Evaluating Coreference Resolvers on Community-based Question Answering: From Rule-based to State of the Art

Haixia Chai[†] **Nafise Sadat Moosavi**[Φ Ψ] **Iryna Gurevych**[Φ] **Michael Strube**[†]
[†] Heidelberg Institute for Theoretical Studies
[Φ] UKP Lab, Technische Universität Darmstadt
[Ψ] Department of Computer Science, The University of Sheffield
{haixia.chai, michael.strube}@h-its.org

Abstract

Coreference resolution is a key step in natural language understanding. Developments in coreference resolution are mainly focused on improving the performance on standard datasets annotated for coreference resolution. However, coreference resolution is an intermediate step for text understanding and it is not clear how these improvements translate into downstream task performance. In this paper, we perform a thorough investigation on the impact of coreference resolvers in multiple settings of a community-based question answering task, i.e., answer selection with long answers. Our settings cover multiple text domains and encompass several answer selection methods. We first inspect extrinsic evaluation of coreference resolvers on answer selection by using coreference relations to decontextualize individual sentences of candidate answers, and then annotate a subset of answers with coreference information for intrinsic evaluation. The results of our extrinsic evaluation show that while there is a significant difference between the performance of the rule-based system vs. state-of-the-art neural model on coreference resolution datasets, we do not observe a considerable difference on their impact on downstream models. Our intrinsic evaluation shows that (i) resolving coreference relations on less-formal text genres is more difficult even for trained annotators, and (ii) the values of linguistic-agnostic coreference evaluation metrics do not correlate with the impact on downstream data.[1]

1 Introduction

Coreference resolution is the task of determining the expressions of the text that refer to the same entity. Modeling coreference relations is a key step for understanding the meaning of the text that can benefit various tasks like machine reading comprehension (Huang et al., 2022), summarization (Huang and Kurohashi, 2021), and dialogue processing (Xu and Choi, 2022).

The progress in coreference resolution is tailored to improve the performance on available coreference resolution datasets (Lee et al., 2017, 2018; Joshi et al., 2019, 2020; Kirstain et al., 2021; Chai and Strube, 2022), but it is not clear how this progress translates to downstream applications.

In this paper, we take a new perspective to directly evaluate the impact of coreference resolvers on a downstream task. First, we implement the extrinsic evaluation of coreference resolvers on the task of community-based question answering (CQA), in which the task is to select the correct answer given a question and a set of candidate answers. Answers in CQA are often very long, and they contain multiple referring expressions in each answer. To do so, we use existing coreference resolvers for decontextualizing candidate answers — i.e., replacing less informative nouns and pronouns with their most informative antecedent — so that the containing information in each individual sentence would be more standalone. To ensure that the resulting effects are not specific to a single dataset, domain, or downstream model, our settings cover multiple text domains and encompass several CQA methods. Second, we provide coreference annotations on a subset of answers from two CQA domains to enable intrinsic evaluation of coreference resolvers on a downstream data.

We evaluate several coreference resolvers from the rule-based system (Lee et al., 2013) to the state-of-the-art coreference resolver (Joshi et al., 2020) using our extrinsic and intrinsic evaluation setups.

The results of our extrinsic evaluation show that (i) rule-based system has a more positive and less negative impact on CQA compared to neural coreference resolvers, (ii) while there is a significant difference between the performance of the rule-based system vs. state-of-the-art neural model on coreference resolution datasets, we do not observe

[1] Our code and coreference annotations on CQA datasets are publicly available at: https://github.com/HaixiaChai/Coref_CQA

Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2022) , pages 61–73
Gyeongju, Republic of Korea, October 16–17, 2022.

a considerable difference on their impact on CQA models. This means that intrinsic evaluation has to be accompanied by extrinsic evaluation, (iii) the impact of coreference resolution is different on various CQA methods. Thus, we suggest to consider the overall impact on multiple CQA models in order to investigate the effect of a coreference resolver on CQA, and (iv) coreference resolvers are most beneficial when both training and test data are decontextualized, and the rule-based system has consistent impact on different domains of the data while the state-of-the-art neural models have a considerable different impact on different domains.

Our extrinsic evaluation results show that (i) resolving coreference relations on less-formal text genres — like ones in the Stack Exchange answers — is more difficult even for trained annotators, and (ii) the results of linguistic-agnostic coreference evaluation metrics do not correlate with the impact of coreference resolvers on downstream data.

2 Coreference for Answer Selection

Given a question, the task of answer selection is to find the correct answer among the set of candidate answers. We use answer selection datasets from community question answering (CQA). CQA questions are non-factoid and they often require answers with descriptions or explanations. Therefore, CQA answers are long multi-sentence texts.

With the length of answers, the use of less informative expressions like pronouns increases. This presents a challenge for answer selection methods that mainly rely on the lexical forms to compute the similarity of the candidate answers to the question. Especially, when CQA data is collected by using a search engine or the answers to the similar questions for candidates, incorrect answers also have high lexical similarity with questions.

Using coreference resolvers for decontextualizing individual sentences in answer selection datasets makes correct answers more similar to the question and incorrect ones more dissimilar. Table 1 shows a sample question and two candidate answers, in which mentions that refer to the same entity are specified by the same index in each of the answers. **A1** and **A2** address two different issues, i.e., the need for a visa from Ireland to UK vs. getting an Irish visa given that your UK visa has been rejected. Both candidate answers contain a similar text sequence that is relevant to the question, i.e., “need to acquire a visa to enter the country” in **A1** and “get the visa to enter the country” in **A2**. These two text sequences can be easily discriminated given coreference information, i.e., “need to acquire a visa to enter UK” in **A1** and “get your Irish visa to enter Ireland” in **A2**.

Q: Do I need a UK visa to enter UK from Ireland?
A1: What is your nationality? According to the [UK]$_1$ government service information website (URL), people from the countries who are mentioned in URL would still need to acquire a visa to enter [the country]$_1$.
A2: Data sharing means only that they share data, so while [the officers in [Ireland]$_6$]$_3$ are able to see details of [your]$_4$ failed UK visa when [they]$_3$ process [[your]$_4$ Irish visa]$_5$, that doesn't mean [you]$_4$ will be refused to get [the visa]$_5$ to enter [the country]$_6$.

Table 1: An example of a question and a correct (**A1**) and an incorrect (**A2**) candidate answer.

3 Extrinsic Evaluation on CQA

The following sections describe different components for the extrinsic evaluation of coreference resolvers using CQA. Figure 1 shows the flow chart.

3.1 Answer Selection Models

Sentence-BERT. We use Sentence-BERT (Reimers and Gurevych, 2019) as an unsupervised baseline for answer selection. Here, we use the pre-trained model, MPNet (Song et al., 2020), to compute sentence embeddings.[2] By computing the sentence embedding of each candidate answer and that of the question, we select the candidate answer with the highest cosine similarity to the question.[3]

CNN. We train a CNN network for computing the semantic representation of candidate answers and questions. Similar to Tan et al. (2016) and Rücklé et al. (2019), we use a max-pooling layer on top of a CNN to get fix-sized representations. The similarity of the candidate answer and question representations is computed by cosine similarity.

Attentive LSTM. Instead of computing independent representations for questions and candidate answers, Tan et al. (2016) propose to use the attentive LSTM model in which the representation of answers is computed based on the question representation.

[2] MPNet shows the best performance at `https://www.sbert.net/docs/pretrained_models.html`.

[3] This approach is the state of the art on the datasets (Rücklé et al., 2019) on which we study the extrinsic evaluation of coreference resolution systems.

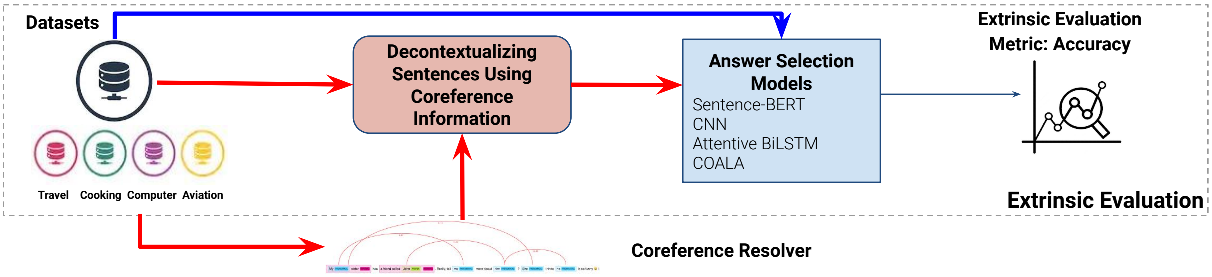

Figure 1: The figure shows our extrinsic evaluation of coreference resolvers on CQA. The red line indicates decontextualizing sentences using coreference information in the data, while the blue line shows the original data.

COALA. The COALA model (Rücklé et al., 2019) first uses a CNN to compute a representation for the local aspect (bi-grams) of both the question and each candidate answer. It selects the candidate answer that covers more aspects of the question.[4]

3.2 Datasets

The datasets (Rücklé et al., 2019) are in English from a diverse set of domains from StackExchange including Travel (Q&A for travelers), Cooking (Q&A for professional and amateur chefs), Computer (Q&A for the users of Apple hardware and software), and Aviation (Q&A for aircraft pilots, mechanics, etc.) communities. Table 2 provides the statistics of these datasets.

Dataset	Number of Questions			Answer
	Train	Valid	Test	Length
Travel	3 572	765	766	214
Cooking	3 692	791	792	189
Computer	5 831	1 249	1 250	114
Aviation	3 035	650	652	281

Table 2: The statistics of the answer selection datasets.

These datasets contain predefined train, validation, and test splits. We train each of the supervised answer selection models on the training split of each of the datasets.[5] For instance, we have four different CNN trained models for each of the datasets. Models that are trained on the travel training set are used for evaluating the effect of coreference resolution on the travel test set.

Note that existing supervised coreference resolvers are trained on the CoNLL-2012 data (Pradhan et al., 2012) that contains different domains including newswire, broadcast news, broadcast conversations, telephone conversations, weblogs, magazines, and Bible domains.

3.3 Incorporating Coreference Relations

To benefit from coreference information in downstream tasks, one can either incorporate coreference relations in the model, e.g., (Dhingra et al., 2018; Du and Cardie, 2018; De Cao et al., 2019; Dua et al., 2020), or in its input data, e.g., (Steinberger et al., 2007; Du and Cardie, 2018), from which we use the second approach. The approach is similar to decontextualization (Choi et al., 2021), in which the goal is to make the meaning of individual sentences standalone in an empty context. Coreference resolution is one of the main steps for decontextualization, and as shown by Choi et al. (2021), it is a valuable preprocessing step for tasks that require document understanding.[6] In addition, using coreference resolvers for decontextualizing input sentences has the following benefits: (1) a single coreference annotated dataset can be used for evaluating various answer selection models, and (2) it does not require developing specialized coreference-aware models for the application.

We first apply the coreference resolver on all candidate answers and get the resulting coreference chains. Then for each mention in the coreference chains, we determine the most representative antecedent[7] using the rules proposed by Lee et al. (2013): if two mentions are of different types, proper names are the most representative mentions and common nouns are more representative than pronouns, e.g., “the UK visa” vs. “it”. Otherwise, the mention containing more words is more representative, “the UK visa” vs. “the visa”.

In our experiments, we examine and report two different settings: (1) **coreference resolution**: replacing all types of referring expressions with their

[4] It has two variants, from which we select the one with higher scores, i.e., COALA p-means.

[5] We use same hyper-parameters as Rücklé et al. (2019).

[6] Note that the full decontextualization of sentences requires more than coreference resolution — e.g., bridging resolution. We aim to evaluate coreference resolvers, so we focus on using coreference resolution for decontextualization.

[7] All coreferring mentions that appear before the current mention are considered as antecedents.

most representative antecedent, and (2) **pronoun resolution**: only replacing pronouns with their most informative antecedent. Meanwhile, we incorporate coreference annotations in two different ways: (1) **only in the test data**: models trained on original training data are evaluated on different coreference annotations on the test data, and (2) **both in the training and test sets**: we train and test the supervised CQA models on the training and test sets that are decontextualized by using coreference relations.

3.4 Extrinsic Evaluation Metric

We use accuracy — i.e., the ratio of correctly selected answers — to measure the performance of answer selection models. The impact of each coreference resolver on answer selection is measured by computing the difference between the accuracy of answer selection models on the coreference annotated test sets vs. the original ones. Table 3 reports the performance of CQA models.

Model	**Dataset**			
	Travel	Cooking	Computer	Aviation
Sentence-BERT	81.98	77.65	64.32	80.06
CNN	34.46	26.01	20.24	26.22
Att.-BiLSTM	43.34	38.38	25.60	36.34
COALA	54.83	47.34	33.52	52.45

Table 3: Accuracy of answer selection models.

4 Intrinsic Evaluation on CQA data

We enable intrinsic evaluation of coreference resolvers on CQA data by annotating coreference relations on a subset of the CQA data.

We annotate a subset of examples from the Travel and Cooking test sets. We use *MMAX2* (Müller and Strube, 2006) for the annotations.[8] The annotations are done by six bachelor and master students with NLP background from the Departments of Computational Linguistics and Computer Science. They received a minimal training for coreference resolution and the *MMAX2* annotation tool. Table 4 presents the statistics of this annotated data. We annotate a subset of 100 answers by two of the annotators and perform an inter-annotator agreement study. The inter-annotator agreement is 0.71 using Krippendorff's α (Krippendorff, 1980) with MASI distance metric (Passonneau, 2006).[9]

	Travel	Cooking
answers	389	558
max words/answer	319	283
coreference chains/answer	4.2	3.4
mentions/answer	14.0	12.3

Table 4: Statistics of our human anotations based on the number of annotated answers, maximum number of words per answer, average number of coreference chains per answer, and average number of annotated mentions per answer in each of the domains.

While our agreement study shows a high inter-annotator agreement, we also perform a manual error analysis on the resulting annotations.[10] Based on our analysis, annotating coreference relations in less-formal genres is more difficult than in the common genres in existing NLP datasets, e.g., news, and their error-free annotations would require expert linguists.[11] In particular, human annotations in Travel contain more errors. This indicates that resolving coreference relations of the answers in the Travel domain, which contains more nominal expressions, is more difficult than Cooking.

5 Examined Coreference Resolvers

We evaluate four different coreference resolvers.

First, the Stanford **rule-based** system (Lee et al., 2013) that uses heuristic rules like string match for resolving coreference relations. There is a considerable gap between its performance and state-of-the-art coreference resolvers on the CoNLL-2012 test set. However, it has a reasonable performance across different domains (Moosavi and Strube, 2017).

Second, **deep-coref** (Clark and Manning, 2016), which is a neural coreference resolver. *deep-coref* is a neural model that first extracts candidate mentions using syntactic information. For each candidate mention, it scores all preceding mentions to select the best scoring one as the antecedent. It also includes a dummy antecedent to determine non-anaphoric mentions, i.e., if the dummy antecedent has the highest score, the mention is non-anaphoric.

Third, **e2e-coref** (Lee et al., 2018) that is an end-to-end neural coreference resolver and the base model for the majority of state-of-the-art coreference resolvers since 2018. Unlike *deep-coref* and

[8] http://mmax2.net/

[9] Details are included in Appendix A.

[10] For the error analysis of the human annotations, we refer to Appendix A.

[11] This is consistent with the previous observation of Chai et al. (2020) that resolving coreference relations in noisy user-generated texts is very challenging.

Coreference	Answer Selection	Travel	Cooking	Computer	Aviation
rule-based	Sentence-BERT	-1.57	-1.39	-0.96	-1.54
	CNN	**1.17**	**0.50**	**0.80**	0.00
	Att.-BiLSTM	***1.17***	***0.63***	***0.88***	***0.92***
	COALA	**0.78**	**0.13**	**1.44**	**0.46**
deep-coref	Sentence-BERT	-0.65	-0.63	-0.48	-1.54
	CNN	***0.52***	***0.75***	***0.40***	0.00
	Att.-BiLSTM	-0.13	**0.63**	**0.16**	**0.92**
	COALA	-0.40	**0.38**	**0.96**	**0.46**
e2e-coref	Sentence-BERT	**0.26**	-1.14	-0.48	-1.23
	CNN	***1.04***	***0.75***	***0.24***	***0.16***
	Att.-BiLSTM	-0.26	**0.50**	-0.24	-0.61
	COALA	**0.52**	-0.12	**0.24**	0.00
bert-coref	Sentence-BERT	-0.13	-1.01	0.00	-1.38
	CNN	**0.78**	-0.38	-0.40	**0.31**
	Att.-BiLSTM	**0.39**	**0.25**	**0.32**	**0.31**
	COALA	***0.39***	0.00	***0.56***	***0.61***

Table 5: Effect of the coreference resolvers on different answer selection models and datasets. Cell values indicate the difference between the accuracy when incorporating coreference annotations on test sets vs. the baseline results. The bold-faced values mean that the coreference resolver has a positive impact on the corresponding CQA models and domains. The values in italic and underline show the answer selection models on which each coreference system has the best impact.

rule-based systems, *e2e-coref* does not use syntactic information or a separate modules to determine candidate mentions. It jointly determines mention spans as well as their corresponding coreference relations by an end-to-end neural model.

Last, **bert-coref** (Joshi et al., 2020) that is one of the most recent state-of-the-art coreference resolvers on the CoNLL-2012 dataset. *bert-coref* is an extension of *e2e-coref* by replacing the bidirectional LSTM encoder with SpanBERT encodings. Concretely, we use the SpanBERT-large language model, which has a novel span masking pretraining objective that predicts the entire masked span instead of individual tokens.

For the reported extrinsic and intrinsic evaluations, the supervised coreference resolvers are trained on the English CoNLL-2012 dataset. Table 6 presents the scores of these coreference resolvers on the CoNLL-2012 test set based on the standard coreference evaluation metrics, i.e., MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF$_e$ (Luo, 2005), and LEA (Moosavi and Strube, 2016).

Metric	rule-based	deep-coref	e2e-coref	bert-coref
MUC	64.7	74.2	80.4	**85.3**
B^3	52.7	63.0	70.8	**78.1**
CEAF$_e$	49.3	58.7	67.6	**75.3**
LEA	47.3	59.5	67.7	**75.9**

Table 6: Performance of examined coreference resolvers on the English CoNLL-2012 test set based on coreference evaluation metrics.

6 Results

6.1 Extrinsic Evaluation

Evaluating CQA using coreference annotations in the test data. Table 5 shows the results of using the examined coreference resolvers on the CQA models and domains in the **coreference resolution** setting, i.e., replacing all referring expressions with their most representative antecedent.[12]

First, we observe that compared to state-of-the-art coreference resolvers, *rule-based* has a more positive impact and less negative impact on CQA.[13]

[12] Appendix C includes results of the **pronoun resolution**.

[13] We compute the statistical significance of *rule-based* and *bert-coref* by using Wilcoxon's Signed Rank test on all CQA models and two domains. For the travel domain the differences are statistically significant ($p \leq 0.05$), while in the cooking domain the results are not.

To further investigate this result, we report the total number of resolved mentions and pronouns by the examined resolvers across all CQA domains in Table 7. We observe that *rule-based* resolves the highest number of mentions (99k) and the lowest percentage of pronouns (64%), i.e., the ratio of pronouns in all resolved mentions, indicating that *rule-based* resolves more nominal mentions than the other coreference resolvers. Based on this observation, we hypothesize that resolving more nominal mentions and improving the precision of resolved pronouns will improve the effectiveness of state-of-the-art coreference resolvers on downstream applications.

Resolver	**Mentions**	**Pronouns**	**% of Pronouns**
rule-based	99k	63k	64%
deep-coref	70k	51k	73%
e2e-coref	72k	56k	77%
bert-coref	81k	57k	70%

Table 7: The statistics of total mentions and pronouns resolved by coreference resolvers on all domains.

Second, while there is a significant difference between performance of coreference resolvers on the CoNLL-2012 coreference dataset, e.g., ≈ 20 point difference between *rule-based* and *bert-coref* based on various coreference metrics in Table 6, we do not observe a considerable difference in their impact on CQA models. This suggests that intrinsic evaluation on CoNLL should be accompanied by extrinsic evaluation to approximate the utility of the coreference resolvers for end tasks.

Finally, we find that CNN and COALA that encode the text based on the local context have better performance with neural coreference resolvers, Attentive LSTM which encodes the context globally performs best with *rule-based*, and no coreference resolvers have a clear positive impact on Sentence-BERT[14] in Table 5. In general, the impact of coreference resolvers varies for different CQA models. So, we suggest to consider the overall impact on multiple CQA models to investigate the effect of a coreference resolver on CQA.

Table 10 shows an example of replaced coreference relations in a candidate answer.

Evaluating CQA using coreference annotations in both training and test data. For the above experiments, we only evaluate the impact of coreference resolvers by incorporating coreference information only on the test data. However, this may results in disparity between the data that models are trained on vs. testing data. We also investigate the impact of incorporating coreference relations on both training and test CQA data. Table 8 presents the results of this experiment for the *rule-based* and *bert-coref* systems and for the two representative domains, Travel and Cooking. For each of the experiments, we train and test the CQA models on the training and test data in which referring expressions are replaced with their most representative detected antecedent.

Resolver	**CQA**	**Travel**	**Cooking**
rule-based	CNN	-0.78	1.26
	Att.-BiLSTM	2.35	0.13
	COALA	0.91	0.63
bert-coref	CNN	2.22	0.63
	Att.-BiLSTM	2.09	-2.27
	COALA	0.13	-2.14

Table 8: Evaluating the impact of coreference resolution on supervised CQA models when the coreference information is used both in training and test sets.

Based on the results, incorporating coreference relations in both training and test datasets results in higher improvements compared to only incorporating them in the test data since the models see similar data formats during training and evaluation. From both challenging domains, we observe that *bert-coref* performs better on the Travel domain, while *rule-based* shows most positive results on both domains, even on Cooking that has shorter texts and contains more disfluent and ungrammatical expressions compared to the Travel domain. Thus, we encourage people to research more on diverse domains or genres beyond well-structured narrative texts.

6.2 Intrinsic Evaluation

Table 9 shows the evaluation of the examined coreference resolvers on our CQA coreference data described in §4 based on standard coreference resolution evaluation metrics as well as Application Related Coreference Scores (ARCS). ARCS is proposed by Tuggener (2014) for evaluating coreference resolvers based on their potential impact on downstream applications.

As mentioned in §5, all systems are trained on

[14] It is shown that pretrained models, like SentenceBERT, capture linguistic structures like anaphoric coreference to some extent (Manning et al., 2020), that may be the reason that using the incorporating the noisy output of coreference resolvers does not improve the performance of such systems.

the CoNLL-2012 training data, which contains different genres that those in our CQA data.

Metric	rule-based	deep-coref	e2e-coref	bert-coref
		Travel		
MUC	28.07	**55.36**	34.90	39.53
B^3	28.81	**50.66**	34.28	39.31
$CEAF_e$	33.56	**45.83**	38.95	44.62
LEA	23.19	**46.86**	30.19	35.29
ARCS	18.24	23.99	29.47	**36.80**
		Cooking		
MUC	31.58	**59.43**	37.82	43.07
B^3	30.99	**54.85**	36.17	40.70
$CEAF_e$	34.77	**52.42**	41.36	45.11
LEA	24.47	**50.01**	30.88	36.04
ARCS	15.49	24.37	26.27	**34.17**

Table 9: Intrinsic evaluation of examined coreference resolvers on our CQA coreference data.

As we see from the results, all standard coreference evaluation metrics — including MUC, B^3, $CEAF_e$, and LEA — agree on the ranking of the examined resolvers on both domains, based on which *deep-coref* performs better than the other systems.[15] ARCS, on the other hand, ranks *bert-coref* higher than the rest of the systems on both domains. Interestingly, none of the above rankings is consistent with our extrinsic evaluations in Table 5, e.g., the *rule-based* system receives the lowest ranking based on all metrics in intrinsic evaluations while its overall impact on CQA models is better than that of *bert-coref*.

Note that existing coreference resolution evaluation metrics are linguistic-agnostic, i.e., they do not discriminate the resolution of different types of mentions. This can be a potential reason that existing metrics do not correlate with the performance on a downstream task. For instance, as shown by Agarwal et al. (2019) resolving the corresponding proper name of each entity is more important than the resolution of other relations for certain downstream tasks.

7 Related Work

Task-oriented evaluation of coreference resolution. Tuggener classifies the use of coreference resolution in higher-level applications into three classes and proposes a different evaluation metric for each usecase:

- *Modeling entity distributions* to determine the exact sequence of each entity occurrence, which is useful in applications like modeling local coherence (Barzilay and Lapata, 2008). For such use-cases, Tuggener proposes to evaluate the detection of the immediate antecedent of each mention.

- *Inferring local entities* to determine the closest nominal antecedent of each mention. This use-case can be useful in applications like machine translation and summarization in which resolving pronouns with a nominal antecedent reduces ambiguity of the text. The proposed evaluation for this category is to only evaluate the closest preceding nominal antecedent of each mention.[16]

- *Finding context for a specific entity* to determine all references to the entity. This is useful for finding parts of the context that are related to a given question. Tuggener proposes to evaluate this setting by first finding the most representative mention of each coreference chain, called the anchor mention. He then computes the number of correct and incorrect references for each anchor mention in order to measure the performance.

Evaluation metrics of Tuggener (2014) are applicable on coreference annotated datasets. However, (1) existing coreference resolvers do not generalize well to new datasets and the performance in in-domain vs. out-of-domain settings may be completely different, and (2) as we saw in §6.2, they do not necessarily correlate with the impact on downstream applications.

Coreference for question answering. The use of coreference resolution in answer selection has been explored by various work, e.g., (Morton, 1999, 2000; Vicedo and Ferrández, 2000, 2008; Wang, 2010).[17] Morton (1999) proposes to rank candidate answers based on their coreference relations with the question, so that answers having more common entities with the question would get a higher rank. Stuckhardt (2003) and Wang et al. (2010) use anaphora resolution to detect common entities between the question and the candidate document for improving QA.

Morton (2000) evaluates the use of coreference resolution for QA. In order to compute the relevance of each sentence to the given question, he

[15] Based on our analysis, *deep-coref* resolves fewer informative mentions and more repeated pronouns compared to other systems.

[16] ARCS used in §6.2 refers to this metric.

[17] For the use of coreference resolution for other NLP applications refer to Stuckardt (2016).

Original text: Short answer, you can't. However, you can at least make sure they have an official license, and any other accreditation which might lend some credence to their claims. Look for **ones that are licensed by the <URL>**, and consider <**URL**>, to see if anyone has mentioned **them**$_1$ or complained about **them**$_2$. All you can do is research, and ask around when you get there as well. Or consider approaching **the companies** and ask **them**$_3$ directly-I 'm sure you'd not be the first, even if **it** is rather brazen;)
rule-based: {them$_1$, them$_2$} ← ones that are licensed by the <URL>; {them$_3$} ← the companies; {it} ← <URL>
deep-coref: NIL
e2e-coref: {them$_1$, them$_2$} ← ones that are licensed by the <URL>; {them$_3$} ← the companies
bert-coref: {them$_1$, them$_2$} ← ones that are licensed by the <URL>; {them$_3$} ← the companies
human-annot: {them$_1$, them$_2$} ← ones that are licensed by the <URL>, <URL>; {them$_3$} ← the companies

Table 10: An example from the replacements made by each of the examined coreference resolvers.

considers all the other mentions beyond the boundary of the sentence itself, that are coreferent with any of the sentence mentions. Vicedo and Ferrández (2000) evaluate the use of pronoun resolution in QA, and more specifically answer selection in QA. They show that incorporating information regarding the antecedent of pronouns improves, and in some cases is essential, for QA.

Aforementioned works, which show that coreference is beneficial for QA, use small-scale evaluations and simple QA models, e.g., TF-IDF, and coreference resolvers, e.g., rule-based systems. In this work, we investigate the use of coreference using recent answer selection models and coreference resolvers as well as multiple large-scale datasets.

Du and Cardie (2018) incorporate coreference information both at the input- and model-level for QA. At the input-level, they add the most informative antecedent of the pronouns to the input. At the model-level, they add coreference position feature embeddings to the model that specify the position of pronouns and their corresponding antecedents. They incorporate a gating mechanism to refine position embeddings based on the corresponding coreference score of each antecedent-pronoun relation.

These methods are costly to train, and therefore, they are not suitable to facilitate an efficient evaluation of various coreference resolvers on different CQA models and domains, e.g., their experiments are based on a single coreference output.

Quoref (Dasigi et al., 2019) is a question answering dataset that is designed based on coreference relations, i.e., answering the question requires resolving the coreference relation between two mentions in the context. However, it is shown that answering questions in Quoref does not necessarily require coreference resolution and the questions may be answered by using simple shortcuts in the dataset (Wu et al., 2020). Dua et al. (2020) annotate the required coreference relations for answering questions in a subset of Quoref examples. They then propose a model that jointly predict coreference relations and the final answer. They show that this joint prediction improves the result of the question answering model. They use gold annotations in their study, and they only annotate the relations that are related to the question. This work does not explicitly use a coreference resolver to obtain coreference relations and does not aim to resolve all coreference relations.

8 Discussion

As mentioned in §7, there are many ways to incorporate coreference information in QA. In this work, we make it at the input-level by decontextualizing the input sentences. This makes the extrinsic evaluations efficient and enables evaluating any coreference resolvers on any downstream models and datasets. On the downside, the decontextualization results in unnatural sentences in some examples, which may negatively impact the downstream model. For instance, we observe that most coreference resolvers have a negative impact on Sentence-BERT in Table 5. Meanwhile, we find that the other three CQA models are more robust on the revised data especially for the *rule-based* system. Overall, evaluating coreference resolution systems in downstream tasks is a complicated task. Various evaluation methods could result in very different extrinsic evaluation results on different downstream models and datasets that could be similar or dissimilar with standard coreference datasets. In this paper, our method evaluates coreference resolvers more on the out-of-domain corpora with less-formal text in a downstream task, community-based question answering.

9 Conclusions

Coreference resolution is an important step for text understanding. The main shortcoming of recent developments in coreference resolution is that they mainly target improving the performance in standard coreference datasets. However, coreference resolution is not an end-application and it is not clear how the progress in in-domain evaluations translates into downstream tasks performance. In this work, we enable direct extrinsic and intrinsic evaluation of coreference resolvers on downstream models and data, respectively. For the extrinsic evaluations, we use coreference resolvers for decontextualizing the input sentences in community-based question answering (CQA) task. For intrinsic evaluation, we have annotated a subset of CQA data with coreference relations. Our extrinsic evaluations suggest that (1) while there is a significant gap on the performances of state-of-the-art coreference resolver and the rule-based system on coreference datasets, the rule-based system has a more consistent and positive impact on CQA while the impact of the state-of-the-art model can considerably vary based on the domain of the downstream data, and (2) using coreference resolvers for decontextualizing both training and test datasets is more beneficial than decontextualizing the test data. Our intrinsic evaluations suggest that there is a discrepancy between the rankings of existing coreference resolution evaluation metrics and the resulting rankings from the extrinsic evaluations. This suggests that intrinsic evaluation on CoNLL should be accompanied by extrinsic evaluation to approximate the utility of the coreference resolvers for downstream tasks.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and suggestions. We also thank Friederike Lenke, Michael Tilli, Muhammad Usman and Fabian Dueker for coreference annotations on the CQA data. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies PhD scholarship.

References

Oshin Agarwal, Sanjay Subramanian, Ani Nenkova, and Dan Roth. 2019. Evaluation of named entity coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–7, Minneapolis, USA. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference, Volume 1*, pages 563–566. Citeseer.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1).

Haixia Chai and Michael Strube. 2022. Incorporating centering theory into neural coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2996–3002, Seattle, United States. Association for Computational Linguistics.

Haixia Chai, Wei Zhao, Steffen Eger, and Michael Strube. 2020. Evaluation of coreference resolution systems under adversarial attacks. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 154–159, Online. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Kevin Clark and Christopher D. Manning. 2016. 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)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Michael Collins. 2003. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637.

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.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. 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 2306–2317, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 42–48, New Orleans, Louisiana. Association for Computational Linguistics.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. 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.

Baorong Huang, Zhuosheng Zhang, and Hai Zhao. 2022. Tracing origins: Coreference-aware machine reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1281–1292, Dublin, Ireland. Association for Computational Linguistics.

Yin Jou Huang and Sadao Kurohashi. 2021. Extractive summarization considering discourse and coreference relations based on heterogeneous graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052, Online. Association for Computational Linguistics.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

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

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

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

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. CA: Sage Publications, Beverly Hills.

Klaus Krippendorff. 2004. Content analysis: An introduction to its methodology (2nd edition).

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.

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

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

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. volume 117, pages 30046–30054.

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

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Thomas S. Morton. 1999. Using coreference for question answering. In *Coreference and Its Applications*.

Thomas S. Morton. 2000. Coreference for NLP applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 173–180, Hong Kong. Association for Computational Linguistics.

C. Müller and M. Strube. 2006. Multi-level annotation of linguistic data with MMAX 2.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation.

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

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*. Association for Computational Linguistics.

Andreas Rücklé, Nafise Sadat Moosavi, and Iryna Gurevych. 2019. COALA: A neural coverage-based approach for long answer selection with small data. In *Proc. of AAAI-19*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek. 2007. Two uses of anaphora resolution in summarization. *Information Processing & Management*, 43(6):1663–1680.

Roland Stuckardt. 2016. Towards a procedure model for developing anaphora processing applications. In *Anaphora Resolution, Massimo Poesio, Roland Stuckardt and Yannick Versley*, pages 457–484. Springer.

Roland Stuckhardt. 2003. Coreference-based summarization and question answering: A case for high precision anaphor resolution. In *Proc. of the 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization*, pages 33–42.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473, Berlin, Germany. Association for Computational Linguistics.

Don Tuggener. 2014. Coreference resolution evaluation for higher level applications. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 231–235, Gothenburg, Sweden. Association for Computational Linguistics.

José L. Vicedo and Antonio Ferrández. 2000. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 555–562, Hong Kong. Association for Computational Linguistics.

José L. Vicedo and Antonio Ferrández. 2008. Coreference in Q&A. In *Advances in Open Domain Question Answering*, pages 71–96.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Kai Wang. 2010. *Retrieving questions and answers in community-based question answering services*. Ph.D. thesis.

Kai Wang, Zhao-Yan Ming, Xia Hu, and Tat-Seng Chua. 2010. Segmentation of multi-sentence questions: towards effective question retrieval in cqa services. In *Proc. of ACM-SIGIR-10*, pages 387–394.

Mingzhu Wu, Nafise Sadat Moosavi, Dan Roth, and Iryna Gurevych. 2020. Coreference reasoning in machine reading comprehension. *arXiv preprint arXiv:2012.15573*.

Liyan Xu and Jinho D. Choi. 2022. Online coreference resolution for dialogue processing: Improving mention-linking on real-time conversations. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 341–347, Seattle, Washington. Association for Computational Linguistics.

A Human Annotation Study

Annotation guidelines. Annotators were required to detect all noun phrase markables, except pronouns which were highlighted in advance through a chunker. Since the replacement for incorporating coreference information applies to coreferent mentions, non-referring markables (singletons) are not requested to be annotated.

Annotation procedure. The annotation process was run in three stages. First, we gave a training to the annotators to show the annotation guidelines and how to use the *MMAX2*. Then, home exercises containing 5 candidate answers from Travel domain were assigned to the annotators, so that we can point out problems they made promptly.

Finally, all annotators independently labeled annotations on their assigned work.

A.1 Inter-Annotator Agreement

To evaluate the reliability of the human annotations, we use Krippendorff's α (Krippendorff, 1980) to measure inter-annotator agreement, which allows for partial agreement among coreference chains by using distance metrics as weights. The alpha value can be affected by 'too strict' or 'too generous' distance metrics applied (Artstein and Poesio, 2008), so we report three different distance metrics, MASI (Passonneau, 2006), Jaccard (Jaccard, 1912) and Dice (Dice, 1945) for references. Since annotators can freely decide the boundary of markables, we use head-finding algorithm (Collins, 2003) for the overlapped markables identified by annotators to verify if they agree the markables are identical.

We randomly select two annotators to annotate the same 100 candidate answers from Travel domain. The final inter-annotator agreement was computed by the average of Krippendorff's α value of all answers. As showing in Table 11, our results are greater than 0.66 which was suggested as acceptable by Krippendorff (2004).

	MASI	Jaccard	Dice
Krippendorff's α	0.71	0.78	0.82

Table 11: Inter-Annotator agreement.

A.2 Error Analysis

In Table 12, we show three annotations from two annotators and one expert linguist on one answer from Travel domain. While two human annotators have similar coreference resolution results, the expert linguist resolves one more cluster that the annotators do not recognize. In addition, without the questions context, the annotation is sometimes harder for annotators.

H1: [I]$_1$ am from croatia and [I]$_1$ find their site confusing as well. Maybe [<url>]$_2$ can help [you]$_3$. imho, on [this link]$_2$ [you]$_3$ have very clear timetable for selected date if that is what [you]$_3$ want to find.
H2: [I]$_1$ am from croatia and [I]$_1$ find their site confusing as well. Maybe [<url>]$_2$ can help [you]$_3$. imho, on this [link]$_2$ [you]$_3$ have very clear timetable for selected date if that is what [you]$_3$ want to find.
L: [I]$_1$ am from croatia and [I]$_1$ find their site confusing as well. Maybe [<url>]$_2$ can help [you]$_3$. imho, on [this link]$_2$ [you]$_3$ have [very clear timetable for selected date]$_4$ if [that]$_4$ is what [you]$_3$ want to find.

Table 12: An example of human annotations on Travel domain by two annotators (**H1**) and (**H2**) and one expert linguist (**L**).

B Why applying coreference resolvers on candidate answers?

To incorporate coreference resolution, we can apply the coreference resolver on (1) the question, (2) the candidate answer, or (3) the concatenation of the question and each candidate answer. We examined all the above settings in our preliminary experiments, and we find out that the second one, i.e., resolving coreference relations of the candidate answers, is the most beneficial one. Questions are usually too short and do not contain coreference relations, so it is not useful to apply coreference resolvers on them.

To examine the third setting, we concatenate the question in the beginning of each candidate answer so that the model would be able to resolve intra-coreference relations among mentions of the candidate answer as well as inter-relations among the answer and the question. However, based on our experiments, the use of this setting results in lower performance in answer selection compared to the second one. The reason is that resolving coreference relations between candidate answers and the question makes many incorrect candidate answers more similar to the question by resolving the pronouns of the incorrect answer to the named entities of the question.[18] In addition, the question and answer have different speakers, which makes the resolution of first- and second-person pronouns more difficult across question-answer. Therefore, we only apply coreference resolvers to resolve the coreference relations of candidate answers.

C Results

Table 13 below shows the impact of pronoun resolution of the examined coreference resolvers on the answer selection models and domains. In this setting, we only replace pronouns with their most informative antecedent.

[18] For instance, the pronoun "it" from the incorrect candidate answer "You can get it by going to the closest grocery store", which is the answer of the question "where can I buy tomatoes?", can be resolved to "UK visa" from the other question, and makes the candidate answer more similar to this question.

Coreference	Answer Selection	Travel	Cooking	Computer	Aviation
rule-based	Sentence-BERT	-0.39	-1.14	-0.56	-0.77
	CNN	1.31	0.75	0.80	-0.61
	Att.-BiLSTM	1.17	1.14	0.80	0.92
	COALA	0.26	0.51	0.56	-0.15
deep-coref	Sentence-BERT	-0.39	-0.63	-0.40	-0.77
	CNN	0.39	0.37	0.40	-0.61
	Att.-BiLSTM	-0.66	0.76	0.16	0.92
	COALA	0.00	0.13	0.80	-0.15
e2e-coref	Sentence-BERT	0.13	-0.89	-0.16	-0.62
	CNN	0.78	0.88	0.16	0.46
	Att.-BiLSTM	-0.79	0.50	-0.16	0.46
	COALA	0.52	-0.38	0.00	0.46
bert-coref	Sentence-BERT	0.13	-1.01	-0.08	-0.31
	CNN	1.04	-0.26	-0.48	0.46
	Att.-BiLSTM	-0.53	0.13	-0.16	0.31
	COALA	0.13	-0.25	0.08	0.15

Table 13: Effect of the examined pronoun resolution on the answer selection models and datasets. Cell values indicate the difference in accuracy when incorporating pronoun resolution on test sets compared to the baseline results.