PCR4ALL: A Comprehensive Evaluation Benchmark for Pronoun Coreference Resolution in English

Xinran Zhao[1,3], **Hongming Zhang**[2,3], **Yangqiu Song**[3]
[1]Stanford University
[2]Tencent AI Lab, Seattle
[3]HKUST
xzhaoar@stanford.edu, hongmzhang@tencent.com, yqsong@cse.ust.hk

Abstract

Pronoun Coreference Resolution (PCR) is the task of resolving pronominal expressions to all mentions they refer to. The correct resolution of pronouns typically involves the complex inference over both linguistic knowledge and general world knowledge. Recently, with the help of pre-trained language representation models, the community has made significant progress on various PCR tasks. However, as most existing works focus on developing PCR models for specific datasets and measuring the accuracy or F1 alone, it is still unclear whether current PCR systems are reliable in real applications. Motivated by this, we propose PCR4ALL, a new benchmark and a toolbox that evaluates and analyzes the performance of PCR systems from different perspectives (i.e., knowledge source, domain, data size, frequency, relevance, and polarity). Experiments demonstrate notable performance differences when the models are examined from different angles. We hope that PCR4ALL can motivate the community to pay more attention to solving the overall PCR problem and understand the performance comprehensively. All data and codes are available at: https://github.com/HKUST-KnowComp/PCR4ALL.

Keywords: PCR, Format Unification, Multi-perspective Evaluation

1. Introduction

The question of how human beings resolve pronouns has long been of interest to both linguistic and natural language processing (NLP) communities. The situation that a pronoun itself only has weak semantic meaning brings challenges to natural language understanding systems. To explore solutions for that question, pronoun coreference resolution (PCR) (Hobbs, 1978) was proposed[1]. The first and second personal pronouns are typically not considered as they often refer to the current speakers, which are normally out of the conversation or document. Conventional PCR works (Ng, 2005; Zhang et al., 2019a; Zhang et al., 2019b) mostly focus on identifying coreference relations between pronouns and noun phrases rather than relations between pronouns. As a challenging yet vital natural language understanding task, PCR is to find the correct reference for a given pronominal anaphor in the context and has shown to be crucial for a series of downstream tasks, including machine translation (Mitkov et al., 1995), summarization (Steinberger et al., 2007), information extraction (Edens et al., 2003), and dialog systems (Strube and Müller, 2003).

The correct resolution of pronouns typically requires reasoning over both linguistic knowledge (e.g., "they" can only refer to plural objects) and commonsense knowledge (e.g., in sentence "The fish ate the worm, it was hungry", "it" refers to "fish" because hungry things tend to eat rather than being eaten). To investigate the possibility for machines to understand pronouns, many datasets were developed. However, due to the limitation of existing datasets, the performance on these datasets cannot effectively reflect the reliability of PCR systems in real applications. Current evaluation benchmarks mainly have four drawbacks: (1) Existing datasets mostly focus on a specific domain (e.g., CoNLL-2012 (Pradhan et al., 2012) for news and I2b2 (Uzuner et al., 2012) for medical) or specific reasoning types (e.g., WSC (Levesque et al., 2012) for commonsense) rather than providing an overall evaluation. (2) As most existing datasets follow the traditional machine learning setting (i.e., the training and test data follow the same distribution), it is unclear whether the progress on these datasets comes from understanding pronouns or naively capturing the distribution of the datasets. (3) With the reported accuracy on a single test set, it remains unclear if the PCR model is reliable under different circumstances, such as when the training data is small or different to the test data. (4) Different datasets may have different formats, which makes it challenging to train and evaluate PCR models across these datasets.

There is a growing body of literature that focuses on enhancing the model reliability through benchmark unification (Raganato et al., 2017) and measurement beyond accuracy (Ribeiro et al., 2020). Motivated by these works, in this paper, to address the limitation of existing PCR evaluation benchmarks, we propose PCR4ALL to help researchers analyze and compare PCR model performance under different circumstances

[1]Previous studies (Ng, 2005; Zhang et al., 2019b) mainly focus on three kinds of pronouns: third personal pronoun (e.g., *she*, *her*, *he*, *him*, *them*, *they*, *it*), possessive pronoun (e.g., *his*, *hers*, *its*, *their*, *theirs*), and demonstrative pronoun (e.g., *this*, *that*, *these*, *those*).

(e.g., when test cases are from other domains or with different relevance to the training data, etc.) To ensure the broad coverage, we include all major PCR datasets and standardize them into a unified format. To be informative, we propose to evaluate PCR systems from different perspectives rather than just the overall performance (i.e., a single test set accuracy). Carefully selected large-scale training, development, and test sets are provided to minimize the influence of artifacts. Performance of models on the standard set can then be viewed from other perspectives using specifically designed add-up tests, including how the performance changes when the data domain changes, data size grows, data being infrequent or irrelevant, or data being close to the distribution of the train set. These tests, as an evaluation toolbox or checklist, can then be used for further analysis or fair comparison across different PCR models.

Experiments demonstrate that even though we have made great progress on multiple datasets, none of the current systems can handle all perspectives of resolving pronouns very well. In addition, extensive model performance reports with multiple domains and perspectives lead to valuable and detailed understanding of the characteristics of various PCR models and datasets.

To summarize, the contribution of this paper is twofold: (1) we propose a new PCR evaluation benchmark PCR4ALL, which unifies the format of existing PCR datasets and provides automatically evaluation checklist on different perspectives of a pronoun coreference resolution model or system; (2) we conduct extensive experiments to point out the strengths and limitations of representative existing PCR models and systems.

2. Related Works

2.1. Previous PCR Datasets

Throughout the years, researchers in the NLP community have devoted great efforts to developing high-quality coreference resolution datasets (Grishman and Sundheim, 1996; Chinchor, 1998; Doddington et al., 2004; Pradhan et al., 2011; Pradhan et al., 2012)[2]. These general PCR datasets are mostly developed with expert annotations and focus on the newswire domain. Among these datasets, CoNLL-2012 (Pradhan et al., 2012) is the most popular one as it provides clear train, development, and test set separation as well as the official evaluation tool.

Another important line of work is the hard PCR datasets. Different from the general PCR task, the hard PCR datasets eliminate the effect of all commonly used linguistic knowledge (e.g., gender and plurality) via careful design, and focus on evaluating how models can understand commonsense knowledge that is required to resolve the pronouns. The most popular dataset is the Winograd Schema Challenge (WSC) (Levesque et al., 2012), which contains 273 carefully selected PCR questions[3]. Recently, to address the small scale problem of WSC, several similar datasets (i.e., DPR (Rahman and Ng, 2012), KnowRef (Emami et al., 2019), and WinoGrande (Sakaguchi et al., 2020)) have been proposed.

Last but not least, several other PCR datasets have been proposed to address different scenarios or perspectives related to the pronoun coreference resolution. For example, I2b2 (Uzuner et al., 2012) focuses on the medical domain, CIC (Chen and Choi, 2016) focuses on pronouns in multi-party dialogues, and WinoGender (Rudinger et al., 2018) studies the gender bias phenomenon in the process of pronoun resolution. In this work, to achieve a comprehensive understanding of the tested PCR systems, PCR4ALL includes all representative datasets and convert them into a unified format.

2.2. Existing PCR Models

Before the rise of deep learning, human-designed rules (Hobbs, 1978; Raghunathan et al., 2010; Chang et al., 2013) and features (Ng, 2005; Bengtson and Roth, 2008; Clark and Manning, 2015) dominated the general coreference resolution and PCR tasks. However, these features mainly reflect linguistic knowledge and cannot handle pronouns that require the correct understanding of semantics and background commonsense knowledge. To better represent the contextual semantics, an end-to-end deep model (Lee et al., 2017) was proposed and achieved surprisingly good performance without any human-defined rules. On top of it, several works (Lee et al., 2018; Zhang et al., 2019a; Zhang et al., 2019b; Kantor and Globerson, 2019; Joshi et al., 2020; Wu et al., 2020) have been proposed to address different limitations of the original end-to-end model. Recently, the pre-trained language models have almost dominated all NLP tasks including the benchmarks in PCR (Joshi et al., 2020; Sakaguchi et al., 2020). In this work, we will evaluate both the traditional and recent deep models with the proposed PCR4ALL benchmark to analyze their strengths and limitations.

3. PCR4ALL

PCR4ALL is a large-scale multi-perspective PCR benchmark with examples directly created by or automatically transformed from human annotations. PCR4ALL aims to evaluate the system performance on identifying the correct candidate that the pronoun refers to. Besides training and testing on a specific dataset, we also expect models to demonstrate their comparative generalizability and robustness in the multi-perspective evaluation. On the other hand, for any new dataset

[2]Some datasets (e.g., CoNLL-2012 shared task (Pradhan et al., 2012)) are originally designed for the general coreference (e.g., coreference among noun phrases) resolution task. Nonetheless, we can easily convert them into a PCR task.

[3]The latest version of WSC has 284 questions, but as most of the following works are evaluated based on the 273-question version, we still use the 273-question version in PCR4ALL.

Source	Context Sentence	Candidate 1	Candidate 2	Label
CoNLL-2012	With **their** unique charm , these cartoon images once again caused Hong Kong to be a focus of worldwide attention.	Hong Kong	these cartoon images	2
WSC style	Seymour sought Johnson's support , but _ long remained silent on the campaign.	Johnson	Seymour	1
I2b2	**He** was loaded with Dilantin ; serial head CT scans were performed on this young patient.	this young patient	serial head CT	1
CIC	Rachel Green says, Please . I haven't heard from her in seven months , and now **she** calls me ? ... She was my best Mindy Hunter ... she taught me how to kiss ..	Rachel Green	Mindy Hunter	2
WinoGender	The technician told the customer that _ can pay with cash.	technician	customer	2

Table 1: Examples from different datasets in the unified examples. WSC style datasets include the datasets with the same format of WSC (i.e., WSC, DPR, WinoGrande, and KnowRef). Target pronouns are in blue color. Correct and wrong candidates are indicated with the green and red colors, respectively. For datasets that do not contain the real pronouns, we use a pronoun place holder (i.e., "_") to represent the target pronoun. WinoGender dataset evaluates if the predictions are consistent when the pronoun is "he" or "she".

proposed in this area, we can also compare the multi-angle performance to test if the knowledge that models learned from this dataset can be easily transferred. The construction details of PCR4ALL are as follows.

3.1. Dataset Creation

We first select a few PCR representative datasets as the source of PCR4ALL before unification. We consider datasets from different domains (e.g., medical reports, TV series, and etc.) and formats (referred candidates span across a document or a sentence). In total, PCR4ALL is built from eight high-quality existing PCR datasets with details as follows: (1) **CoNLL-2012**: To address the scale problem of previous coreference datasets, CoNLL-2012 (Pradhan et al., 2012) was proposed. CoNLL-2012 focuses on the newswire domain and has been one of the most popular PCR benchmarks. (2) **Winograd Schema Challenge (WSC)**: To investigate if current coreference models can understand the commonsense knowledge needed for resolving pronouns, WSC (Levesque et al., 2012) removes the effect of all linguistic knowledge (e.g., gender and plurality) and formalize the PCR problem as a multiple-choice problem. The most widely used version of WSC contains 273 such questions. The PCR problem that only relies on commonsense knowledge is also called the hard pronoun coreference task in some other works. (3) **Definite Pronoun Resolution (DPR)**: Another hard PCR dataset is the definite pronoun resolution dataset (DPR) (Rahman and Ng, 2012). Different from WSC, DPR leveraged undergraduates rather than experts to create the dataset. In total, DPR collected 1,886 relatively simpler questions than WSC[4]. (4) **WinoGrande**: One common problem of WSC and DPR is their small scales. To create a larger scale data, WinoGrande (Sakaguchi et al., 2020) was proposed. By leveraging annotators from Amazon Mechanical Turk, WinoGrande collected 53 thousand WSC-like questions. (5) **KnowRef**: Similar to WinoGrande, KnowRef (Emami et al., 2019) aimed at creating a larger scale WSC dataset but with a different approach. Instead of using a crowd-sourcing + adversarial filtering framework, KnowRef tried to extract WSC-like questions from raw sentences. As a result, KnowRef collected 8 thousand WSC-like questions. (6) **I2b2** (Uzuner et al., 2012): A dataset focuses on identifying coreference relations in electronic medical records. As a dataset in a relatively narrow domain, the usage of domain knowledge is commonly considered as important. (7) **CIC** (Chen and Choi, 2016): A dataset focuses on identifying coreference relations in multi-party conversations. Compared with the ordinary PCR tasks, which are mostly annotated on formal textual data (e.g., newswire), identifying coreference relation in conversation is more challenging since the wording can be more casual. (8) **WinoGender** (Rudinger et al., 2018): it is proposed to study the gender bias phenomenon of current coreference systems in a similar setting with WSC. Different from accuracy used in WSC, the task requires model to report the consistency of queries with pronouns representing different genders. All the datasets mentioned above will be categorized and analyzed in the later sections.

[4]This dataset is also referred to as WSCR in some works.

3.2. Format Unification

In PCR4ALL, we convert all selected datasets into the same formulation such that we can easily and fairly evaluate different PCR models. As the research focus

of PCR is coreference resolution rather than mention detection (Zhang et al., 2020), we adopt the problem formulation of WSC (i.e., the setting of DPR, WinoGrande, KnowRef, and WinoGender): Given a text span, which contains a target pronoun and two candidate noun phrases, the task is to figure out which candidate does the target pronoun refers to. Converted examples in the unified format from different datasets are presented in Table 1. Although we conduct this pioneer research in English, all conversion rules can also be used for other languages.

For CoNLL-2012 and I2b2, each document contains several segments (one or two sentences) and the annotated co-referred mentions in different clusters. Since we focus on the pronoun coreference task, we only select the clusters that contain both pronouns and noun phrases. We then retrieve the segments that contain these pronouns as the target sentence and use the noun phrase as the positive example. We finally select the negative examples from other clusters and randomly assign the order of options of two candidates.

For CIC, for each scene in the TV show, we divide it into several sentence segments. For each segment, we conduct two kinds of formatting: (1) easy: we find the segments with pronouns as the annotated mentions. We mask the pronoun and use its referred character as the answer. Then, in the same sentence, we extract the noun phrases that are not referred to as negative examples; (2) hard: we generate questions with both options as character names. We extract the segments that contain two characters (e.g., A and B). We find the sentence where A is the speaker and B is referred to by both a pronoun and a noun phrase. We then replace the noun phase mention with character B's name, use "A says" as the prefix, and generate a question that distinguishes the pronoun coreference from the speaker (A) and the character (B).

3.3. Evaluation Toolbox

As aforementioned, a critical limitation of previous PCR benchmarks is that they only produce an overall performance (e.g., accuracy or F1) on a limited test set.To address this issue, in PCR4ALL, besides the overall performance, we also produce detailed evaluation results based on different perspectives (knowledge source, domain, data size, frequency, relevance, and polarity) of the PCR problem.

As a starting point, for each selected question, we will label it with one or several categories by the knowledge sources and domains (perspectives may not be mutually exclusive). When we report the performance from one perspective, we will only evaluate based on the corresponding questions. The detailed definition and how we select the questions are as follows: (1) **Linguistic Knowledge**: First of all, it is crucial to know how well PCR models can understand linguistic knowledge such as "he" normally can only refer to male and "they" to plural nouns. For each question, if only one candidate fits the linguistic requirement of the target pronoun, we will select this question. For the annotation, we design a rule-based system to automatically annotate if some linguistic requirements exist in the question. (2) **World Knowledge**: Based on the lower bound of a semantic theory (Katz and Fodor, 1963), language understanding needs the "speakers' knowledge of the language" and "his knowledge about the world". If linguistic knowledge is not a constraint, the understanding of language will depend on the world knowledge[5]. Thus, questions with both candidates fit linguistic requirements can be grouped into this category. (3) **Formal Language**: To evaluate how well models can handle formal language situations, where grammar errors or typos appear rarely (e.g., newswire, medical reports, and expert reviewed documents), this category contains all expert-annotated datasets. Questions are annotated to this category if their sources are expert-annotated datasets. (4) **Casual Language**: Besides the formal language, casual language is more popular in real applications (e.g., daily dialogue or online platforms). An important challenge of casual language are the potential typos, grammar errors, and incomplete sentences. To study how well PCR models can handle these cases, we select questions from dialogues (i.e., CIC) and crowd-sourced questions (i.e., WinoGrande) for the casual language category. (5) **Medical**: To study how well PCR models can handle corpus that requires domain-specific knowledge. We use the medical domain as an example and report based on the questions from the I2b2 dataset. (6) **Gender Bias**: Besides the accuracy, fairness is also a critical evaluation metric for current AI systems, and the PCR model should not be an exception. Motivated by this, we include questions from WinoGender (Zhao et al., 2018) to evaluate how well models can treat different gender fairly.

Besides the domain separation, we also regard other perspectives (data size, frequency, relevance, and polarity) as important measurements to help researchers understand the model robustness in different settings and conduct a fair and controlled comparison. Since paired identical sentences are evaluated in WinoGender, the gender bias issue is not evaluated for the perspectives below (except for data size). Included perspectives are defined as follows: (1) **Data Size**: To evaluate how the training data size influences the final performance, we set 6 thresholds on the randomly shuffled overall training data: 1, 10, 1,000, 10,000, 50,000, and 100,000 to reveal the expected performance on data with different sizes. (2) **Frequency**: To evaluate how performance differs on frequent and infrequent candidates, we break down the test set into subsets by the candidate frequency which is defined as the sum of token occurrences in the train set (excluding the stop words). (3) **Relevance**: As indicated by (Emami et al., 2020),

[5]World Knowledge is a bigger term than the commonsense knowledge because it also includes fact knowledge like "Obama was the president of US".

	Train	Dev	Test	Overall
Linguistic K.	12,422	1,537	1,514	15,473
World K.	90,918	11,381	11,404	113,703
Formal	50,118	6,221	6,392	62,731
Casual	53,222	6,697	6,526	66,445
Medical	15,270	1,887	1,986	19,143
Gender Bias	-	-	480	480
Overall	103,340	12,918	13,398	129,656

Table 2: PCR4ALL statistics. Number of training, development, and test set for each types are reported. K. denotes knowledge.

the relevance between train and test data will largely influence the PCR model performance. Following their work, we use BM25 (Amati, 2009) to score the dataset overlap. BM25 is a bag-of-words based approach to represent the document and score the relevance between a query and the document with weighted token coverage. In this metric, a lower score means less relevance. We break down the test set into subsets with similar sizes by the BM25 scores for each example. (4) **Polarity**: Ensuring the models are not simply remembering the train set is another important factor to revealing the model robustness. Motivated by this, we further compute the correlation between the candidates' polarity and model prediction. We define the candidate polarity as the sum of token probabilities of appearing in the correct candidates in the train set.

3.4. Statistics and Evaluation

We randomly split all collected questions, except those from WinoGender, into the training, development, and testing set based on the standard 8:1:1 separation. For WinoGender, we follow the original paper and use the whole dataset as the test set. Statistics of resulted PCR4ALL are presented in Table 2. For all evaluation perspectives except gender bias, examples evenly distribute over the train, development, and test sets.
All examples are formalized as binary choice problems. Since (Sakaguchi et al., 2020) has reached over 90% accuracy in a similar setting through fine-tuning pre-trained language models, we report error rate (instead of accuracy) to ensure clearer performance demonstration. For gender bias evaluation, we follow WinoGender that evaluates how genders affect prediction by the consistency of model predictions regarding different gender pronouns. For each question, we replace the pronoun with "he" and "she", test if the models give the same prediction, and report the consistency.

4. Experiments

Through the years, many approaches such as rule-based or feature-based systems (Hobbs, 1978; Raghunathan et al., 2010; Chang et al., 2013; Ng, 2005; Bengtson and Roth, 2008; Clark and Manning, 2015) have been proposed to resolve pronouns. Besides those PCR-oriented approaches, fine-tuned pre-trained language models also serve as important baselines to solve coreference resolution (Lee et al., 2018; Kantor and Globerson, 2019; Joshi et al., 2020; Wu et al., 2020). An important reason for its success is that rich semantic and world knowledge is learned via the pre-training (Petroni et al., 2019) and with the task-specific fine-tuning, the model can then learn how to use the acquired knowledge for the target task. Notice that (Joshi et al., 2020) is not compared as a baseline since it mainly focuses on improving mention detection, instead of pronoun coreference resolution. To clearly show the effect of both steps, we conduct experiments on both the vanilla language representation models and the fine-tuned ones with PCR4ALL. The details are as follows.

4.1. Existing Coreference systems

As the beginning of the evaluation, we test how the existing pronoun coreference systems perform on the dataset. The selected systems are as follows: (1) Stanford CoreNLP (Clark and Manning, 2016a; Clark and Manning, 2016b): Stanford CoreNLP group has provided a neural coreference system with reinforcement learning and entity-level distributed representations; (2) SpaCy (Honnibal et al., 2020): The SpaCy team adopts the hugginface's[6] implementation of a coreference resolution module with SpaCy parser and neural net scoring model based on (Clark and Manning, 2016a); (3) AllenNLP[7]: AllenNLP team implements another coreference resolution model with an end-to-end manner based on (Lee et al., 2017). All of these systems described above take a text span as input and output the identified clusters of text spans that refer to the same entity. To apply these systems to the PCR4ALL questions, we set the criterion of a successful identification as: both the pronoun and the correct candidate are in the same cluster. We report the error rate as the final results.

4.2. Vanilla LM

Recently, pre-trained contextualized language representation models have shown significant improvement over multiple NLP tasks including the mentioned PCR tasks. Besides the existing models, we also evaluate if these pre-trained models can also identify the more plausible options in the candidates. Following and adapted from (Sakaguchi et al., 2020), we utilize the representation models as multiple-choice solvers with two candidates as the options. We treat the candidates as options and add these options as apposition of the pronouns (e.g., replace "it" with "it, candidate, "). Then we treat the option with the highest plausibility as the answer.
The covered models are as follows: (1) **BERT** (Devlin et al., 2019): as a powerful contextualized lan-

[6] https://github.com/huggingface/neuralcoref
[7] https://demo.allennlp.org/coreference-resolution

Model	Linguistic K (Error Rate↓)	World K (Error Rate↓)	Formal (Error Rate↓)	Casual (Error Rate↓)	Medical (Error Rate↓)	Gender Bias (Consistency↑)
CoreNLP	63.29%	80.97%	61.28%	96.58%	82.43%	**94.58%**
SpaCy	51.61%	73.10%	46.30%	94.92%	61.58%	79.58%
AllenNLP	**44.83%**	65.53%	**32.97%**	93.24%	**46.68%**	84.58%
BERT	61.96%	57.15%	64.98%	50.46%	56.19%	**94.58%**
RoBERTa	47.13%	**48.07%**	47.31%	**48.61%**	50.40%	91.67%

Table 3: Overall Performance (error rate and consistency) of existing systems and vanilla language models. CoreNLP denotes the Stanford CoreNLP package. ↓ / ↑ indicates that a lower/higher score in this metric means better performance. The best-performing entries for each category are marked in bold.

Model	Train set	Linguistic K (Error Rate↓)	World K (Error Rate↓)	Formal (Error Rate↓)	Casual (Error Rate↓)	Medical (Error Rate↓)	Gender Bias (Consistency↑)
BERT	All Data	2.10%	5.72%	2.77%	8.02%	0.55%	91.67%
	Linguistic	2.59%	30.06%	14.09%	39.81%	10.78%	72.08%
	World	2.59%	6.16%	2.91%	8.59%	0.60%	85.42%
	Formal	11.19%	22.99%	3.63%	39.58%	0.70%	88.33%
	Casual	10.48%	30.45%	33.02%	23.45%	36.81%	87.92%
	Medical	20.56%	27.90%	8.08%	45.94%	0.76%	85.83%
RoBERTa	All Data	**1.68%**	4.06%	2.57%	5.00%	0.55%	94.58%
	Linguistic	1.75%	24.47%	9.36%	34.42%	6.65%	81.25%
	World	2.73%	**3.57%**	**2.33%**	4.62%	0.60%	**95.42%**
	Formal	10.28%	12.86%	2.40%	22.63%	0.60%	92.50%
	Casual	6.01%	9.27%	13.65%	**4.20%**	10.78%	90.42%
	Medical	15.03%	22.69%	5.17%	38.38%	**0.40%**	85.42%

Table 4: Performance in the inter-/intra-domain setting. ↓ / ↑ indicates that a lower/higher score in this metric means better performance. The best-performing entries for each category are marked in bold.

guage representation model, BERT-based models have become the state-of-the-art for many downstream NLP tasks. (2) **RoBERTa** (Liu et al., 2019): RoBERTa is an improved version of BERT with larger amount of training data and other techniques including dynamic masking. We use the large versions of both models.

4.3. Finetuned LMs

Besides the unsupervised vanilla models, fine-tuning is also an important technique to boost the performance, as indicated by (Kocijan et al., 2019). We also fine-tune the language representation models with different training data from different domains (e.g., linguistic knowledge, world knowledge, and etc.) and test on the test data from all domains. Notice that since we perform train-test split over the whole PCR4ALL dataset, train/test sets are mutually exclusive. We denote the model trained on all the training data as the practical upper bound for each language representation model. However, since in the real applications, test data might be from completely different domains than the training data, we produce both the intra-domain and inter-domain fine-tuning experiments as more detailed and generalizable analysis on how good the models can transfer learned knowledge among domains. For all the experiments, we use learning rate 1×10^{-5}, batch size 16, and epoch 3. With 4 GTX 1080 Ti GPUs, training on the whole PCR4ALL typically takes 24 hours. We perform uniform sampling to select the hyperparameters from 5 trials and final choices will be reported in the project code.

5. Performances and Analysis

5.1. Existing Systems

Table 3 presents the performance of existing systems on PCR4ALL. We can observe that current models that are not fine-tuned on the perspectives cannot solve the task well. Despite the good consistency on gender bias data, existing coreference systems surpass 50% error rate for most domains, especially for casual language data. One explanation is that these systems are mainly trained on formal documents (CoNLL-2012), which leads to relatively poor generalizability to casual conversations from TV series. LMs that are pre-trained to predict the plausibility of the sentences still can not distinguish the correct candidates that pronouns refer to without fine-tuning.

5.2. Influence of Domains

Table 4 presents the multi-perspective performance of different LMs trained on training data from different domains. The *All data* rows can be considered as both a reference for the model performances and a standard machine learning evaluation setting with a train test

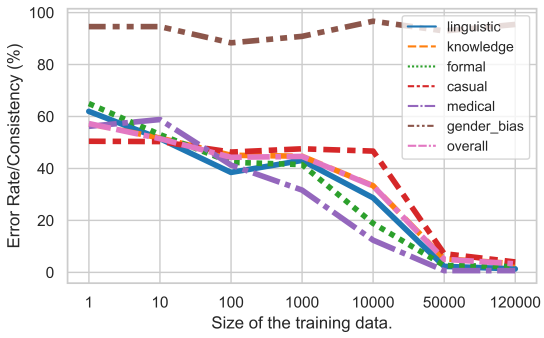

Figure 1: Performances on different perspectives with different sizes of the overall train set when finetuning RoBERTa. Lines with different colors indicate different subsets of the test data for different domains.

split. From the results, we can make following observations:

(1) In general, the consistency for gender bias data is similarly good across models. However, in many cases, fine-tuning leads to less consistency than the vanilla language models demonstrated in Table 4. For example, the consistency decreases significantly (-23.34% and -14.17% for BERT and RoBERTa, respectively, comparing with the best-performing model) with models trained on linguistic data, which suggests the importance of designing unbiased dataset that does not rely on gender heuristics.

(2) Models trained from different training data usually perform the best in their own domain. However, there are different levels of hardness for knowledge learned from different domains to be transferred to some other domains. Comparing the rows in the table, we can observe that training from *World knowledge* acquires the best domain transfer performance. In contrast, knowledge learned the *Linguistics* domain is harder to be transferred to others. For example, training the RoBERTa model with linguistic domain data, the error rates are around four to seven times larger than training with all data. Another interesting finding is that there exist a few domain pairs that lead to bad transfer learning performance (e.g., training on *Formal* and test on *Casual*, and vice versa). which further suggests the significance of the division of domains.

(3) On the other hand, some domains are harder than others to be solved by the knowledge learned from other domains. Comparing different columns of Table 4, taking results from RoBERTa as an example, we can observe that *Casual* test data is extremely hard to be solved in a transfer learning setting. Models fine-tuned from *Linguistic* and *Medical* data report 34.42% and 38.38% error rates on this test case, which is close to the majority votes. Considering that casual conversations can likely become the use case for a real-world PCR model, above results suggest the importance of involving test data from different domains, especially for those that are less tested in previous literature.

Model	Test Data	Error Rate ↓
BERT	All	5.40%
	MFC	**4.44%**
	LFC	5.93%
	Zero-Shot	8.92%
RoBERTa	All	3.79%
	MFC	**2.94%**
	LFC	4.28%
	Zero-Shot	5.84%

Table 5: Performance over subsets with different candidate frequencies. MFC and LFC denote the test sentences with the more frequent or less frequent candidates, respectively. Zero-shot denotes the sentences with the candidates that have never appeared in the training data.

5.3. Influence of Training Data Size

Figure 1 presents how the performance changes with the increasing size of training data used. We can observe that the general performance on gender consistency does not vary much with the increase of training data. Although in general the error rates from all the domains diminish quickly as the training data size increases, test data from these domains still show different levels of sensitivity towards the size change. For example, error rate on test data from medical domain (purple line in Figure 1) drops much faster than data from casual domain. There is a long flat curve for data from casual domain when the training data is small. One possible explanation is that, while medical reports have similar patterns and terms, daily conversations have much more abundance in word use and sentence structuring. Models are required to learn more data to capture the coreference pattern in daily conversations in the casual data domain. This observation suggests that when the training data is small, the innate domain characteristics of the test data can result in unclear performance comparison between domains. Uniformed large-scale training data is crucial if viewing the performance from different perspectives is expected.

For the future use of PCR4ALL benchmark, models can be tested on different size of the training data with the same order in this experiment. Then the performance can be compared with the data points provided to acquire a detailed understanding of how likely the model will work with different amount of data over different domains.

5.4. Influence of Frequency

To further analyze the performance, we divide the test set into more frequent candidates (MFC), less frequent candidates (LFC), and unseen candidates (Zero-Shot) subsets by the mean candidate frequency of both candidates in each question.

Table 5 presents how the test set performance changes with the candidate frequency. In general, we can ob-

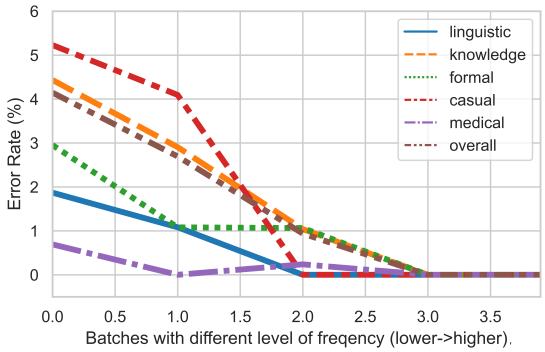

Figure 2: Performances on different perspectives with test subsets divided by their word frequency when fine-tuning RoBERTa with all training data.

serve that both models perform better on the examples with candidates that appear more often during training. In addition, while the performance gain from an improved pre-trained model (RoBERTa) is not restricted to a single subset (RoBERTa reduces the error rate by a large margin for both *MFC* and *Zero-shot*), it also slightly relieves the performance drop for unseen candidates, where the error rate gap between *MFC* and *Zero-shot* drops from 4.48% to 2.90%.

Furthermore, as a detailed study, Figure 2 presents how the performance against candidate frequency varies from different data domains, where the x-axis is a normalized indicator of the frequency (larger means higher frequency). The data points are grouped into batches, for example, examples with frequency 0 to 0.5 (right exclusive) will be considered as the group with frequency 0. In general, for most domains, the error rate drops quickly with the increase of examples' candidate frequency.

5.5. Influence of Relevance

With the aforementioned BM25 scores (details in Section 3.3), we empirically divide the test set into four batches with similar sizes (the score thresholds for splitting these batches are 47, 71, and 120; the relevance increases as the score increase).

Table 6 presents the influence of the relevance between train and test data. We can observe that for both the overall test data and most of its perspectives, the increase of BM25 scores will lead to a large decrease in the error rate. The significant effect of relevance suggests the importance of including the relevance view for a fair comparison between datasets.

Figure 3 presents the detailed analysis on different domains for the relevance perspective. We could observe that, though the performances on subsets from different perspectives generally increases as the relevance increases, *Linguistic* perspective examples are particularly sensitive. The reason behind can be that unseen linguistic patterns can be hard for the models.

We could also observe that there are some small peaks

Model	Test Data	Error Rate ↓
BERT	All	5.40%
	BM25 ∈ [0, 47]	10.16%
	BM25 ∈ (47, 71]	7.56%
	BM25 ∈ (71, 120]	2.98%
	BM25 ∈ (120, +∞)	**0.75%**
RoBERTa	All	3.79%
	BM25 ∈ [0, 47]	6.11%
	BM25 ∈ (47, 71]	5.56%
	BM25 ∈ (71, 120]	2.55%
	BM25 ∈ (120, +∞)	**0.88%**

Table 6: Performance over subsets with different relevance. Lower BM25 score indicates lower relevance.

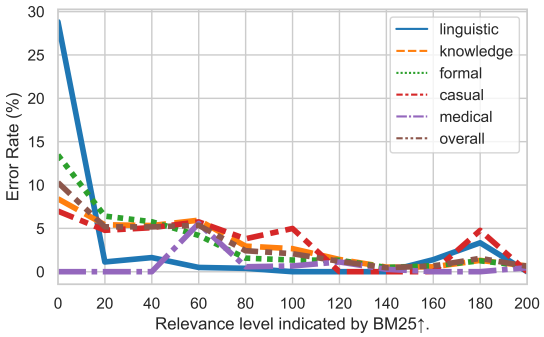

Figure 3: Performances on test subsets that have different levels of relevance to the training data when fine-tuning RoBERTa with all training data. The relevance level is indicated by aforementioned BM25.

with high relevance. The reason can be that the models predicted the "remebered" label for a similar test example with a flipped label, which we can commonly observe in WSC-style datasets.

5.6. Influence of Polarity

Table 7 demonstrates how the candidate polarity correlates with the model predictions. We can observe that positive correlation commonly exists on all models and most domains, which suggests that the label distribution of each word captured during training correlates with the model prediction during testing. Also, the differences in correlation mainly come from the change of test data domain, instead of the used model.

Medical data show larger correlation than data from other domains. One possible explanation is that many patterns occur in the training data can also appear often in the test data as the medical reports are highly formatted. In detail, there are two types of co-references in the original annotations of I2b2, where one is the ordinary pairs such as "he" refers to "the patient" and the other is about the syndromes or other terms such as "the syndrome" refers to "COPD" (Chronic Obstructive Pulmonary Disease). Tokens from the first type reoccur much more often than the second among all documents, which leads to higher polarity for tokens like

Test Data	BERT (Correlation)	RoBERTa (Correlation)
All	0.18138 (<.001)	0.18144 (<.001)
Linguistic K	0.14861 (<.001)	0.14719 (<.001)
World K	0.18677 (<.001)	0.18691 (<.001)
Formal	0.25789 (<.001)	0.25805 (<.001)
Casual	0.08343 (<.001)	0.08272 (<.001)
Medical	0.50276 (<.001)	0.50245 (<.001)

Table 7: The correlation between the candidate polarity and model predictions with Spearman's correlation (and two-tailed p-values).

"patient". Since the score for the model based on LMs could be viewed as which candidate occurs to be a better substitute for the pronoun in the sentence, LMs naturally assign higher probability to the common tokens instead of the rare ones (e.g., patient vs. vancomycin). Then the correlation between polarity and predictions is higher than other datasets.

We could also observe that questions CIC (from TV drama, Friends) achieves the lowest correlation, where the candidates are usually from the same pool of names and has no above-mentioned issue in I2b2. For evaluation, we only need to compare the polarity across models as part of the correlation comes from characteristics of datasets.

6. Conclusion

In this paper, we propose PCR4ALL, a unified large-scale benchmark for pronoun coreference resolution task that evaluates PCR systems from multiple perspectives, including knowledge source, domain, data size, frequency, relevance, and polarity. Multi-angle experiments included in the benchmark are bundled as a comprehensive evaluation toolbox to allow deep understanding of the performance and applicability of the systems beyond the overall accuracy or F1 score alone. We also further point out the strengths and limitations of current models through extensive experiments, such as the gap for zero-shot examples and the reliance on the data relevance.

PCR4LL provides a unified easy-to-use benchmark through careful split and labeling on the overall dataset. The unification allows fair and detailed comparison over newly proposed models or datasets. More importantly, the unification method and toolbox can be easily applied to other uncovered PCR tasks to extend PCR4ALL collaboratively. We hope that this benchmark can inspire the community to improve the evaluation of the robustness and real-world-applicability of PCR models. One potential future direction is to further extend the binary-choice problem to candidate ranking problem to further challenge current PCR methods.

7. Acknowledgement

We would like to thank the anonymous reviewers for their insightful advice and comprehensive knowledge over the area. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020- 19 and R6021-20) and the GRF (16211520) from RGC of Hong Kong, the MHKJFS (MHP/001/19) from ITC of Hong Kong and the National Key RD Program of China (2019YFE0198200) with special thanks to HKMAAC and CUSBLT, and the Jiangsu Province Science and Technology Collaboration Fund (BZ2021065).

8. Bibliographical References

Amati, G., (2009). *Encyclopedia of Database Systems: BM25*, pages 257–260. Springer US, Boston, MA.

Bengtson, E. and Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proceedings of EMNLP 2008*, pages 294–303.

Chang, K., Samdani, R., and Roth, D. (2013). A constrained latent variable model for coreference resolution. In *Proceedings of EMNLP 2013*, pages 601–612.

Chen, Y. and Choi, J. D. (2016). Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of SIGDIAL 2016*, pages 90–100.

Chinchor, N. A. (1998). Overview of MUC-7/MET-2. In *the Seventh Message Understanding Conference(MUC7)*.

Clark, K. and Manning, C. D. (2015). Entity-centric coreference resolution with model stacking. In *Proceedings of ACL 2015*, pages 1405–1415.

Clark, K. and Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas, November. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (2016b). 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, August. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 837–840, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Edens, R. J., Gaylard, H. L., Jones, G. J. F., and Lam-Adesina, A. M. (2003). An investigation of broad coverage automatic pronoun resolution for information retrieval. In *Proceedings of SIGIR 2003*, pages 381–382.

Emami, A., Trichelair, P., Trischler, A., Suleman, K., Schulz, H., and Cheung, J. C. K. (2019). The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of ACL 2019*, pages 3952–3961.

Emami, A., Suleman, K., Trischler, A., and Cheung, J. C. K. (2020). An analysis of dataset overlap on Winograd-style tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5855–5865, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Grishman, R. and Sundheim, B. (1996). Message understanding conference- 6: A brief history. In *Proceedings of COLING 1996*, pages 466–471.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4):311–338.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Kantor, B. and Globerson, A. (2019). Coreference resolution with entity equalization. In *Proceedings of ACL 2019*, pages 673–677.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *Language*, 39(2):170–210.

Kocijan, V., Cretu, A., Camburu, O., Yordanov, Y., and Lukasiewicz, T. (2019). A surprisingly robust trick for the winograd schema challenge. In *Proceedings of ACL 2019*, pages 4837–4842.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. In *Proceedings of EMNLP 2017*, pages 188–197.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of NAACL-HLT 2018*, pages 687–692.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The Winograd Schema Challenge. In *Proceeedings of KRR 2012*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mitkov et al., R. (1995). Anaphora resolution in machine translation. In *TMMT*.

Ng, V. (2005). Supervised ranking for pronoun resolution: Some recent improvements. In *Proceedings of AAAI 2005*, pages 1081–1086.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P. S. H., Bakhtin, A., Wu, Y., and Miller, A. H. (2019). Language models as knowledge bases? In Kentaro Inui, et al., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Pradhan, S., Ramshaw, L. A., Marcus, M. P., Palmer, M., Weischedel, R. M., and Xue, N. (2011). Conll-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of CoNLL 2011*, pages 1–27.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). Conll-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of CoNLL 2012*, pages 1–40.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April. Association for Computational Linguistics.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA, October. Association for Computational Linguistics.

Rahman, A. and Ng, V. (2012). Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of EMNLP-CoNLL 2012*, pages 777–789.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July. Association for Computational Linguistics.

Rudinger, R., Naradowsky, J., Leonard, B., and Durme, B. V. (2018). Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 8–14.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). WinoGrande: an adversarial winograd schema challenge at scale. In *Proceedings of AAAI 2020*, pages 99–106.

Steinberger, J., Poesio, M., Kabadjov, M. A., and Jeek, K. (2007). Two uses of anaphora reso-

lution in summarization. *Inf. Process. Manage.*, 43(6):1663–1680, nov.

Strube, M. and Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL 2003*, pages 168–175.

Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., and South, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *J. Am. Medical Informatics Assoc.*, 19(5):786–791.

Wu, W., Wang, F., Yuan, A., Wu, F., and Li, J. (2020). Corefqa: Coreference resolution as query-based span prediction. In Dan Jurafsky, et al., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6953–6963. Association for Computational Linguistics.

Zhang, H., Song, Y., and Song, Y. (2019a). Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of NAACL-HLT 2019*, pages 872–881.

Zhang, H., Song, Y., Song, Y., and Yu, D. (2019b). Knowledge-aware pronoun coreference resolution. In *Proceedings of ACL 2019*, pages 867–876.

Zhang, H., Zhao, X., and Song, Y. (2020). A brief survey and comparative study of recent development of pronoun coreference resolution. *CoRR*, abs/2009.12721.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT 2018*, pages 15–20.