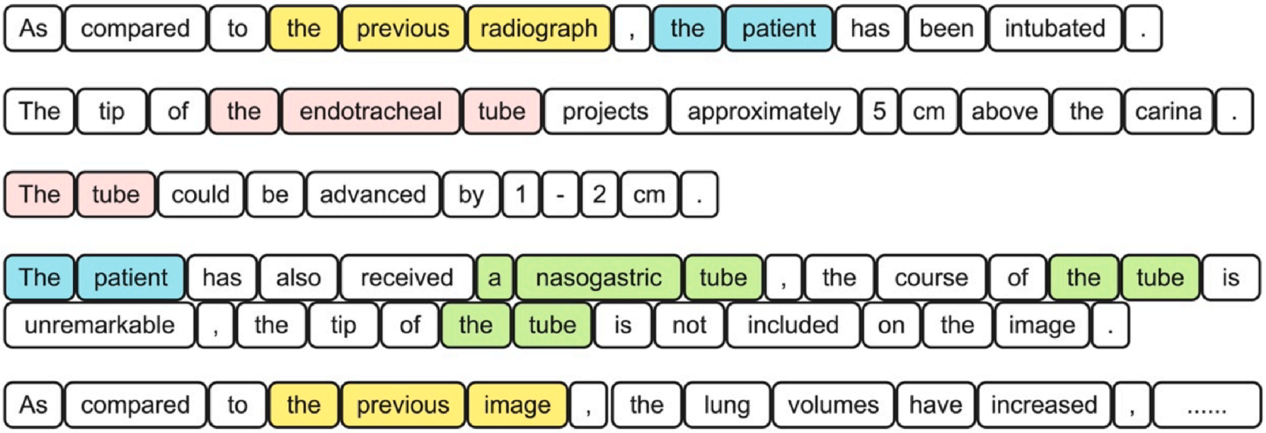

As compared to the previous radiograph , the patient has been intubated .
The tip of the endotracheal tube projects approximately 5 cm above the carina .
The tube could be advanced by 1 - 2 cm .
The patient has also received a nasogastric tube , the course of the tube is unremarkable , the tip of the tube is not included on the image .
As compared to the previous image , the lung volumes have increased , ...

Fig. 1. Examples of coreferences in a radiology report. A coreference mention is represented as a continuous span. Coreference mentions with the same colour refer to the same entity, and thus belong to the same coreference cluster.

clinical domain. Given their similar performance, we selected Longdoc as it naturally supports joint training with better generalization ability and better adaptation for longer text.

Longdoc has three components: a document encoder, a mention proposer and a mention cluster predictor. It uses Longformer-large [25], a long-sequence transformer-based model, to encode documents. The mention proposer and the mention cluster predictor are two stacked feed-forward neural networks (FFNNs). Given a candidate mention span x, and a mention entity e that is currently being tracked by the model, the proposer and predictor work as follow:

$$s_m(x) = FFNN_m(g_x) \tag{1}$$

$$s_c(x,e) = FFNN_c(g_x, g_e, g_x \odot g_e, \phi(x,e)) + s_m(x) \tag{2}$$

where g_x is the representation of a span x produced by the document encoder, $s_m(x)$ is the mention score representing the likelihood that the span x constitutes a mention, and $s_c(x,y)$ is the pairwise scoring function that decides whether the span x refers to any of the active entities e. The symbol $\odot$ represents an element-wise product, and $\phi(x,e)$ concatenates the feature embeddings of the span x and the entity e. After that, a memory-based decision is activated if the span x does not refer to any entity. We selected the unbounded memory setting in our study. The whole model was jointly fine-tuned on three general domain datasets: OntoNotes [20], PreCo [27], and LitBank [28]. Further details are available in the original studies [23,29].

3.2. Data sources

We identified three public datasets relevant to this study, including OntoNotes 5.0 [20], i2b2 [30] and MIMIC-CXR (Medical Information Mart for Intensive Care Chest X-ray) [31,32].

OntoNotes 5.0 [20] is a widely-used general-domain dataset for developing CR models. We used its English subset which has 3,493 manually annotated documents comprising seven genres of texts such as news weblogs, broadcast and so on. There are 2808 training, 343 validation, and 348 test documents.

Recall that our aim is to use radiology reports as a case study to demonstrate how a CR model can be fine-tuned to perform well in a new clinical domain. Therefore, we need to introduce clinical data into the training phase. The fifth i2b2/VA challenge [30] was based on a clinical-domain dataset with manually annotated coreferences. It originally contained 814 de-identified hospital discharge summaries, but only 424 of these documents were available for download via DBMI Data Portal [33]. We split the dataset into training, validation and test sets, which contain 296, 84 and 44 documents, respectively.

As discharge summaries differ in their content and structure from radiology reports, we naturally wanted to incorporate radiology reports into the training data but also evaluate the performance of the CR model in this domain. Not surprisingly, we were not able to identify a pre-annotated dataset in this domain. The most appropriate raw data set available was MIMIC-CXR, which consists of 377,110 images and free-text report pairs corresponding to 227,835 radiographic exams. Only the reports were used in our study. More specifically, we segmented them into sections, and only retained the *Findings* and the *Impression* sections, which provided a total of 156,011 and 189,465 text snippets, respectively. The raw text was tokenized by spaCy [9]. The following section describes how this dataset was annotated.

3.3. Data annotation

In machine learning, annotated data are typically used to evaluate the performance of predictive models. A gold standard is commonly created by manual annotation whose quality is measured by inter-annotator agreement. Alternatively, a silver standard can be created by annotating data automatically by one or more appropriate models.

3.3.1. Silver standard

Zhang et al. [17] successfully transferred a general model of linguistic analysis to the clinical domain. They did so by adding automatically annotated data from MIMIC-III [34] to manually annotated English Web Treebank [35] in order to re-train the model [36] as shown in Fig. 2. We followed this approach to create a silver standard we refer to as MIMIC-Silver-Neural using Longdoc as it demonstrated better generalization ability than the SpanBERT + c2f model [8,22]. Both approaches have already been described in the related work.

The most problematic aspect of this approach is the possibility of annotation errors introduced by the original model, which may be accumulated by the re-trained model. To improve the accuracy of the silver standard, we employed a variety of models for annotation and combined their results using an ensemble approach as illustrated in Fig. 2. The key idea is that when models trained on different data using different methods make the same prediction, the prediction itself is more likely to be correct.

Many open-source NLP tools support the functionality of CR. Most recent implementations have opted for a neural-based approach to CR, but some tools, including CoreNLP [4], still offer alternative approaches. We selected a subset of CR models that employ fundamentally different approaches including a neural approach, a statistical machine learning approach and a traditional rule-based approach. We have already described the chosen neural-based approach, which was used to create MIMIC-Silver-Neural as the first silver standard. As for the machine learning model, we used a two-stacked model proposed by Clark and Manning [5], including a logistic classifier for mention pair prediction and an agent based on an imitation learning algorithm [37] to merge coreferring pairs. Finally, we used a multi-pass sieve model originally proposed by Raghunathan et al. [6] as a representative of rule-based CR approaches.

Before we can reconcile CR predictions made by the three models, we need to understand how coreferences are formally represented. The results of CR are represented as clusters of coreferring mentions. Each cluster contains one or more mentions that refer to the same entity. Each mention is simply a span of text that consists of one or more tokens. Since the output of a CR system can be decomposed into three levels, an ensemble algorithm is designed bottom-up to consider outputs across all levels. The pseudocode of this algorithm is provided in Supplementary material. Fig. 3 exemplifies the key phases of the algorithm.

The boundaries of automatically recognized coreferring mentions may vary across different models. For example, in the sentence *"There is volume loss in both lower lobes. Compared to the prior study, the volume loss has increased."*, one of the coreference clusters contains *"volume loss in both lower lobes"* and *"the volume loss"*, yet it is possible that a CR system clustered *"volume loss"* and *"the volume loss"*. To reconcile these differences, we proposed a token-level voting. A token receives a vote whenever it appears in the output of a CR model. In this way, each mention has an average vote ratio based on the votes its tokens received, ranging from 0 to 1. A value of 1 means that all the CR systems agree that it represents a valid coreferring mention, and 0 indicates that the mention is invalid. Considering that we had three independent CR systems as raters, we set a token-level threshold to 0.66 to include any mention accepted by the majority of raters.

Once less likely individual mentions have been removed, the problem of reconciling mention clusters still remains. We first extract all pairs of mentions that belong to the same cluster. Each pair then receives a vote whenever it occurs in any of the CR systems' outputs. The higher the number of votes, the more likely that the two mentions corefer. Again, the threshold was set to 0.66 to indicate the majority of votes. Finally, we group back the remaining mention pairs to form clusters of coreferring mentions. We used the results to create another silver standard we refer to as MIMIC-Silver-Ensemble. The size of both silver standards was chosen according to the training configuration of the Longdoc model on OntoNotes, where a total of 1000, 344 and 348 documents were used for training, validation and testing, respectively.

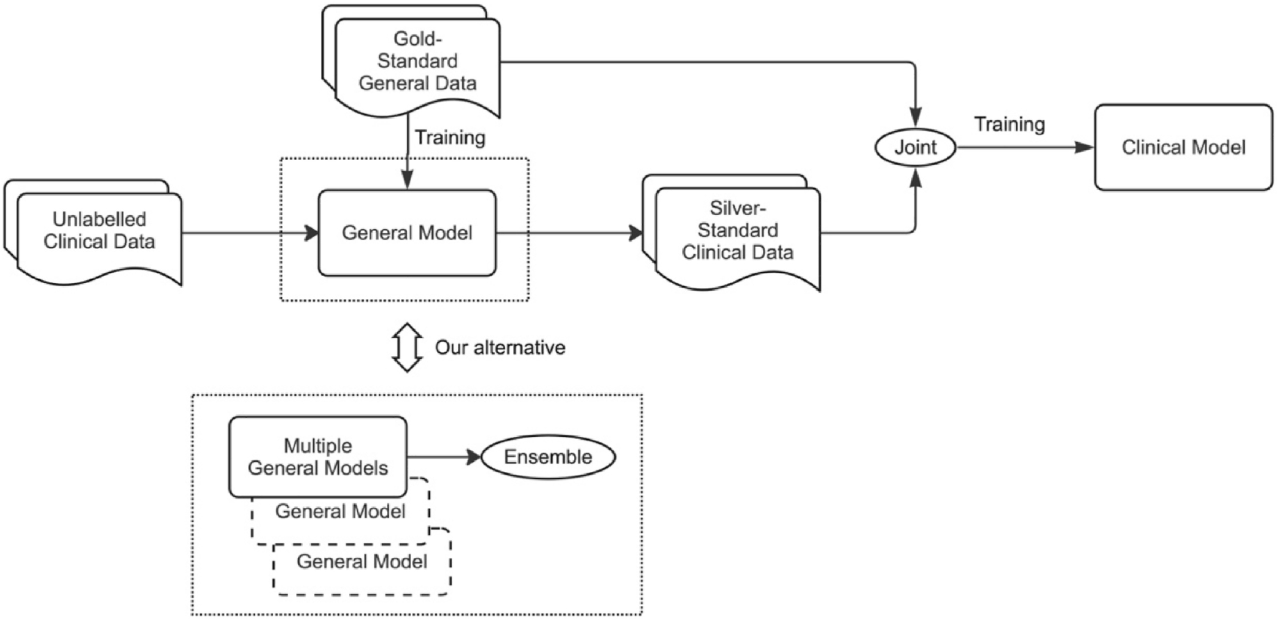
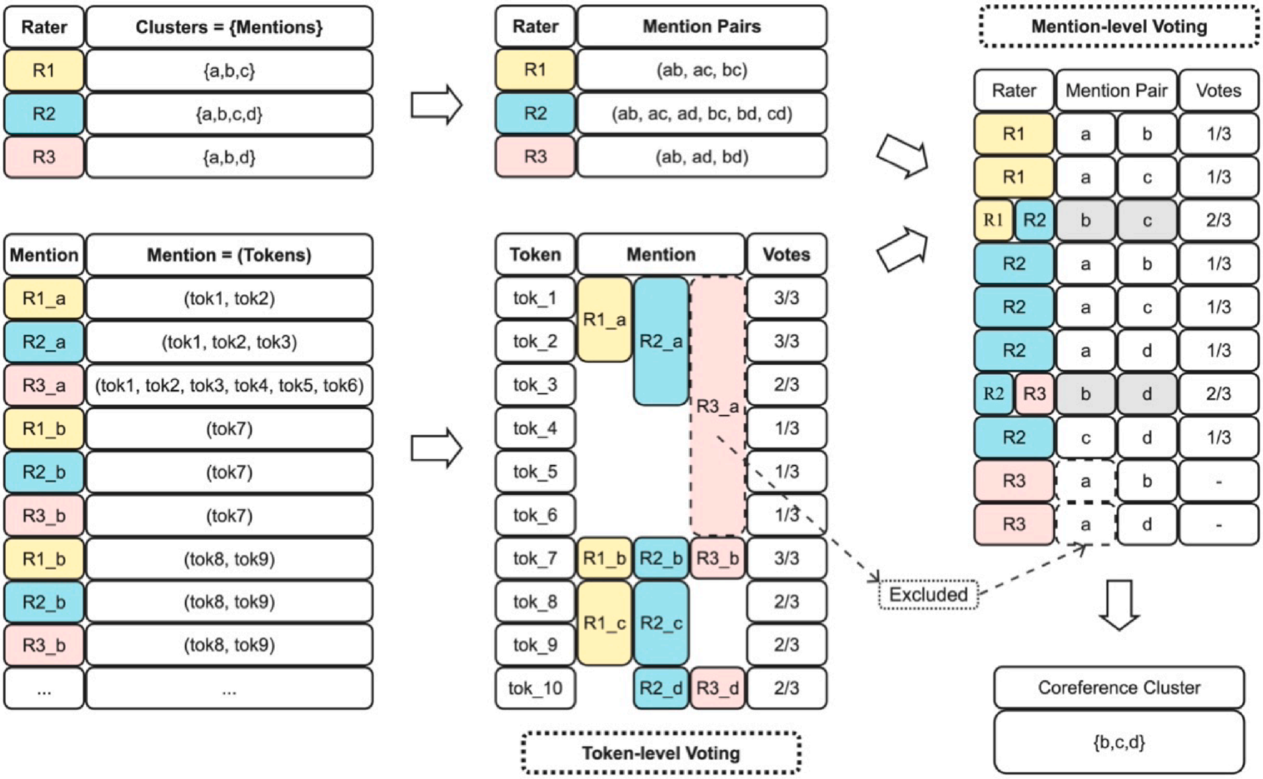


Fig. 2. The original silver standard approach for syntactic analysis [17] (above) and our adaptation for coreference resolution (below).



Rater	Clusters = {Mentions}
R1	{a,b,c}
R2	{a,b,c,d}
R3	{a,b,d}

Rater	Mention Pairs
R1	(ab, ac, bc)
R2	(ab, ac, ad, bc, bd, cd)
R3	(ab, ad, bd)

Mention	Mention = (Tokens)
R1_a	(tok1, tok2)
R2_a	(tok1, tok2, tok3)
R3_a	(tok1, tok2, tok3, tok4, tok5, tok6)
R1_b	(tok7)
R2_b	(tok7)
R3_b	(tok7)
R1_b	(tok8, tok9)
R2_b	(tok8, tok9)
R3_b	(tok8, tok9)
...	...

Token	Mention			Votes
tok_1	R1_a	R2_a	R3_a	3/3
tok_2				3/3
tok_3				2/3
tok_4				1/3
tok_5				1/3
tok_6				1/3
tok_7	R1_b	R2_b	R3_b	3/3
tok_8	R1_c	R2_c		2/3
tok_9				2/3
tok_10		R2_d	R3_d	2/3

Token-level Voting

Mention-level Voting

Rater	Mention Pair		Votes
R1	a	b	1/3
R1	a	c	1/3
R1 R2	b	c	2/3
R2	a	b	1/3
R2	a	c	1/3
R2	a	d	1/3
R2 R3	b	d	2/3
R2	c	d	1/3
R3	a	b	-
R3	a	d	-

Coreference Cluster
{b,c,d}

Fig. 3. A run-though example of the ensemble algorithm. The colors indicate the outputs of different models shown here as raters.

3.3.2. *Gold standard*

A silver standard is automatically annotated, but not manually verified. Manual annotation of the gold standard data is a known bottleneck in machine learning approaches to clinical NLP [13]. Nonetheless, gold standards are regarded more reliable as they are created by experts who possess the knowledge necessary to interpret and follow a specific set of annotation guidelines. To reduce the time and resources associated with manual annotation, we extended the idea used to create the silver standard to pre-annotate the data and manually curate these annotations and making any additional annotations if necessary. We named it a semi-automatic annotation approach. This serves to improve not only the efficiency of manual annotation but also its accuracy. Namely, we noticed that annotators commonly miss some coreferring mentions. Although employing multiple independent human annotators could alleviate this problem, false negatives will still accumulate as the result of fatigue incurred by long-drawn-out annotation.

Fig. 4 illustrates our approach to semi-automatic annotation. We first sampled a small amount of data and manually annotated them. We then

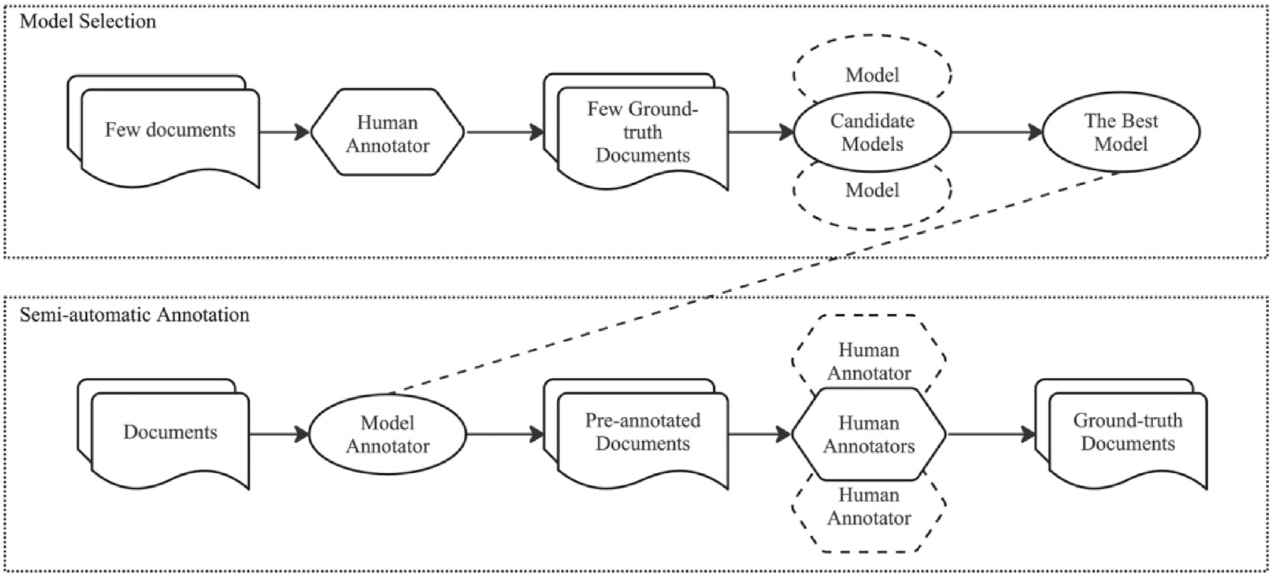


Fig. 4. The semi-automatic annotation pipeline and the corresponding model selection process.

used them to evaluate the performance of three existing CR models as well as the models previously fine-tuned on a silver standard. We selected the most accurate one, which was a Longdoc model that was jointly fine-tuned on OntoNotes and i2b2 data, to pre-annotate the data. These data were then imported into the Brat Rapid Annotation Tool (BRAT), a web-based tool for text annotation [38], where they were manually curated.

To randomly sample documents for annotation, we first stratified them by their estimated number of coreference clusters, which were automatically annotated using the Longdoc model. We first excluded documents that contained no coreference clusters. We then randomly sampled documents from each stratum.

As an alternative to random sampling, we also considered active learning, which is considered to be effective in reducing annotation cost [39]. Following Yuan et al. [18], we divided the dataset into an annotated pool and an unannotated pool and split the model training process into multiple iterations. Each iteration used an updated annotated pool to train the same initial model. Subsequently, the trained model selected top-k samples that may help it gain the most performance improvement for manual annotation and added them to the annotated pool for the next iteration.

For the query strategy, we adapted the highest mention detection entropy proposed by Yuan et al. [18]. Given mention span x, let $X \in \{0, 1\}$ indicating whether x is a mention, the highest mention detection entropy is:

$$H(x) = -\sum_{i=0}^{1} P(X = i) log P(X = i), \tag{3}$$

where the probability $P(X)$ is computed as $f(s_m(x), 0)$ in which f is a softmax function, $s_m(x)$ is a mention score function from Equation (1) and 0 denotes the threshold to determine whether x is a mention.

Overall, three gold standards were created, which we refer to as MIMIC-Gold-Random, MIMIC-Gold-Active-Learning and MIMIC-Gold-Test. MIMIC-Gold-Random and MIMIC-Gold-Active-Learning were annotated by a single human annotator in an attempt to mimic an agile development scenario where only one human annotator is available. Both MIMIC-Gold-Random and MIMIC-Gold-Active-Learning were sampled from the same subset of MIMIC-CXR. MIMIC-Gold-Random was created by interval random sampling from the stratified dataset (see above). MIMIC-Gold-Active-Learning was created by an entropy-based sampling method from the whole dataset. MIMIC-Gold-Random was annotated in five iterations of 100 documents, resulting in a total of 500 annotated documents. MIMIC-Gold-Active-Learning was annotated in 13 iterations, resulting in a total of 475 annotated documents.

The feasibility of using a single annotator for providing training data was evaluated by testing the model on data annotated by multiple annotators. This dataset, called MIMIC-Gold-Test, contains 200 documents. For each document, two human annotators were asked to annotate it independently. Their results were merged manually by the third annotator who also had access to the pre-annotated document. Any disagreements were solved by discussion. The annotators were trained for 10 min to use BRAT and to familiarise with the annotation schema.

The annotated dataset described here has been shared with the community on the PhysioNet platform where it is accessible by registered users who completed their credentialing process and signed a data use agreement [40].

3.4. Model training

The original model can be re-trained by combining newly annotated domain-specific data with the original training data. Different training strategies can be used to fine-tune the model on domain-specific data. Joint training is a simple yet effective method that collates multiple datasets to create a new dataset and use it to train a new model from scratch. This method can effectively improve the generalization ability of a CR model [25]. Alternatively, continued training utilizes a pre-trained model to initialize a model to be fine-tuned on a target dataset. It has been successfully used for rapid transfer of CR models from one dataset to another [41].

3.5. Evaluation methods

3.5.1. Data annotation

Krippendorff's alpha [42] is commonly used to measure the inter-annotator agreement (IAA), but in its original form it is too rigid for measuring the agreement on co-reference clusters as it only considers whether two clusters are identical or not. However, it can easily accommodate distance metrics to assign different weights to different relationships between coreference clusters. We adopted a distance metrics proposed by Passonneau [43], which considers four binary relationships between coreference clusters, including identical, subsume, intersect and disjunct. They are assigned the weights of 0, 0.33, 0.67 and 1, respectively. In the remainder of the article, we refer to these two versions of Krippendorff's alpha as weighted and unweighted Krippendorff's alpha, respectively.

3.5.2. Model performance

Most commonly, CR models are evaluated using the average F1-score among MUC, B^3 and CEAF$_e$ as proposed in CoNLL-2012 Shared Task [44]. We followed this convention in our study. Given a set of ground truth entities $GT = \{e_1^{GT}, e_2^{GT}, \cdots, e_N^{GT}\}$ and a set of predicted entities $PD = \{e_1^{PD}, e_2^{PD}, \cdots, e_N^{PD}\}$, where each entity corresponds to a coreference cluster, which consists of a set of mentions $e_i = \{m_{i,1}, m_{i,2}, \cdots, m_{i,K}\}$, these measures are calculated as follows:

$$P_{MUC} = \frac{\sum_{i=1}^{|PD|}\left(|e_i^{PD}| - \left|cluster\left(e_i^{PD}, e_i^{GT}\right)\right|\right)}{\sum_{i=1}^{|PD|}\left(|e_i^{PD}| - 1\right)}, R_{MUC} = \frac{\sum_{i=1}^{|GT|}\left(|e_i^{GT}| - \left|cluster\left(e_i^{GT}, e_i^{PD}\right)\right|\right)}{\sum_{i=1}^{|GT|}\left(|e_i^{GT}| - 1\right)} \tag{6}$$

$$P_{B^3} = \frac{\sum_{i=1}^{|PD|}\sum_{j=1}^{|GT|}\frac{\left|e_i^{PD}\cap e_j^{GT}\right|^2}{\left|e_i^{PD}\right|}}{\sum_{i=1}^{|PD|}\left|e_i^{PD}\right|}, R_{B^3} = \frac{\sum_{i=1}^{|PD|}\sum_{j=1}^{|GT|}\frac{\left|e_i^{PD}\cap e_j^{GT}\right|^2}{\left|e_j^{GT}\right|}}{\sum_{i=1}^{|GT|}\left|e_i^{GT}\right|} \tag{7}$$

$$P_{CEAF_e} = \frac{\sum_{e_i^{PD}, e_j^{GT}\in align(PD,GT)}\frac{2\times\left|e_i^{PD}\cap e_j^{GT}\right|}{\left|e_i^{PD}\right| + \left|e_j^{GT}\right|}}{|PD|}, R_{CEAF_e} = \frac{\sum_{e_i^{PD}, e_j^{GT}\in align(PD,GT)}\frac{2\times\left|e_i^{PD}\cap e_j^{GT}\right|}{\left|e_i^{PD}\right| + \left|e_j^{GT}\right|}}{|GT|} \tag{8}$$

where $cluster\left(e_i^{PD}, e_i^{GT}\right)$ groups the mentions in e_i^{PD} according to the mentions in e_i^{GT} and $align(PD, GT)$ represents the optimal one-to-one mapping between PD and GT using the Kuhn–Munkres algorithm [45], where e_i^{PD} aligns to at most one e_i^{GT}.

4. Experiments and results

4.1. Data annotation

We compared two annotators against each other. We also compared each annotator against the agreed ground truth. The corresponding results are shown in Fig. 5. Not surprisingly, the IAA was consistently higher for the semi-automatic approach. The semi-automatic approach also improved the efficiency of human annotators. Of note, data were annotated manually ahead of the semi-automatic annotation, which may have contributed to the efficiency of the annotators as they

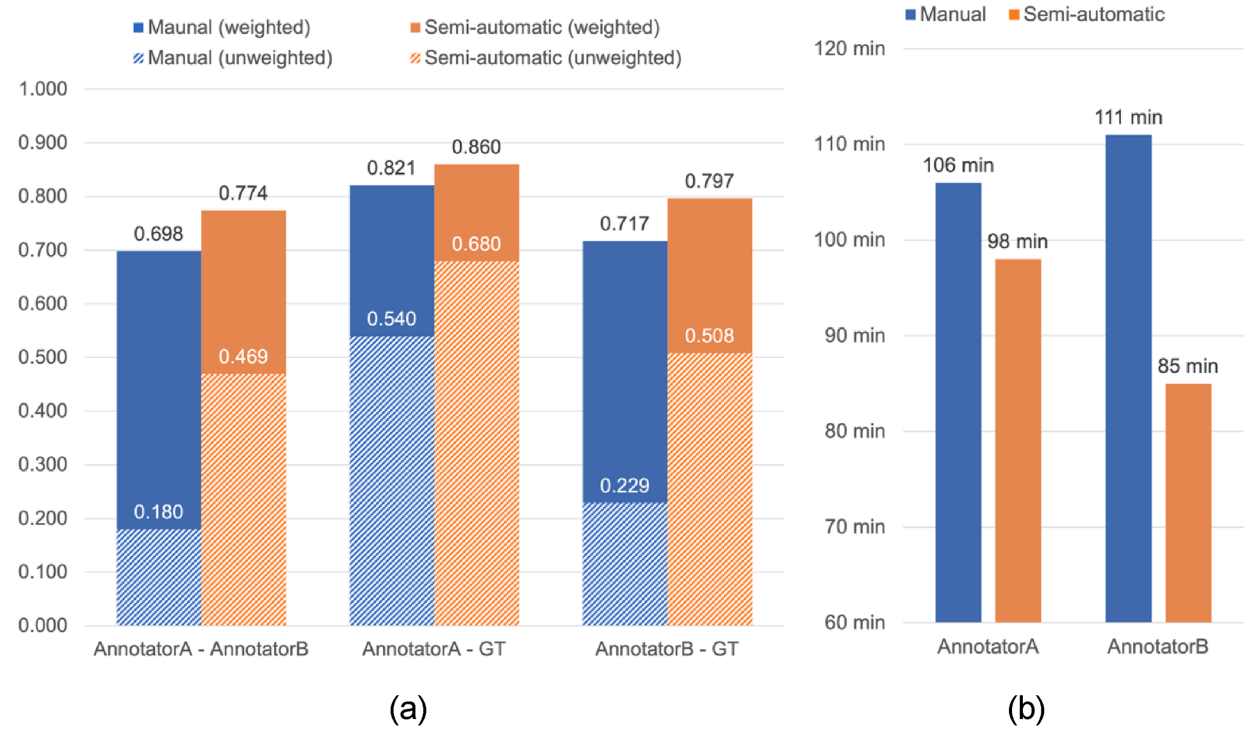


Fig. 5. (a) The inter-annotator agreement measured by Krippendroff's alpha (b) Time spent on annotation. Ground truth (GT) was created by reconciling the results of the two annotators A and B by the third independent annotator. Weighted and unweighted stand for the corresponding versions of Krippendroff's alpha.

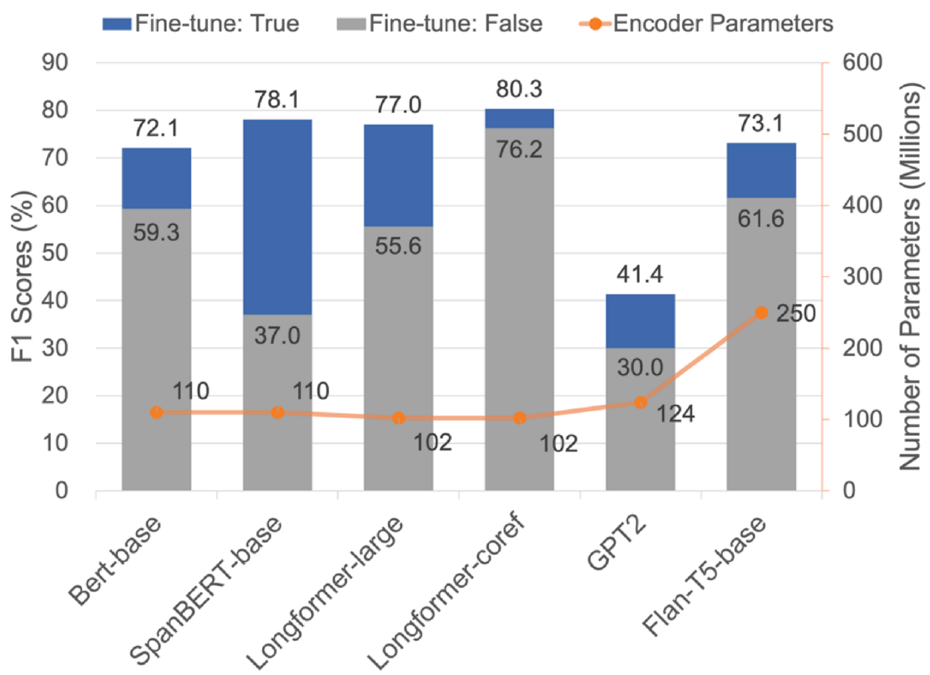


Fig. 6. CR performance using various LLM as document encoders. The CR models were trained on MIMIC-Gold-Random and tested on MIMIC-Gold-Test. The True/False values of the fine-tune option indicate whether the encoder's parameters were updated/frozen during training.

familiarized themselves with the annotation process. On the other hand, we noticed that the design of BRAT visualization negatively affected the efficiency of coreference annotation, especially for longer documents featuring a larger number and size of coreference clusters. Therefore, an annotation tool better suited for the CR task may be able to better exploit the increased efficiency that comes with the semi-automatic approach.

4.2. Large language model

Longdoc has three components: a document encoder, a mention proposer and a mention cluster predictor. It uses a Longformer-large model [25] as a document encoder. In the context of the recent advances in LLMs, we performed a series of experiments by replacing the original encoder and evaluating the performance of the corresponding fine-tuned CR model. We considered a total of six LLMs including: (1) BERT [46], an early transformer encoder, (2) SpanBERT [8], which is designed to better represent and predict text spans, (3) Longformer [25], which is based on RoBERTa [26] and is able to handle longer documents, (4) Longformer-coref [23], a fine-tuned version of Longformer for general domain CR, (5) GPT2 [47], a generative model based on the transformer decoder, and (6) Flan-T5 [48], an encoder-decoder model trained by prompting for a wide range of NLP tasks.

Fig. 6 shows that Longformer-coref outperformed all other encoders. Longformer-coref accepts input consisting of 4096 tokens, which is far longer than 1024 tokens accepted by GPT2 and 512 tokens accepted by the remaining encoders. This means that Longformer-coref can utilize longer context for the training of CR. Interestingly, the results suggest that a generative model such as GPT2 is not fit for the task at hand. Interestingly, the higher number of parameters in LLM (e.g. GPT2 and Flan-T5) does not necessarily improve the performance on specific NLP tasks such as CR.

4.3. Model training

We designed 13 experiments whose settings and the corresponding results evaluated against the MIMIC-Gold-Test dataset are shown in

Table 1
Experimental settings with respect to the training data, training methods and initial models. "√" indicates the dataset used for training (fine-tuning). The performance is given as F1-score based on the three metrices described in Section 3.5.2. These values were then averaged to provide an overall performance. In the initial models of experiments EX1-EX7, all parameters apart from the document encoder were set randomly.

Experiment	OntoNotes	i2b2	MIMIC				Training		F1			
			Silver		Gold							
			Neural	Ensemble	Random	Active learning	Method	Initial model	MUC	B^3	$CEAF_e$	Avg
EX1	√	√					Joint	Random	64.6	65.4	67.5	65.9
EX2	√		√				Joint	Random	61.6	64.2	68.0	64.6
EX3	√	√	√				Joint	Random	61.2	64.1	68.2	64.5
EX4	√			√			Joint	Random	61.1	61.3	63.2	61.9
EX5	√	√		√			Joint	Random	58.1	58.2	59.6	58.6
EX6					√		General	Random	74.0	74.9	76.5	75.1
EX7	√	√			√		Joint	Random	76.4	77.4	79.1	77.6
EX8					√		Continued	Longdoc	76.0	77.0	78.6	77.2
EX9					√		Continued	EX1	75.9	76.9	78.9	77.2
EX10					√		Continued	EX2	76.2	76.8	78.4	77.2
EX11					√		Continued	EX4	77.0	78.1	79.9	78.4
EX12						√	Continued	EX4	76.5	77.6	79.7	77.9
EX13					√	√	Continued	EX4	77.8	79.2	81.4	79.5

Table 1. Experiment EX1 sets the baseline by incorporating readily available annotated clinical datasets but without using data (MIMIC) from the target domain (radiology reports). Experiments EX2-EX5 were designed to observe the effect of the silver standard on fine-tuning a model using joint training. Experiment EX6 sets the baseline for the remaining experiments that trained CR models using the gold standard data. It does not take advantage of either joint or continued training. Experiment EX7 combines readily available annotated clinical datasets with data (MIMIC) from the target domain (radiology reports) using joint training. Experiments EX8-EX11 were designed to explore the effectiveness of continued training and the impact of different initial models on such training. Experiments EX11-EX12 explored the effectiveness of active learning for training data sampling. Finally, experiment EX13 combined the two gold standards to train the final model. All experiments used Longformer-coref to encode documents except for experiment EX8, whose initial model (Longdoc) uses Longformer to encode documents by default.

With F1 = 58.6–64.6 %, the silver standard approach (see rows EX2-EX5) proved ineffective in fine-tuning a model relative to the baseline (see row EX1, F1 = 65.9 %). Nevertheless, when the corresponding model (e.g. EX4) was used to initialize continued training (see row EX11), the performance (F1 = 78.4 %) improved compared to other models trained on gold standard (see rows EX8-EX10, F1 = 77.2 %).

When training data are limited in size due to the manual data annotation bottleneck, we need to optimize the ratio between the subsets used for training and validation, respectively. We used the settings of experiment EX11, to conduct further experiments with different splits of data into training and validation subsets. Starting with our default of 100 out of 500 documents used for validation, we reduced the number of documents reserved for validation and evaluated the corresponding models on the test data. Fig. 7 provides the corresponding results. Even though some authors suggest maximizing the number of training documents at the expense of the validation ones, going as far as using a single document for validation (e.g. [41]), our results suggest that a minimum of 25 validation documents were required. Any further increases did not improve the performance. However, it is worth noticing that our dataset is very different from the PreCo dataset used by Xia and Van Durme [41]. MIMIC-Gold-Random has on average 101 words and 4 mentions per document, while PreCo has 330 words and 105.6 mentions per document [27]. Therefore, both approaches are in agreement that at least 100 (25 × 4 ≈ 1 × 105.6) mentions should be used to validate the model. Nonetheless, we recommend conducting a series of experiments with different training/validation splits to optimize the performance of a fine-tuned CR model.

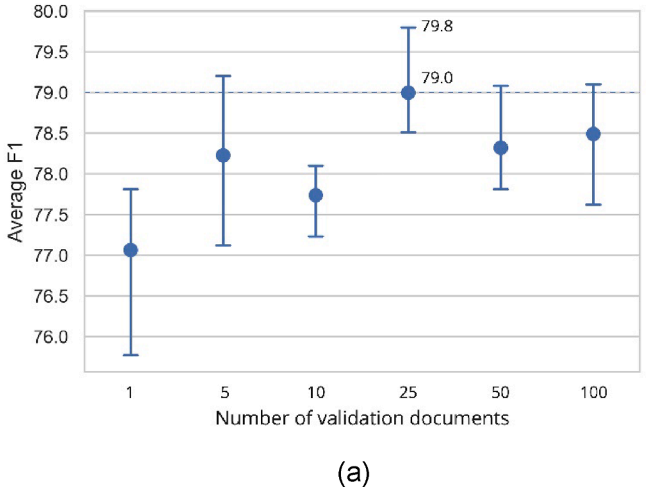


Fig. 7. The effect of training/validation split ratios on model performance. The validation documents were sampled from a total of 500 documents. The remaining documents were used for training. Each bar indicates the polar and mean values via short lines and a circular point, respectively.

Table 2
External baseline results. The performance is given as F1-score based on the three metrices described in Section 3.5.2. These values were then averaged to provide an overall performance.

Model	OntoNotes	MIMIC	Library
Deterministic [6,50–52]	59.3	46.7	CoreNLP [4]
Statistical [5]	63.0	51.6	CoreNLP [4]
Neuralcoref [53]	63.9	50.1	spaCy [9]
c2f + SpanBERT [8]	79.6	64.9	AllenNLP [10]
Longdoc + Longformer [23]	79.6	64.1	
LingMess + Longformer [54]	81.4	63.4	
CAW + RoBERTa [55]	81.6	69.1	

Starting from the experiment EX11, we changed the ratio between training/validation sets from 400:100 to 475:25. This improved the F1-score of EX11 from 78.4 % to 79.8 % (see Fig. 7). The same settings were then used to conduct experiment EX12 on gold standard data sampled using active learning. We can see that at F1 = 79.8 % random sampling outperformed active learning (see row EX12, F1 = 77.9 %) even when both gold standards were combined (see row EX13, F1 = 79.5 %). From the results reported in Table 1, we can see that the best performance was just below 80 %. Without any fine-tuning, the original Longdoc model achieved only F1 = 64.1 %.

Finally, we wanted to compare the performance of our fined-tuned model to external baseline methods (see Table 2). We selected several high-performing models from a community leaderboard, which is maintained specifically for coreference resolution on OntoNotes [49]. Given that these models have been routinely employed by open-source NLP tools [4,9,10], we also included other CR models supported by these tools. For example, the most recent update found in AllenNLP [10] was based on SpanBERT [8] and c2f-coref model [22], which achieved 79.6 % F1-score on OntoNotes.

The results presented in Table 2 compare the performance on OntoNotes against the performance achieved on the MIMIC-Gold-Test. First of all, we can observe a significant drop in performance, which is consistently over 10 percent points across all models. This re-iterates the need to fine-tune CR models for different styles of clinical narratives. Focusing our attention on the performance on MIMIC data, we can see that the best performance was still below 70 %. Specifically, deterministic CR methods [6,43–45] performed poorly achieving only F1 = 46.7 % on average. At F1 = 51.6 %, a statistical CR model [5] performed only slightly better. The best neural models with performance on OntoNotes around 80 % (the last four rows), failed to reproduce these results on MIMIC data. On the other hand, at F1 = 80.6 %, our best result obtained by combining three fine-tuned models EX11-EX13, was in line with the state-of-the-art results (F1 = 79.6 %-81.6 %). Therefore, we conclude that our approach was successful in fine-tuning a generic CR model for clinical domain.

To summarize, the most practical route to rapidly fine-tuning an existing CR model is to randomly sample data and pre-annotate them with the given model or, even better, with an ensemble of different models. Use these silver-standard annotations to estimate the number of coreferring mentions aiming to reserve at least 100 mentions for validation. If necessary, randomly sample additional data for training. Use the silver standard together with other relevant, readily available, annotated data, to fine-tune an initial model. Manually curate the silver standard to establish a gold standard. Continue training the initial model on the gold standard to obtain the final CR model.

5. Conclusion

We discussed a practical challenge of quickly utilizing existing models to address CR on an unseen dataset. We first attempted to fine-tune a CR model with a silver standard, which was assembled without human intervention. Our experiments asserted that this approach is not suitable for fine-tuning a CR model. However, the by-product of applying this method – an ensemble algorithm that we designed to reduce the bias of any single model annotation – achieves a significant improvement in precision albeit at the expense of the recall. We then designed a semi-automatic annotation approach that can improve the consistency and efficiency of manual annotation.

We created two small gold-standard datasets via random sampling and active learning respectively and compared their performance. We discovered that active learning should be applied with caution. Applying active learning in practice is challenging because datasets with different characteristics are sensitive to query strategies. The feasible matching between datasets and query strategies is still unclear. In a low-resource scenario, an arbitrary choice of query strategy does not necessarily outperform random sampling. Based on our experiments, there may be a compromise when it is difficult to choose the right query strategy, which first uses randomly sampled data to make the model have a stable decision boundary and then employs active learning to refine the decision boundary and increase the upper limits of the performance.

We observed that in low-resource settings without introducing silver standards and ensemble methods, the performance of continued training is typically close to joint training, yet both approaches outperform the training from scratch. We conclude that using continued training with a relatively small, annotated dataset is adequate for transferring a CR model.

Finally, we conclude that the semi-automatic annotation approach combined with continued training is well suited for the rapid transfer of CR models under low-resource conditions. The ensemble approach has the potential to further enhance the quality of model outputs when a set of transferred models are available. Overall, we have effectively transferred a general CR model to the clinical domain and comprehensively demonstrated our outcomes as well as any setbacks encountered.

6. Statement of significance

Problem:

Existing open-source Coreference Resolution (CR) tools may not generalize well on unseen clinical data.

What is already known:

Existing CR models are typically generic and as such their performance does not necessarily transfer into clinical domains. Even within a single domain, sublanguages can vary, which means that when a model is trained within a domain, its performance may still vary across different types of documents.

What this paper adds:

This study has effectively transferred a general CR model to the clinical domain. Our findings based on extensive experimentation have been summarized into practical suggestions to the research community in terms of rapidly applying CR to different styles of clinical narratives.

CRediT authorship contribution statement

Yuxiang Liao: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Hantao Liu:** Data curation, Funding acquisition, Supervision, Writing – review & editing. **Irena Spasić:** Data curation, Formal analysis, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yuxiang Liao reports a relationship with China Scholarship Council that includes: funding grants.

Acknowledgements

This work is part of a PhD project funded by China Scholarship Council-Cardiff University Scholarship (CSC202108140022). The scholarship had been awarded to Y.L. The project is supervised by I.S. and H.L.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104578.

References

[1] D. Ganeshan, P.-A.T. Duong, L. Probyn, et al., Structured reporting in radiology, Acad. Radiol. 25 (1) (2018) 66–73, https://doi.org/10.1016/j.acra.2017.08.005.

[2] M. Pourreza Shahri, A. Tahmasebi, B. Ye, H. Zhu, J. Aslam, T. Ferris, An Ensemble Approach for Automatic Structuring of Radiology Reports, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, 2020, pp. 249–258.

[3] P. Lu, M. Poesio, Coreference Resolution for the Biomedical Domain: A Survey, in: Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 12–23.

[4] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland, Association for Computational Linguistics, 2014, pp. 55–60.

[5] K. Clark, C.D. Manning, Entity-Centric Coreference Resolution with Model Stacking, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics, Beijing, China, 2015, pp. 1405–1415.

[6] K. Raghunathan, H. Lee, S. Rangarajan, et al., A Multi-Pass Sieve for Coreference Resolution, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA, Association for Computational Linguistics, 2010, pp. 492–501.

[7] K. Clark, C.D. Manning, Improving Coreference Resolution by Learning Entity-Level Distributed Representations, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Association for Computational Linguistics, 2016, pp. 643–653.

[8] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: improving pre-training by representing and predicting spans, Trans. Assoc. Comput. Linguist. 8 (2020) 64–77, https://doi.org/10.1162/tacl_a_00300.

[9] Explosion, SpaCy: Industrial-Strength Natural Language Processing, https://spacy.io/ (accessed 16 Nov 2022).

[10] The Allen Institute for Artificial Intelligence, AllenNLP, https://allenai.org/allennlp (accessed 17 Nov 2022).

[11] I. Temnikova, W.A. Baumgartner Jr., N.D. Hailu, et al., Sublanguage Corpus Analysis Toolkit: A Tool for Assessing the Representativeness and Sublanguage Characteristics of Corpora, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA), 2014, pp. 1714–1718.

[12] C. Friedman, P. Kra, A. Rzhetsky, Two biomedical sublanguages: a description based on the theories of Zellig Harris, J. Biomed. Inf. 35 (4) (2002) 222–235, https://doi.org/10.1016/s1532-0464(03)00012-1.

[13] I. Spasic, G. Nenadic, Clinical text data in machine learning: systematic review, JMIR Med. Inform. 8 (3) (2020), e17984, https://doi.org/10.2196/17984.

[14] K.B. Wagholikar, H. Estiri, M. Murphy, S.N. Murphy, Polar labeling: silver standard algorithm for training disease classifiers, Bioinformatics 36 (10) (2020) 3200–3326, https://doi.org/10.1093/bioinformatics/btaa088.

[15] A. Oellrich, N. Collier, D. Smedley, T. Groza, Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes, PLoS One 10 (1) (2015), e0116040.

[16] I. Korkontzelos, D. Piliouras, A.W. Dowsey, S. Ananiadou, Boosting drug named entity recognition using an aggregate classifier, Artif. Intell. Med. 65 (2) (2015) 145–153.

[17] Y. Zhang, Y. Zhang, P. Qi, C.D. Manning, C.P. Langlotz, Biomedical and clinical English model packages for the Stanza Python NLP library, J. Am. Med. Inform. Assoc. 28 (9) (2021) 1892–2189, https://doi.org/10.1093/jamia/ocab090.

[18] M. Yuan, P. Xia, C. May, B. Van Durme, J. Boyd-Graber, Adapting Coreference Resolution Models through Active Learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, Association for Computational Linguistics, 2022, pp. 7533–7549.

[19] K. Lee, L. He, M. Lewis, L. Zettlemoyer, End-to-end Neural Coreference Resolution, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 188–197.

[20] R. Weischedel, M. Palmer, M. Marcus, et al., OntoNotes Release 5.0, Linguistic Data Consortium, 2013.

[21] V. Dobrovolskii, Word-Level Coreference Resolution, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021, pp. 7670–7675.

[22] K. Lee, L. He, L. Zettlemoyer, 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), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 687–692.

[23] S. Toshniwal, P. Xia, S. Wiseman, K. Livescu, K. Gimpel, On Generalization in Coreference Resolution, in: Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference, Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021, pp. 111–120.

[24] J. Nivre, C.-T. Fang, Universal Dependency Evaluation, in: Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), Gothenburg, Sweden, Association for Computational Linguistics, 2017, pp. 86–95.

[25] I. Beltagy, M.E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150, 2020.

[26] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A Robustly Optimized BERT Pre-training Approach with Post-training, in: Proceedings of the 20th Chinese National Conference on Computational Linguistics, Huhhot, China, Chinese Information Processing Society of China, 2021, pp. 1218–1227.

[27] H. Chen, Z. Fan, H. Lu, A. Yuille, S. Rong, PreCo: A Large-scale Dataset in Preschool Vocabulary for Coreference Resolution, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Association for Computational Linguistics, 2018, pp. 172–181.

[28] D. Bamman, O. Lewke, A. Mansoor, An Annotated Dataset of Coreference in English Literature, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, European Language Resources Association, 2020, pp. 44–54.

[29] S. Toshniwal, S. Wiseman, A. Ettinger, K. Livescu, K. Gimpel, Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 8519–8526.

[30] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, B.R. South, Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc. 19 (5) (2012) 786–791, https://doi.org/10.1136/amiajnl-2011-000784.

[31] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, et al., MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (1) (2019) 317, https://doi.org/10.1038/s41597-019-0322-0.

[32] A.E.W. Johnson, T.J. Pollard, N.R. Greenbaum, et al., MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019, https://ui.adsabs.harvard.edu/abs/2019arXiv190107042J (accessed 01 January 2019).

[33] Harvard Medical School, DBMI Data Portal, https://portal.dbmi.hms.harvard.edu/ (accessed 15 June 2023).

[34] A.E.W. Johnson, T.J. Pollard, L. Shen, et al., MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016), 160035, https://doi.org/10.1038/sdata.2016.35.

[35] A. Bies, J. Mott, C. Warner, S. Kulick, English Web Treebank, LDC2012T13, Linguistic Data Consortium, Philadelphia, 2012.

[36] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C.D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2020, pp. 101–108.

[37] S. Ross, G. Gordon, D. Bagnell, A reduction of imitation learning and structured prediction to no-regret online learning, in: G. Geoffrey, D. David, D. Miroslav (Eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, 2011, pp. 627–635.

[38] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, T. Ji, brat: a Web-based Tool for NLP-Assisted Text Annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, Association for Computational Linguistics, 2012, pp. 102–107.

[39] Z. Zhang, E. Strubell, E. Hovy, A Survey of Active Learning for Natural Language Processing, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, Association for Computational Linguistics, 2022, pp. 6166–6190.

[40] Y. Liao, H. Liu, Spasic. I. RadCoref: Fine-tuning coreference resolution for different styles of clinical narratives (version 1.0.0), PhysioNet 2023 (in press).

[41] P. Xia, B. Van Durme, Moving on from OntoNotes: Coreference Resolution Model Transfer, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, Association for Computational Linguistics, 2021, pp. 5241–5256.

[42] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, fourth edition, Thousand Oaks, California, 2019.

[43] R.J. Passonneau, Computing Reliability for Coreference Annotation, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, European Language Resources Association (ELRA), 2004, pp. 1503–1506.

[44] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, in: Joint Conference on EMNLP and CoNLL - Shared Task, Jeju Island, Korea, Association for Computational Linguistics, 2012, pp. 1–40.

[45] H.W. Kuhn, The Hungarian method for the assignment problem, Nav. Res. Logist. Q. 2 (1–2) (1955) 83–97, https://doi.org/10.1002/nav.3800020109.

[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Minneapolis, Minnesota, Association for Computational Linguistics, 2019, pp. 4171–4186.

[47] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.

[48] H.W. Chung, L. Hou, S. Longpre, et al., Scaling Instruction-Finetuned Language Models, 2022, https://ui.adsabs.harvard.edu/abs/2022arXiv221011416C (accessed October 01, 2022).

[49] Meta AI, Papers with Code: Coreference Resolution on OntoNotes - Leaderboard. (accessed 19 Nov 2023).

[50] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, D. Jurafsky, Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 28–34.

[51] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, D. Jurafsky, Deterministic coreference resolution based on entity-centric, Precision-ranked rules, Comput. Linguist. 39 (4) (2013) 885–916, https://doi.org/10.1162/COLI_a_00152.

[52] M. Recasens, M.-C. de Marneffe, C. Potts, 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, Atlanta, Georgia, Association for Computational Linguistics, 2013, pp. 627–633.

[53] K. Clark, C.D. Manning, Deep Reinforcement Learning for Mention-Ranking Coreference Models, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Association for Computational Linguistics, 2016, pp. 2256–2262.

[54] S. Otmazgin, A. Cattan, Y. Goldberg, LingMess, Linguistically Informed Multi Expert Scorers for Coreference Resolution, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, Association for Computational Linguistics, 2023, pp. 2752–2760.

[55] K. D'Oosterlinck, S. Kiros Bitew, B. Papineau, C. Potts, T. Demeester, C. Develder, CAW-coref: Conjunction-Aware Word-level Coreference Resolution, 2023, https://ui.adsabs.harvard.edu/abs/2023arXiv231006165D (accessed October 01, 2023).