Coreference Reasoning in Machine Reading Comprehension

Mingzhu Wu[1], **Nafise Sadat Moosavi**[1], **Dan Roth**[2], **Iryna Gurevych**[1]

[1]UKP Lab, Technische Universitat Darmstadt
[2]Department of Computer and Information Science, UPenn
[1]https://www.ukp.tu-darmstadt.de
[2] https://www.seas.upenn.edu/~danroth/

Abstract

Coreference resolution is essential for natural language understanding and has been long studied in NLP. In recent years, as the format of Question Answering (QA) became a standard for machine reading comprehension (MRC), there have been data collection efforts, e.g., Dasigi et al. (2019), that attempt to evaluate the ability of MRC models to reason about coreference. However, as we show, coreference reasoning in MRC is a greater challenge than earlier thought; MRC datasets do not reflect the natural distribution and, consequently, the challenges of coreference reasoning. Specifically, success on these datasets does not reflect a model's proficiency in coreference reasoning. We propose a methodology for creating MRC datasets that better reflect the challenges of coreference reasoning and use it to create a sample evaluation set. The results on our dataset show that state-of-the-art models still struggle with these phenomena. Furthermore, we develop an effective way to use naturally occurring coreference phenomena from existing coreference resolution datasets when training MRC models. This allows us to show an improvement in the coreference reasoning abilities of state-of-the-art models.[1]

1 Introduction

Machine reading comprehension is the ability to read and understand the given passages and answer questions about them. Coreference resolution is the task of finding different expressions that refer to the same real-world entity. The tasks of coreference resolution and machine reading comprehension have moved closer to each other. Converting coreference-related datasets into an MRC format improves the performance on some coreference-related datasets (Wu et al., 2020b; Aralikatte et al., 2019). There are also various datasets for the task of reading comprehension on which the model requires to perform coreference reasoning to answer some of the questions, e.g., DROP (Dua et al., 2019), DuoRC (Saha et al., 2018), MultiRC (Khashabi et al., 2018), etc.

Quoref (Dasigi et al., 2019) is a dataset that is particularly designed for evaluating coreference understanding of MRC models. Figure 1 shows a QA sample from Quoref in which the model needs to resolve the coreference relation between "his" and "John Motteux" to answer the question.

context: "In 1834, Henry Hoste Henley died without issue, and the estate was bought at auction by John Motteux, a London merchant. Motteux was also without heirs and bequeathed Sandringham, together with another Norfolk estate and a property in Surrey, to the third son of his close friend, Emily Lamb, the wife of Lord Palmerston..."
question: "What is the first name of the person who was close friends with Lamb?"
gold answer: "John"

Figure 1: A sample from the Quoref dataset.

Recent large pre-trained language models reached high performance on Quoref. However, our results and analyses suggest that this dataset contains artifacts and does not reflect the natural distribution and, therefore, the challenges of coreference reasoning. As a result, high performances on Quoref do not necessarily reflect the coreference reasoning capabilities of the examined models and answering questions that require coreference reasoning might be a greater challenge than current scores suggest.

In this paper, we propose two solutions to address this issue. First, we propose a methodology for creating MRC datasets that better reflect the coreference reasoning challenge. We release a sample challenging evaluation set containing 200 examples by asking an annotator to create new question-

[1]The code and the resulting dataset are available at https://github.com/UKPLab/coref-reasoning-in-qa.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 5768–5781
August 1–6, 2021.

answer pairs using our methodology and based on existing passages in Quoref. We show that this dataset contains fewer annotation artifacts, and its distribution of biases is closer to a coreference resolution dataset. The performance of state-of-the-art models on Quoref considerably drops on our evaluation set suggesting that (1) coreference reasoning is still an open problem for MRC models, and (2) our methodology opens a promising direction to create future challenging MRC datasets.

Second, we propose to directly use coreference resolution datasets for training MRC models to improve their coreference reasoning. We automatically create a question whose answer is a coreferring expression m_1 using the BART model (Lewis et al., 2020). We then consider this question, m_1's antecedent, and the corresponding document as a new (question, answer, context) tuple. This data helps the model learning to resolve the coreference relation between m_1 and its antecedent to answer the question. We show that incorporating these additional data improves the performance of the state-of-the-art models on our new evaluation set.

Our main contributions are as follows:

- We show that Quoref does not reflect the natural challenges of coreference reasoning and propose a methodology for creating MRC datasets that better reflect this challenge.
- We release a sample challenging dataset that is manually created by an annotator using our methodology. The results of state-of-the-art MRC models on our evaluation set show that, despite the high performance of MRC models on Quoref, answering questions based on coreference reasoning is still an open challenge.
- We propose an approach to use existing coreference resolution datasets for training MRC models. We show that, while coreference resolution and MRC datasets are independent and belong to different domains, our approach improves the coreference reasoning of state-of-the-art MRC models.

2 Related Work

2.1 Artifacts in NLP datasets

One of the known drawbacks of many NLP datasets is that they contain artifacts.[2] Models tend to exploit these easy-to-learn patterns in the early stages of training (Arpit et al., 2017; Liu et al., 2020; Utama et al., 2020b), and therefore, they may not focus on learning harder patterns of the data that are useful for solving the underlying task. As a result, overfitting to dataset-specific artifacts limits the robustness and generalization of NLP models.

There are two general approaches to tackle such artifacts: (1) adversarial filtering of biased examples, i.e., examples that contain artifacts, and (2) debiasing methods. In the first approach, potentially biased examples are discarded from the dataset, either after the dataset creation (Zellers et al., 2018; Yang et al., 2018a; Le Bras et al., 2020; Bartolo et al., 2020), or while creating the dataset (Dua et al., 2019; Chen et al., 2019; Nie et al., 2020).

In the second approach, they first recognize examples that contain artifacts, and use this knowledge in the training objective to either skip or downweight biased examples (He et al., 2019; Clark et al., 2019a), or to regularize the confidence of the model on those examples (Utama et al., 2020a). The use of this information in the training objective improves the robustness of the model on adversarial datasets (He et al., 2019; Clark et al., 2019a; Utama et al., 2020a), i.e., datasets that contain counterexamples in which relying on the bias results in an incorrect prediction. In addition, it can also improve in-domain performances as well as generalization across various datasets that represent the same task (Wu et al., 2020a; Utama et al., 2020b).

While there is an emerging trend of including adversarial models in data collection, their effectiveness is not yet compared with using debiasing methods, e.g., whether they are still beneficial when we use debiasing methods or vice versa.

2.2 Joint QA and Coreference Reasoning

There are a few studies on the joint understanding of coreference relations and reading comprehension. Wu et al. (2020b) propose to formulate coreference resolution as a span-prediction task by generating a query for each mention using the surrounding context, thus converting coreference resolution to a reading comprehension problem. They leverage the plethora of existing MRC datasets for data augmentation and improve the generalization of the coreference model. In parallel to Wu et al. (2020b), Aralikatte et al. (2019) also cast ellipsis and coreference resolution as reading comprehension tasks. They leverage the existing neural archi-

[2] I.e., the conditional distribution of the target label based on specific attributes of the training domain diverges while testing on other domains.

tectures designed for MRC for ellipsis resolution and outperform the previous best results. In a similar direction, Hou (2020) propose to cast bridging anaphora resolution as question answering and present a question answering framework for this task. However, none of the above works investigate the impact of using coreference data on QA.

Dua et al. (2020) use Amazon Mechanical Turkers to annotate the corresponding coreference chains of the answers in the passages of Quoref for 2,000 QA pairs. They then use this additional coreference annotation for training a model on Quoref. They show that including these additional coreference annotations improves the overall performance on Quoref. The proposed method by Dua et al. (2020) requires annotating additional coreference relations on every new coreference-aware QA dataset. Contrary to this, our approach uses existing coreference resolution datasets, and therefore, applies to any new QA dataset without introducing any additional cost.

3 How Well Quoref Presents Coreference Reasoning?

For creating the Quoref dataset, annotators first identify coreferring expressions and then ask questions that connect the two coreferring expressions. Dasigi et al. (2019) use a BERT-base model (Devlin et al., 2019) that is fine-tuned on the SQuAD dataset (Rajpurkar et al., 2016) as an adversarial model to exclude QA samples that the adversarial model can already answer. The goal of using this adversarial model is to avoid including question-answer pairs that can be solved using surface cues. They claim that most examples in Quoref cannot be answered without coreference reasoning.

If we fine-tune a RoBERTa-large model on Quoref, it achieves 78 F1 score while the estimated human performance is around 93 F1 score (Dasigi et al., 2019). This high performance, given that RoBERTa can only predict continuous span answers while Quoref also contains discontinuous answers, indicates that either (1) Quoref presents coreference-aware QA very well so that the model can properly learn coreference reasoning from the training data, (2) pretrained transformer-based models have already learned coreference reasoning during their pre-training, e.g., as suggested by Tenney et al. (2019) and Clark et al. (2019b), or (3) coreference reasoning is not necessarily required for solving most examples.

In this section, we investigate whether Quoref contains the known artifacts of QA datasets, and therefore, models can solve some of the QA pairs without performing coreference reasoning. Figure 2 shows such an example where simple lexical cues are enough to answer the question despite the fact that coreference expressions "Frankie" and "his" were included in the corresponding context.

context: "...In turn, Frankie stalks Ralph back to his tenement and strangles him to death following a violent brawl between the two. Losing his nerve, Frankie calls up his employers to tell them he wants to quit the job. Unsympathetic, the supervisor tells him he has until New Year's Eve to perform the hit...."
question: "What is the first name of the person who wants to quit their job?"
gold answer: "Frankie"

Figure 2: A QA example that relies on simple lexical overlap without requiring coreference reasoning.

We investigate five artifacts (biases) as follows:

- Random named entity: the majority of answers in Quoref are person names. To evaluate this artifact, we randomly select a PERSON named entity from the context as the answer.[3]
- Wh-word (Weissenborn et al., 2017): to recognize the QA pairs that can be answered by only using the interrogative adverbs from the question, we train a model on a variation of the training dataset in which questions only contain interrogative adverbs.
- Empty question (Sugawara et al., 2020): to recognize QA pairs that are answerable without considering the question,[4] we train a QA model only on the contexts and without questions.
- Semantic overlap (Jia and Liang, 2017): for this artifact, we report the ratio of the QA pairs whose answers lie in the sentence of the context that has the highest semantic similarity to the question. We use sentence-BERT (Reimers and Gurevych, 2019) to find the most similar sentence.
- Short distance reasoning: for this bias, we train a model only using the sentence of the context that is the most similar to the question, instead of the whole context. We exclude the question-answer pairs in which the most similar sentence does not contain the answer. This model will not learn to perform coreference reasoning when the related coreferring pairs are not in the same sentence.

[3] We use spaCy (Honnibal and Johnson, 2015) for NER.

[4] E.g., this can indicate the bias of the model to select the most frequent named entity in the context as the answer.

For wh-word, empty question, and short distance reasoning, we use the TASE model (Segal et al., 2020) to learn the bias. Biased examples are then those that can be correctly solved by these models. We only change the training data for biased example detection, if necessary, and the development set is unchanged. The *Quoref* column in Table 1 reports the proportion of biased examples in the Quoref development set.

Bias	**Quoref**	**CoNLL$_{bart}$**
random named entity	9.39	1.52
wh-word	22.99	13.12
empty question	21.51	11.60
semantic overlap	28.66	21.38
short-distance reasoning	**50.70**	9.86

Table 1: The proportion of examples in the Quoref development set and CoNLL-2012 coreference resolution dataset that contain each of the examined biases.

We also investigate whether these biases have similar ratios in a coreference resolution dataset. We use the CoNLL-2012 coreference resolution dataset (Pradhan et al., 2012a) and convert it to a reading comprehension format, i.e., CoNLL$_{bart}$ in Section 5.[5] This data contains question-answer pairs in which the question is created based on a coreferring expression in CoNLL-2012, and the answer is its closest antecedent. We split this data into training and test sets and train bias models on the training split. The *CoNLL$_{bart}$* column in Table 1 shows the bias proportions on this data.

As we see, the short distance reasoning is the most prominent bias in the Quoref dataset. However, the ratio of such biased examples is only around 10% in CoNLL-2012. Therefore, apart from the examples that can be solved without coreference reasoning,[6] the difficulty of the required coreference reasoning in the remaining examples is also not comparable with naturally occurring coreference relations in a coreference resolution dataset.

As a result, high performance on Quoref does not necessarily indicate that the model is adept at performing coreference reasoning.

[5]We report the bias ratios of CoNLL$_{dec}$ in Section 5 in the appendix.

[6]E.g., about 20% of examples can be answered without considering the question.

4 Creating an MRC Dataset that Better Reflects Coreference Reasoning

There is a growing trend in using adversarial models for data creation to make the dataset more challenging or discard examples that can be solved using surface cues (Bartolo et al., 2020; Nie et al., 2020; Yang et al., 2018a; Zellers et al., 2018; Yang et al., 2018b; Dua et al., 2019; Chen et al., 2019; Dasigi et al., 2019).

Quoref is also created using an adversarial data collection method to discard examples that can be solved using simple lexical cues. The assumption is that it is hard to avoid simple lexical cues by which the model can answer questions without coreference reasoning. Therefore, an adversarial model (A) is used to discard examples that contain such lexical cues. While this adversarial filtering removes examples that are easy to solve by A, it does not ensure that the remaining examples do not contain shortcuts that are not explored by A. First, the adversarial model in Quoref is trained on another dataset, i.e., SQuAD. Thus, the failure of A on Quoref examples may be due to (1) Quoref having different lexical cues than those in SQuAD, or (2) domain shift. Second, and more importantly, as argued by Dunietz et al. (2020), making the task challenging by focusing on examples that are more difficult for existing models is not a solution for more useful reading comprehension.[7]

We instead propose a methodology for creating question-answer pairs as follows:

- Annotators should create a question that connects the referring expression m_1 to its antecedent m_2 so that (1) m_2 is more informative than m_1,[8] and (2) m_1 and m_2 reside in a different sentence.
- Candidate passages for creating QA pairs are selected according to their number of named entities and pronouns. The number of distinct named entities is an indicator of the number of entities in the text. Therefore, there would be more candidate entities for resolving referring expressions. The number of pronouns indicates that we have enough candidate m_1s that have more informative antecedents.

We provide this guideline to a student from the

[7]As put by them: "the dominant MRC research paradigm is like trying to become a professional sprinter by glancing around the gym and adopting any exercises that look hard".

[8]Proper names are more informative than common nouns, and they are more informative than pronouns (Lee et al., 2013).

Context Snippet	Question	Gold Answer
"Diamonds" was certified sextuple platinum by the Recording Industry Association of America (RIAA). In Canada, the song debuted at number nine on the Canadian Hot 100 for the issue dated October 13, 2012 [...] It remained atop of it for four consecutive weeks [...]	What is the full name of the chart of which Diamonds remained atop for four consecutive weeks?	Canadian Hot 10
The ever-winding path of John Frusciante's solo career is a confusing one to say the least [...] The album of the same name is Frusciante's first experimenting with the acid house genre. He previously released an EP, Sect In Sgt under this alias in 2012.	Who did release an EP called Sect In Sgt?	John Frusciante

Table 2: Examples from our dataset. The context is cropped to only show the relevant parts.

Computer Science department for generating new QA pairs from the existing passages in the Quoref development set. We use Quoref passages to ensure that the source of performance differences on our dataset vs. Quoref is not due to domain differences. This results in 200 new QA pairs. Table 2 presents examples from our dataset.

Table 3 shows the results of the examined biases on our dataset. By comparing Table 3 and Table 1, we observe that the examined biases are less strong in our dataset, and their distribution is closer to those in CoNLL-2012. As we will see in Table 5, the performance of state-of-the-art models on Quoref drops more than 10 points, i.e., 13-18 points, on our challenge dataset.[9]

Bias	Ours
random named entity	3.03
wh-word	13.64
empty question	11.62
semantic overlap	24.50
short-distance reasoning	35.35

Table 3: Proportion of biased examples in our dataset.

5 Improving Coreference Reasoning

While we do not have access to many coreference annotations for the task of coreference-aware MRC, there are various datasets for the task of coreference resolution. Coreference resolution datasets contain the annotation of expressions that refer to the same entity. In this paper, we hypothesize that we can directly use coreference resolution corpora to improve the coreference reasoning of MRC models. We propose an effective approach to convert coreference annotations into QA pairs so that models learn to perform coreference resolution by answering those questions. In our experiments, we use the CoNLL-2012 dataset (Pradhan et al., 2012b) that is the largest annotated dataset with coreference information.

5.1 Coreference-to-QA Conversion

The existing approach to convert coreference annotations into (question, context, answer) tuples, which is used to improve coreference resolution performance (Wu et al., 2020b; Aralikatte et al., 2019), is to use the sentence of the anaphor as a declarative query, and its closest antecedent as the answer. The format of these queries is not compatible with questions in MRC datasets, and therefore, the impact of this data on MRC models may be limited. In this work, we instead generate questions from those declarative queries using an automatic question generation model. We use the BART model (Lewis et al., 2020) that is one of the state-of-the-art text generation models. Below we explain the details of each of these two approaches for creating QA data from CoNLL-2012. Table 4 shows examples from both approaches.

CoNLL$_{dec}$: Wu et al. (2020b) and Aralikatte et al. (2019) choose a sentence that contains an anaphor as a declarative query, the closest non-pronominal antecedent of that anaphor as the answer, and the corresponding document of the expressions as the context.[10] We remove the tuples in which the anaphor and its antecedent are identical. The reason is that (1) Quoref already contains many examples in which the coreference relation is between two mentions with the same string, and (2) even after removing such examples, CoNLL$_{dec}$ contains around four times more QA pairs than the Quoref training data.

CoNLL$_{bart}$: we use a fine-tuned BART model (Lewis et al., 2020) released by Durmus et al.

[9] We examine 50 randomly selected examples from our challenge set, and they were all answerable by a human.

[10] We use the code provided by Aralikatte et al. (2019).

Passage in CoNLL	Mention Cluster	CoNLL$_{dec}$ Quesion	CoNLL$_{bart}$ Question	Gold Answer
My mother was Thelma Wahl [...] She was a very good mother. She was at Huntingdon because she needed care [...]	[My mother, She, She, she]	She was at Huntingdon because <ref> she </ref> needed care.	who was at huntingdon because she needed care?	My mother
The angel also held a large chain in his hand [...] The angel tied the dragon with the chain for 1000 years.	[a large chain, the chain]	The angel tied the dragon with <ref> the chain </ref> for 1000 years.	what did the angel tie the dragon with for 1000 years?	a large chain

Table 4: Coreference-to-QA conversion examples using CoNLL$_{dec}$ and CoNLL$_{bart}$ approaches.

(2020) for question generation and apply it on the declarative queries in CoNLL$_{dec}$. The BART model specifies potential answers by masking noun phrases or named entities in the query and then generates questions for each masked text span. We only keep questions whose answer, i.e., the masked expression, is a coreferring expression and replace that answer with its closest non-pronominal antecedent. We only keep questions in which the masked expression and its antecedent are not identical. Such QA pairs enforce the model to resolve the coreference relation between the two coreferring expressions to answer generated questions.

5.2 Experimental Setups

We use two recent models from the Quoref leaderboard: RoBERTa (Liu et al., 2019) and TASE (Segal et al., 2020), from which TASE has the state-of-the-art results. We use RoBERTa-large from HuggingFace (Wolf et al., 2020). TASE casts MRC as a sequence tagging problem to handle questions with multi-span answers. It assigns a tag to every token of the context indicating whether the token is a part of the answer. We use the TASE$_{IO}$+SSE setup that is a combination of their multi-span architecture and single-span extraction with IO tagging.We use the same configuration and hyper-parameters for TASE$_{IO}$+SSE as described in Segal et al. (2020). We train all models for two epochs in all experiments.[11] We use the F1 score that calculates the number of shared words between predictions and gold answers for evaluation.

Training Strategies. To include the additional training data that we create from CoNLL-2012 using coreference-to-CoNLL conversion methods, we use two different strategies:

- *Joint*: we concatenate the training examples from Quoref and CoNLL-to-QA converted datasets. Therefore, the model is jointly trained on the examples from both datasets.
- *Transfer*: Since the CoNLL-to-QA data is automatically created and is noisy, we also examine a sequential fine-tuning setting in which we first train the model on the CoNLL-to-QA converted data, and then fine-tune it on Quoref.

5.3 Data

We evaluate all the models on four different QA datasets.

- *Quoref*: the official development and test sets of Quoref, i.e., Quoref$_{dev}$ and Quoref$_{test}$, respectively.
- *Our challenge set*: our new evaluation set described in Section 4.
- *Contrast set*: the evaluation set by Gardner et al. (2020) that is created based on the official Quoref test set. For creating this evaluation set, the authors manually performed small but meaningful perturbations to the test examples in a way that it changes the gold label. This dataset is constructed to evaluate whether models decision boundaries align to true decision boundaries when they are measured around the same point.
- *MultiRC*: Multi-Sentence Reading Comprehension set (Khashabi et al., 2018) is created in a way that answering questions requires a more complex understanding from multiple sentences. Therefore, coreference reasoning can be one of the sources for improving the performance on this dataset. Note that *MultiRC* is from a different domain than the rest of evaluation sets.[12]

[11] The only difference of TASE in our experiments and the reported results in Segal et al. (2020) is the number of training epochs. For a fair comparison, we train all models for the same number of iterations.

[12] To use the MultiRC development set, which is in a multi-choice answer selection format, we convert it to a reading comprehension format by removing QA pairs whose answers cannot be extracted from the context.

Model	Training setup	Quoref$_{dev}$	Quoref$_{test}$	Ours	Contrast set	MultiRC
TASE	Baseline	84.05	84.71	66.48	73.44	51.83
	CoNLL$_{bart}$	34.95	35.76	39.55	26.24	26.51
	Joint-CoNLL$_{dec}$	84.36	85.14	65.92	74.88	44.71
	Transfer-CoNLL$_{dec}$	85.00	85.88	**73.07**	75.69	50.18
	Joint-CoNLL$_{bart}$	84.30	85.93	69.37	74.00	48.26
	Transfer-CoNLL$_{bart}$	**85.13**	85.98	73.01	77.40	51.96
	Transfer-SQuAD	84.70	**87.02**	67.99	**78.28**	**53.51**
RoBERTa	Baseline	79.64	79.69	64.35	**69.95**	37.12
	CoNLL$_{bart}$	28.82	29.10	29.00	17.36	14.81
	Joint-CoNLL$_{dec}$	75.15	74.83	56.94	57.78	29.97
	Transfer-CoNLL$_{dec}$	74.10	73.65	60.09	58.95	30.93
	Joint-CoNLL$_{bart}$	78.70	79.59	**67.07**	66.78	35.43
	Transfer-CoNLL$_{bart}$	78.22	78.33	66.62	66.58	36.84
	Transfer-SQuAD	**80.18**	**79.82**	64.88	69.46	**38.26**

Table 5: Impact of incorporating coreference data in MRC using CoNLL$_{dec}$ and CoNLL$_{bart}$ conversion methods on RoBERTa-large and TASE models. The *Baseline* and *CoNLL$_{bart}$* rows show the results when models are trained on the Quoref training data and the CoNLL$_{bart}$ data, respectively. *Joint* refers to the setting in which the model is jointly trained on Quoref and the converted CoNLL data. *Transfer* refers to the setting in which the model is first trained on the converted CoNLL data and fine-tuned on Quoref. *Transfer-SQuAD* shows the impact of training the model on additional QA data from a similar domain. Results are reported based on F1 scores. The highest F1 scores for each model are boldfaced and scores lower than the *Baseline* are marked in gray.

The *Contrast set* and *MultiRC* datasets are not designed to explicitly evaluate coreference reasoning. However, we include them among our evaluation sets to have a broader view about the impact of using our coreference data in QA.

Training	examples	Test	examples
Quoref train	19399	Quoref dev	2418
CoNLL$_{dec}$	89403	Ours	200
CoNLL$_{bart}$	18906	Contrast set	700
SQuAD	86588	MultiRC	389

Table 6: Number of examples in each dataset.

Table 6 reports the statistics of these QA datasets. In addition, it reports the number of examples in CoNLL$_{dec}$ and CoNLL$_{bart}$ datasets that we create by converting the CoNLL-2012 training data into QA examples. Since the question generation model cannot generate a standard question for every declarative sentence, CoNLL$_{bart}$ contains a smaller number of examples. We also include the statistics of SQuAD in Table 6 as we use it for investigating whether the resulting performance changes are due to using more training data or using coreference-aware additional data.

The language of all the datasets is English.

5.4 Results

Table 5 presents the results of evaluating the impact of using coreference annotations to improve coreference reasoning in MRC. We report the results for both of the examined state-of-the-art models, i.e., TASE and RoBERTa-large, using both training settings: (1) training the model jointly on Quoref and CoNLL-to-QA converted data (Joint), and (2) pre-training the model on CoNLL-to-QA data first and fine-tuning it on Quoref (Transfer). *Baseline* represents the results of the examined models that are only trained on Quoref. *CoNLL$_{bart}$* represents the results of the models that are only trained on the CoNLL$_{bart}$ data. *Transfer-SQuAD* reports the results of the sequential training when the model is first trained on the SQuAD training dataset (Rajpurkar et al., 2016) and is then fine-tuned on Quoref.

Based on the results of Table 5, we make the following observations.

First, the most successful setting for improving coreference reasoning, i.e., improving the performance on our challenge evaluation set, is Transfer-CoNLL$_{bart}$. Pre-training the TASE model on *CoNLL$_{bart}$* improves its performance on all of the examined evaluation sets. However, it only improves the performance of RoBERTa on our challenge set.

Second, SQuAD contains well-formed QA pairs while CoNLL$_{bart}$ and CoNLL$_{dec}$ contain noisy QA. Also, SQuAD and Quoref are both created based on Wikipedia articles, and therefore, have similar domains. However, the genres of the documents in CoNLL-2012 include newswire, broadcast news, broadcast conversations, telephone conversations,

Model	Semantic overlap		¬Semantic overlap		Short reasoning		¬Short reasoning	
	dev	Ours	dev	Ours	dev	Ours	dev	Ours
TASE Baseline	81.69	77.2	84.86	62.96	94.84	89.04	72.95	54.15
Joint-CoNLL$_{dec}$	+2.07	-5.80	-0.30	+1.19	+0.94	-1.65	-0.34	+0.03
Joint-CoNLL$_{bart}$	+0.86	-3.00	+0.03	+4.82	+0.64	+1.20	-0.16	+3.80
Transfer-CoNLL$_{dec}$	+1.29	+8.56	+0.83	+5.94	+1.07	+8.82	+0.83	+5.37
Transfer-CoNLL$_{bart}$	+1.74	+1.23	+0.85	+8.26	+1.54	+4.70	+0.60	+7.51
Transfer-SQuAD	+0.84	+0.58	+1.19	+0.10	-0.91	+2.3	+0.33	+2.15
RoBERTa baseline	78.09	67.39	80.19	63.36	90.04	84.23	68.97	53.48
Joint-CoNLL$_{dec}$	-5.55	-10.04	-4.15	-6.55	-2.53	-7.32	-6.54	-7.46
Joint-CoNLL$_{bart}$	0.00	+1.94	-1.28	+2.97	-0.48	-1.36	-1.43	+4.95
Transfer-CoNLL$_{dec}$	-4.20	+1.79	-6.02	-6.27	-3.52	-7.00	-7.65	-2.77
Transfer-CoNLL$_{bart}$	-1.02	-0.55	-1.58	+3.18	-0.95	-5.06	-1.94	+6.27
Transfer-SQuAD	+1.32	-1.08	+0.25	+1.05	+0.45	-9.46	+0.6	+5.99

Table 7: F1 score differences of various TASE and RoBERTa models on the Quoref$_{dev}$ and our dataset splits that are created based on the semantic overlap and short distance reasoning biases. For instance, *Ours* in the ¬*Semantic overlap* column shows the performance differences of the examined models on the split of our dataset in which examples do not contain the semantic overlap bias. Negative differences are marked in gray.

weblogs, magazines, and Bible, which are very different from those in Quoref. As a result, pretraining on SQuAD has a positive impact on the majority of datasets. However, this impact is less pronounced on our challenge dataset, as it requires coreference reasoning while this skill is not present in SQuAD examples.

Finally, while using the sentence of coreferring mentions as a declarative query (CONLL$_{dec}$) is the common method for converting coreference resolution datasets into QA format in previous studies, our results show using CoNLL$_{bart}$ has a more positive impact compared to using CoNLL$_{dec}$.

5.5 Analysis

To analyze what kind of examples benefit more from incorporating the coreference data, we split Quoref$_{dev}$ and our dataset into different subsets based on the *semantic overlap* and *short distance reasoning* biases, which are the most common types of biases in both datasets.

The *semantic overlap* column in Table 7 represents the results on the subset of the data in which answers reside in the most similar sentence of the context, and the ¬*semantic overlap* column contains the rest of the examples in each of the examined datasets. The *short reasoning* column presents the results on the subset of the data containing examples that can be solved by the short distance reasoning bias model, and ¬ *short reasoning* presents the results on the rest of the examples.

Table 7 shows the performance differences of the TASE and RoBERTa models on these four subsets for each of the two datasets.

Surprisingly, the performance of the baseline models is lower on the *semantic overlap* subset compared to ¬*semantic overlap* on Quoref$_{dev}$. This can indicate that examples in the ¬*semantic overlap* subset of Quoref$_{dev}$ contain other types of biases that make QA less challenging on this subset.

The addition of the coreference resolution annotations in all four training settings reduces the performance gap of the TASE model on the *semantic overlap* and ¬*semantic overlap* subsets for both datasets. Incorporating coreference data for RoBERTa, on the other hand, has a positive impact using the CoNLL$_{bart}$ data and on the harder subsets of our challenge evaluation set, i.e., ¬*semantic overlap* and ¬*short reasoning*.

Finally, there is still a large performance gap between *short reasoning* and ¬ *short reasoning* subsets. In our coreference-to-QA conversion methods, we consider the closest antecedent of each anaphor as the answer. A promising direction for future work is to also create QA pairs based on longer distance coreferring expressions, e.g., to create two QA pairs based on each anaphor, one in which the answer is the closest antecedent, and the other with the first mention of the entity in the text as the answer.

6 Conclusions

We show that the high performance of recent models on the Quoref dataset does not necessarily indicate that they are adept at performing coreference reasoning, and that QA based on coreference reasoning is a greater challenge than current scores

suggest. We then propose a methodology for creating a dataset that better presents the coreference reasoning challenge for MRC. We provide our methodology to an annotator and create a sample dataset. Our analysis shows that our dataset contains fewer biases compared to Quoref, and the performance of state-of-the-art Quoref models drops considerably on this evaluation set.

To improve the coreference reasoning of QA models, we propose to use coreference resolution datasets to train MRC models. We propose a method to convert coreference annotations into an MRC format. We examine the impact of incorporating this coreference data on improving the coreference reasoning of QA models using two top-performing QA systems from the Quoref leaderboard. We show that using coreference datasets improves the performance of both examined models on our evaluation set, indicating their improved coreference reasoning. The results on our evaluation set suggest that there is still room for improvement, and reading comprehension with coreference understanding remains a challenge for existing QA models, especially if the coreference relation is between two distant expressions.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) as part of the QASciInf project (grant GU 798/18-3), and the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Dan Roth's work is partly supported by contract FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA). The authors would like to thank Michael Bugert, Max Glockner, Yevgeniy Puzikov, Nils Reimers, Andreas Rücklé, and the anonymous reviewers for their valuable feedback.

References

Rahul Aralikatte, Matthew Lamm, Daniel Hardt, and Anders Søgaard. 2019. A Simple Transfer Learning Baseline for Ellipsis Resolution. *arXiv preprint arXiv:1908.11141*.

Devansh Arpit, Stanisław Jastrzundefinedbski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 233–242. JMLR.org.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019a. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4067–4080, Hong Kong, China. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019b. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932, Hong Kong, China. Association for Computational Linguistics.

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.

Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of intermediate annotations in reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5627–5634, Online. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*, pages 1078–1088. PMLR.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012a. 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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012b. 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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080, Online. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*. Association for the Advancement of Artificial Intelligence.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020a. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8717–8729, Online. Association for Computational Linguistics.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020b. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online. Association for Computational Linguistics.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mingzhu Wu, Nafise Sadat Moosavi, Andreas Rücklé, and Iryna Gurevych. 2020a. Improving QA generalization by concurrent modeling of multiple biases. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 839–853, Online. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020b. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Mastering the dungeon: Grounded language learning by mechanical turker descent. In *Proceedings of 6th International Conference on Learning Representations (ICLR)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Additional Statistics about Biased Examples

Table 8 shows the proportion of biased examples in the CoNLL$_{dec}$ set. We can see that the results are similar to that of the CoNLL$_{bart}$ set.

To compare the ratio of biased examples between Quoref$_{dev}$ and our challenge set when considering the same number of examples in both datasets, we randomly sample 10 different subsets from Quoref$_{dev}$ and our challenge set with 100 samples in each subset and compute the rations in each subset. Figure 3 shows the results. As we see, in this setting the ratio of all bias types in our evaluation set is still lower than those in Quoref$_{dev}$.

B Additional Experiments

Table 9 shows additional experiments for pre-training the examined models on coreference data. We examine an additional setting for pre-training on both CoNLL$_{dec}$ and CoNLL$_{bart}$ by first training the models on CoNLL$_{dec}$, then on CoNLL$_{bart}$, and finally on Quoref (*Transfer-CoNLL*$_{bart+dec}$). By comparing the results of *Transfer-CoNLL*$_{bart+dec}$ with *Transfer-CoNLL*$_{bart}$ from Table 5, we observe that pre-training the models on both CoNLL$_{dec}$ and CoNLL$_{bart}$ does not result in any additional advantage compared to only using CoNLL$_{bart}$.

C Additional Examples

Table 10 presents more examples from CoNLL$_{dec}$ and CoNLL$_{bart}$.

Bias	**CoNLL**$_{dec}$
random named entity	2.11
wh-word	12.97
empty question	11.24
semantic overlap	35.32
short-distance reasoning	21.05

Table 8: Proportion of biased examples in CoNLL$_{dec}$ dataset.

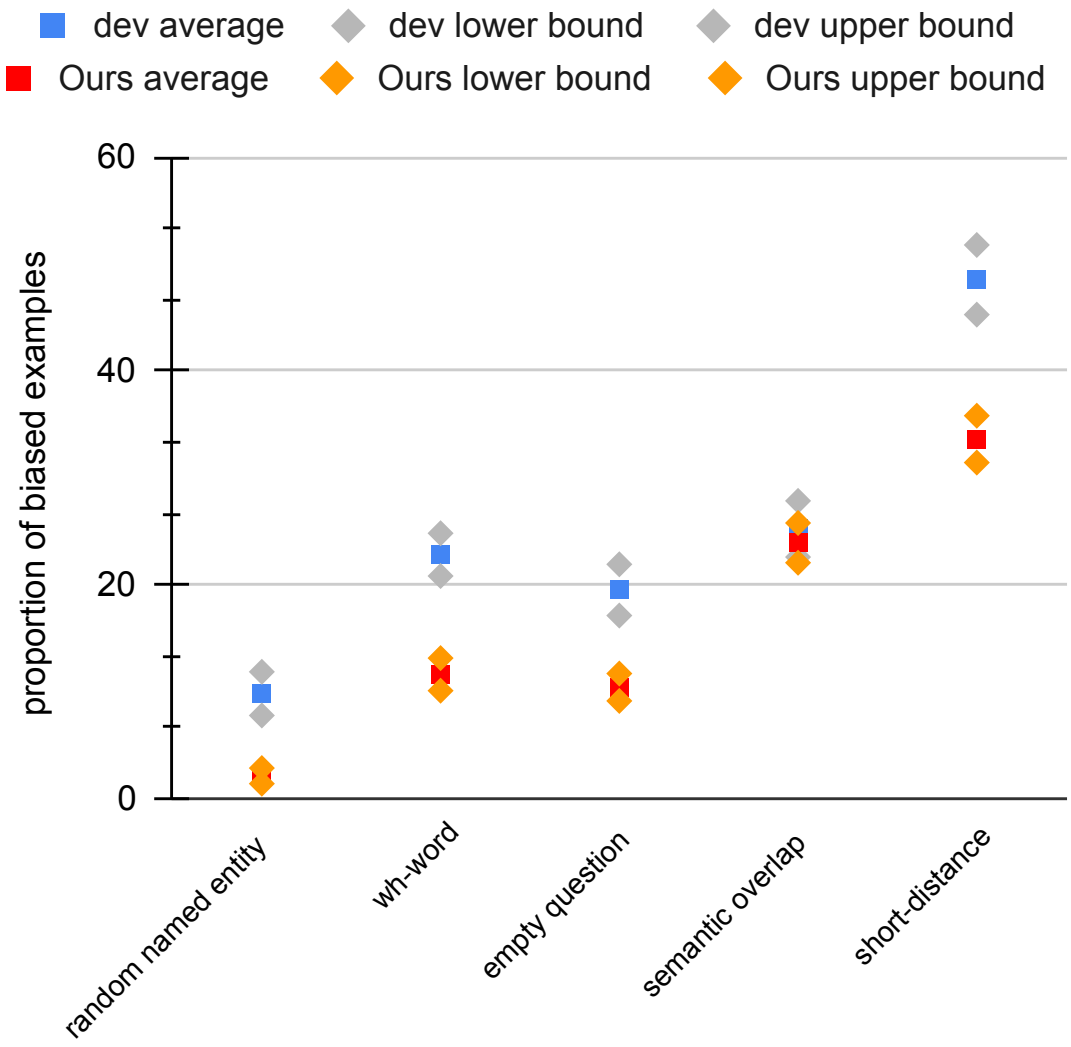

Figure 3: The average, upper and lower bounds of the ratio of biased examples in Quoref$_{dev}$ and our challenge set for the randomly sampled 10 subsets.

Model	Training setup	Quoref$_{dev}$	Quoref$_{test}$	Ours	Contrast set	MultiRC
TASE	Baseline	84.05	84.71	66.48	73.44	51.83
	Transfer-CoNLL$_{bart+dec}$	85.01	85.73	68.06	76.54	49.61
RoBERTa	Baseline	79.64	79.69	64.35	69.95	37.12
	Transfer-CoNLL$_{bart+dec}$	73.29	73.73	58.19	57.18	31.50

Table 9: Additional experiments on using the CoNLL$_{dec}$ and CoNLL$_{bart}$ data for pre-training RoBERTa-large and TASE models. *Transfer-CoNLL$_{bart+dec}$* refers to the setting in which the model is first trained on CoNLL$_{dec}$, then on CoNLL$_{bart}$, and finally on Quoref.

Passage in CoNLL	Mention Cluster	CoNLL$_{dec}$ Quesion	CoNLL$_{bart}$ Question	Gold Answer
After George W. Bush is sworn in, Bill Clinton will head to New York. Mr. Clinton will also spend time at his presidential library in Arkansas. He says he will come to Washington, 'every now and then'.	[Bill Clinton, Mr. Clinton, his, He, he]	He says <ref> he </ref> will come to Washington, 'every now and then.'	who says he will come to washington, 'every now and then'?	Bill Clinton
Paul had already decided not to stop at Ephesus. He did not want to stay too long in Asia. He was hurrying because he wanted to be in Jerusalem on the day of Pentecost if possible.	[Paul, He, He, he]	He was hurrying because <ref> he </ref> wanted to be in Jerusalem on the day of Pentecost if possible.	who was hurrying because they wanted to be in jerusalem on the day of pentecost if possible?	Paul
The KMT vice chairman arrived at party headquarters to meet with KMT Chairman Lien Chan on the afternoon of pw...He said that he will follow Lien Chan as a lifelong volunteer.	[The KMT vice chairman, He, he]	He said that <ref> he </ref> will follow Lien Chan as a lifelong volunteer.	who said that he will follow lien chan as a lifelong volunteer?	The KMT vice chairman
...It also includes a lot of sheep, good clean - living, healthy sheep, and an Italian entrepreneur has an idea about how to make a little money of them...So this guy came up with the idea of having people adopting sheep by an internet.	[an Italian entrepreneur, this guy]	So <ref> this guy </ref> came up with the idea of having people adopting sheep by an internet.	who came up with the idea of having people adopting sheep by an internet?	an Italian entrepreneur
George W. Bush has met with Al Gore in Washington. The two men met for just 15 minutes at the Vice President's official residence...Bush went into the talks with his defeated rival after meeting with President Clinton earlier today.	[Al Gore, his defeated rival]	Bush went into the talks with <ref> his defeated rival </ref> after meeting with President Clinton earlier today.	who did bush go into the talks with after meeting with president clinton earlier today?	Al Gore
Meanwhile Prime Minister Ehud Barak told Israeli television he doubts a peace deal can be reached before Israel's February 6th election. He said he will now focus on suppressing Palestinian violence.	[Prime Minister Ehud Barak, He, he]	He said <ref> he </ref> will now focus on suppressing Palestinian violence.	who said he will now focus on suppressing palestinian violence?	Prime Minister Ehud Barak

Table 10: More examples from coreference-to-QA conversion using CoNLL$_{dec}$ and CoNLL$_{bart}$ approaches.