A Brief Survey and Comparative Study of Recent Development of Pronoun Coreference Resolution in English

Hongming Zhang, Xinran Zhao, Yangqiu Song
Department of Computer Science, HKUST
hzhangal@cse.ust.hk, xzhaoar@connect.ust.hk, yqsong@cse.ust.hk

Abstract

Pronoun Coreference Resolution (PCR) is the task of resolving pronominal expressions to all mentions they refer to. Compared with the general coreference resolution task, the main challenge of PCR is the coreference relation prediction rather than the mention detection. As one important natural language understanding (NLU) component, pronoun resolution is crucial for many downstream tasks and still challenging for existing models, which motivates us to survey existing approaches and think about how to do better. In this survey, we first introduce representative datasets and models for the ordinary pronoun coreference resolution task. Then we focus on recent progress on hard pronoun coreference resolution problems (e.g., Winograd Schema Challenge) to analyze how well current models can understand commonsense. We conduct extensive experiments to show that even though current models are achieving good performance on the standard evaluation set, they are still not ready to be used in real applications (e.g., all SOTA models struggle on correctly resolving pronouns to infrequent objects). All experiment codes will be available upon acceptance.

1 Introduction

The question of how human beings resolve pronouns[1] has long been of interest to both linguistic and natural language processing (NLP) communities, for the reason that a pronoun itself only having weak semantic meaning brings challenges to natural language understanding. To explore solutions for that question, pronoun coreference resolution (PCR) (Hobbs, 1978) was proposed.[2] As a challenging yet vital natural language understanding task, pronoun coreference resolution is to find the correct reference for a given pronominal anaphor in the context and has been shown to be useful for a series of downstream tasks, such as machine translation (Mitkov et al., 1995; Lapshinova-Koltunski et al., 2018), summarization (Steinberger et al., 2007), and dialog systems (Strube and Müller, 2003).

Type	# Pairs	F1 (no mention)	F1 (mention)
NP-NP	25,828	0.690	0.768
NP-P	43,883	0.667	0.707
P-P	41,741	0.754	0.763
Overall	111,452	0.705	0.742

Table 1: The performance of the End-to-end model on the CoNLL-2012 shared task coreference resolution dataset. The model's performances of different coreference types are reported separately.

To investigate the difference between PCR and the general coreference resolution task, which tries to identify not only the coreference relations between noun phrases (NP) and pronouns (P) but also potential coreference relations between noun phrases or coreference relations between pronouns, we conduct experiments with one recent breakthrough model (i.e., End-to-end model (Lee et al., 2017)) on the CoNLL-2012 shard task (Pradhan et al., 2012) under two settings: one without the gold mention and one with the gold mention. In the 'without gold mention' setting, models are required to first identify spans from the documents as the mentions and then predict the coreference relations among these mentions. As a comparison, if gold

[1] Some pronouns may refer to non-nominal antecedents. For example, the pronoun "it" in "It is too cold in the Winter here" does not refer to any real object (Kolhatkar et al., 2018). But in this survey, we only focus on pronouns that refer to nominal antecedents.

[2] PCR is also known as anaphora resolution (Versley et al., 2016). Previous studies (Ng, 2005; Zhang et al., 2019c) 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*). 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. Besides that, conventional PCR works (Ng, 2005; Zhang et al., 2019b,c) mostly focusing on identifying coreference relations between pronouns and noun phrases rather than coreference relation between pronouns.

Proceedings of the 4th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021), pages 1–11
Punta Cana, Dominican Republic, November 10–11, 2021.

mentions are provided, models only need to predict the coreference relations (i.e., the task of distinguishing between referential and non-referential instances is ignored). From the results in Table 1, we can see that, without the gold mention, the model performs well on P-P coreference relations, while not as well on the other two kinds of relations. However, if gold mentions are provided, the model can achieve very good performance on the NP-NP coreference relations. Compared with other kinds of coreference relations, no matter whether the gold mention is provided or not, resolving pronouns to noun phrases is always the most challenging one.

The correct resolution of pronouns typically requires reasoning over both linguistic knowledge (e.g., 'they' typically refers to plural objects[3]) 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 be eaten.). Considering that the ordinary PCR task evaluates the inference over both types of knowledge at the same time, the performance on ordinary PCR tasks cannot clearly reflect models' performance regarding different knowledge types. To address this problem, the Winograd Schema Challenge (WSC) (Levesque et al., 2012) task is proposed. The influence of all commonly used linguistic knowledge is avoided during the creation of WSC such that WSC can be used to reflect how current PCR models can understand commonsense knowledge. In Section 2 and 3, we introduce the progress and remaining challenges on the ordinary PCR and WSC tasks respectively. After that, we introduce other PCR tasks that are developed for different research purposes in Section 4. In the end, we conclude this survey with Section 5. The contribution of this survey is three-fold: (1) we broadly introduce available PCR tasks, datasets, and models; (2) We summarize the main contribution of recent models; (3) We conduct experiments to analyze the limitations of current models, which can help the community think about how to better solve PCR in the future.

2 Ordinary PCR

Ordinary pronoun coreference resolution tasks are often defined over formal textual corpus (e.g., newspaper) and the annotation is usually conducted by domain experts or linguists. The PCR task can be formally defined as follows. Given a text D, which contains a pronoun p, the goal is to identify all the mentions that p refers to. We denote the correct mentions p refers to as $c \in \mathcal{C}$, where $\mathcal{C}$ is the correct mention set. Similarly, each candidate span is denoted as $s \in \mathcal{S}$, where $\mathcal{S}$ is the set of all candidate spans. Note that in the case where no golden mentions are provided, all possible spans in D are used to form $\mathcal{S}$. The task is thus to identify $\mathcal{C}$ out of $\mathcal{S}$. In the rest of this section, we introduce the widely used datasets as well as the progress and limitation of current approaches.

2.1 Datasets

Throughout the years, researchers in the NLP community have devoted great efforts to developing high-quality coreference resolution datasets[4] and we introduce representative ones as follows:

1. **MUC**: MUC-6 (Grishman and Sundheim, 1996) and MUC-7 (Chinchor, 1998), which were developed for the 6^{th} and 7^{th} message understanding conferences respectively, are the earliest coreference resolution datasets. They are focusing on English news articles and are relatively small compared with modern datasets.

2. **ACE**: The ACE dataset (Doddington et al., 2004) was proposed as part of the Automatic Content Extraction program. Compared with MUC datasets, ACE extends the corpus domain from news to other domains like telephonic speeches and broadcast conversations.

3. **CoNLL shared tasks**: CoNLL-2011 (Pradhan et al., 2011) and CoNLL-2012 (Pradhan et al., 2012) shared tasks were proposed to evaluate models' abilities of resolving unrestricted coreference resolution. Among these two, CoNLL-2011 only contains annotations in English and CoNLL-2012 extends the coverage to multiple other languages (i.e., Chinese and Arabic). Compared with MUC and ACE, CoNLL shared tasks have a much larger scale. Moreover, as CoNLL-2012 shared tasks provide clear training, dev, and test set separation as well as the official evaluation tool, it is the most widely

[3] One exception is the entities that are related to organizations. For example, "they" can refer to "the company" (Hardmeier et al., 2018). Another exception is to prevent generic masculine, where "they" can refer to singular entity in gender-neutral language.

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

used evaluation benchmark for the coreference resolution task.

4. **WikiCoref**: Recently, a new coreference dataset WikiCoref (Ghaddar and Langlais, 2016) was proposed as a supplementary of CoNLL shared tasks. Different from CoNLL, where most of the corpus is from the newswire, WikiCoref directly annotates Wikipedia pages, which provides a new way to evaluate models' performances in the out-of-domain setting.

5. **Crowd-sourced Coref**: Poesio et al. (2019) leveraged a crowd-sourced game to collect 2.2 million annotations about 108,000 coreference relations, which makes it one of the largest coreference datasets. Moreover, their annotations also include ambiguous coreference relations.

2.2 Methods

In this subsection, we introduce representative models for the ordinary PCR task. We first briefly introduce conventional approaches that rely on human-designed rules or features and then introduce the end-to-end model, which is a groundbreaking model for solving coreference resolution tasks. After that, we briefly introduce a few recent improvements over the end-to-end model.

2.2.1 Rule and Feature Based Methods

Before the deep learning era, human-designed rules (Hobbs, 1978; Raghunathan et al., 2010), knowledge (Ponzetto and Strube, 2006; Versley et al., 2016), and features (Ng, 2005; Wiseman et al., 2016) dominated the general coreference resolution and PCR tasks. Some rules and features are crucial for correctly resolving pronouns (Lee et al., 2013). For example, 'he' typically refers to males and 'she' typically refers to females; 'it' typically refers to singular objects and 'them' typically refers to plural objects. The performances of these methods heavily rely on the coverage and quality of the manually defined rules and features. Based on these designed features (Bengtson and Roth, 2008), a few more advanced machine learning models were applied to the coreference resolution task. For example, instead of identifying coreference relation pair-wisely, (Clark and Manning, 2015) proposes an entity-centric coreference system that can learn an effective policy for building coreference chains incrementally. Besides that, a novel model was also proposed to predict coreference relations with a deep reinforcement learning framework (Clark and Manning, 2016). Moreover, heuristic rules based on linguistic knowledge can also be incorporated into constraints for machine learning models (Chang et al., 2013).

2.2.2 End-to-end Model

Leveraging human-designed rules or features can help accurately resolve some pronouns, but it is hard to manually design rules to cover all cases. To solve this problem, an end-to-end deep model (Lee et al., 2017) was proposed. Different from other machine learning-based methods, it does not use any human-defined rules, yet achieves surprisingly good performance. Specifically, the end-to-end model first leverages the combination of Bi-directional LSTM and inner-attention modules to encode local context and generate representations for all potential mentions. After that, a standard feed-forward neural network is used to predict the coreference relations. Experiment results show that the proposed model is simple yet effective. Its success proves that current deep models are capable of capturing rich contextual information, which is crucial for resolving coreference relations.

2.2.3 Further Improvements

Recently, on top of the end-to-end model, a few improved works were proposed to address different limitations of the original end-to-end model[5]:

1. **Higher-order Information**: One limitation of the original end-to-end model is that all predictions are based on pairs, which is not sufficient for capturing higher-order coreference relations. To fix this issue, a differentiable approximation module was proposed in (Lee et al., 2018) to provide the higher-order coreference resolution inference ability (i.e., leveraging the coreference cluster to better predict the coreference relations). Moreover, this work first incorporates ELMo (Peters et al., 2018), a kind of deep contextualized word representations, as part of the word representation, which is proven very effective.

2. **Structured Knowledge**: Another limitation of the end-to-end model is that its success heavily relies on the quality and coverage of the training data. However, in real applications, it is labor-intensive and almost impossible to annotate a large-scale dataset to contain all scenar-

[5]These models once achieved better performance either on the general coreference resolution task or the PCR task.

Model	Third Personal (18,147)			Possessive (6,843)			Demonstrative (546)			Overall (25,536)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Deterministic (Raghunathan et al., 2010)	25.5	58.9	35.6	22.9	64.3	33.8	3.4	5.7	4.2	23.4	57.0	33.4
Statistical (Clark and Manning, 2015)	25.8	62.1	36.5	28.9	64.9	40.0	9.8	6.3	7.6	25.4	59.3	36.5
Deep-RL (Clark and Manning, 2016)	78.6	63.9	70.5	73.3	68.9	71.0	3.7	2.9	5.5	76.4	61.2	68.0
End-to-end (Lee et al., 2017)	70.7	77.8	74.1	75.6	74.0	74.8	37.8	71.7	49.5	68.3	76.4	72.1
+ KG (Zhang et al., 2019c)	80.0	75.6	77.7	81.7	72.2	76.7	50.8	**64.6**	**56.9**	77.9	74.0	75.9
+ SpanBERT (Joshi et al., 2020)	**82.4**	**80.5**	**81.5**	**83.9**	**81.0**	**82.4**	**52.0**	61.5	56.4	**82.2**	**80.2**	**81.2**

Table 2: Performances of different models on the CoNLL-2012 shared task. Precision (P), recall (R), and the F1 score are reported. Numbers of different types of pronouns in the test set are shown in the brackets. Best models are indicated with the **bold** font.

Model	Training data	Test data	
		CoNLL	i2b2
End-to-end	CoNLL	72.1	75.2
	i2b2	20.0	92.3
+ KG	CoNLL	75.9	80.9
	i2b2	42.7	95.2
+ SpanBERT	CoNLL	79.6	40.8
	i2b2	28.5	80.5

Table 3: Models' performance (in F1 score) in cross-domain setting on different training/test data.

ios. To solve this problem, two works (Zhang et al., 2019b,c) were proposed to inject external structured knowledge into the end-to-end model. Among these two, (Zhang et al., 2019b) requires converting external knowledge into features while (Zhang et al., 2019c) directly uses external knowledge in the format of triplets.

3. **Stronger Language Representation Models**: Recently, along with the fast development of language representation models, a few works (Kantor and Globerson, 2019; Joshi et al., 2020) have been trying to replace the encoding layer of the original end-to-end model with more powerful language representation models. SpanBERT (Joshi et al., 2020) replaces ELMo with SpanBERT and boosts the performance by 6.6 F1 over the general coreference resolution task.

2.3 Performances and Analysis

We follow the experimental setting of (Zhang et al., 2019c) and test the performance[6] of representative models (Raghunathan et al., 2010; Clark and Manning, 2015, 2016; Lee et al., 2017; Zhang et al., 2019c; Joshi et al., 2020) on the CoNLL-2012 dataset (Pradhan et al., 2012). The experiment setting (both detection the mentions and resolving the coreference relations) and evaluation metric are the same as these previous works on CoNLL-2012. From the results in Table 2, we can observe that with the help of the end-to-end model and further modifications, the community has made great progress on the standard evaluation set. For example, the end-to-end model achieves an F1 score over 70 and adding external knowledge (either in a structured way or a representation way) further boost the performance. Among all pronoun types, all models perform better on third personal and possessive pronouns, and relatively poorly on demonstrative ones. This is mainly because of the imbalanced distribution of the dataset (i.e., third personal and possessive pronouns appear much more than demonstrative ones).

2.3.1 Cross-domain Performance

To investigate whether current PCR models are good enough to be used in real applications, which can be out of the training domain, we conduct experiments on the cross-domain setting. In detail, we select two different PCR datasets from different domains (i.e., CoNLL (Pradhan et al., 2012) from news and i2b2 (Uzuner et al., 2012) from the medical domain) and try to train the model on one dataset and test it on the other. We conduct experiments with three best-performing models and show the results in Table 3, from which we can see that all models[7] perform significantly worse if they

[6] We use the released codes of different models along with their default hyper-parameters to finish the experiments. For the end2end model, we also include ELMo (Peters et al., 2018) as part of the representation and achieve better performance than the original one in Table 1.

[7] SpanBERT performs poorly on i2b2 when it is not trained on it. The reason can be that the medical corpus is too different from the pre-trained corpus of SpanBERT and we use the default hyper-parameters, which might not be the best ones. Since the main contribution of SpanBERT is helping models to identify the mention spans, in our setting focusing on reference detection, such improvement is not necessary

Model	Object Type	P	R	F1
End-to-End	Infrequent	66.5	73.8	70.0
	Frequent	73.0	83.3	77.8
+ KG	Infrequent	77.9	72.5	75.1
	Frequent	78.0	77.7	77.9
+ SpanBERT	Infrequent	71.3	72.4	71.9
	Frequent	83.3	85.3	84.3

Table 4: Influence of the frequency.

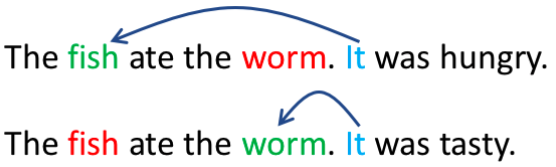

Figure 1: WSC question examples.

are used across domains (i.e., when the domains of training and test data are different). Compared with the baseline method, adding explicit knowledge can help achieve slightly better performance in the cross-domain setting because its training objective allows models to learn to selectively use suitable knowledge rather than just fitting the training data.

2.3.2 Influence of Frequency

To further analyze the performance of existing models, we split the pronouns based on the frequency of the objects they refer to. If an object appears more than ten times in the whole dataset, we denote it as a frequent object. Otherwise, we denote it as an infrequent object. As a result, we collect 1,095 frequent and 470,232 infrequent objects, whose average frequencies are 36.2 and 1.46 respectively. We report the performance of best-performing models on infrequent and frequent objects separately in Table 4. In general, all models perform better on frequent objects because they appear more in the training data. Another interesting observation is that, even though adding external KG and a stronger language representation model can both boost the performance, their improvements come from different types of objects. For example, the main contribution of adding KG is on infrequent objects because even though they are less frequent in the training data, they can still be covered by some external knowledge. As a comparison, using a strong language representation model mainly benefits the frequent objects because it has a stronger ability to fit the training data. This observation is consistent with our previous observations that adding external KG has more effect on those relatively rare pronouns (i.e., demonstrative pronouns).

3 Hard PCR

As aforementioned, the correct resolution of pronouns requires the inference over both linguistic knowledge and commonsense knowledge. To clearly reflect how models can resolve pronouns that require the inference over commonsense knowledge, the hard PCR task was proposed. As Winograd Schema Challenge (WSC) is one of the most popular hard PCR tasks, we use the task definition in WSC to define the hard PCR task. For each question q, a sentence s is given, which contains a pronoun p and two candidates n_1, n_2. The task is to find out which of the candidates p refers to. Different from the ordinary PCR task, the influence of all commonly observed features (e.g., gender or plurality) are removed via careful expert design. In WSC, all questions are paired up such that questions in each pair have only minor differences (mostly one-word difference), but the answers are reversed. One pair of the WSC instances is shown in Figure 1. Solving these questions typically requires the support of complex commonsense knowledge. For example, human beings can know that the pronoun 'it' in the first sentence refers to 'fish' while the one in the second sentence refers to 'worm' because 'hungry' is a common property of something eating while 'tasty' is a common property of something being eaten. Without the support of such commonsense knowledge, answering these questions becomes challenging because both the fish and worm can be hungry or tasty by themselves.

3.1 Datasets

We introduce datasets as follows:

1. **Winograd Schema Challenge**: Among all the hard pronoun coreference resolution tasks, WSC is among the most popular ones. In total, WSC has 273 questions[8]. Its small size determines that it cannot be used to train a good supervised model and can only be used as the evaluation set.

2. **Definite Pronoun Resolution**: Another hard pronoun coreference resolution dataset

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

is the definite pronoun resolution dataset (DPR)[9] (Rahman and Ng, 2012). Different from WSC, DPR leveraged undergraduates rather than experts to create the dataset. In total, DPR collected 1,886 questions, which is a slightly larger scale than the official WSC. However, as DPR can not guarantee that all DPR questions follow the strict design guideline of WSC, questions in DPR are relatively simpler.

3. **WinoGrande**: One common problem of WSC and DPR is their small scales. To create a larger scale dataset, WinoGrande (Sakaguchi et al., 2020) was proposed. By leveraging annotators from Amazon Mechanical Turk, WinoGrande collected 53 thousand WSC-like questions. Moreover, to make sure of the dataset quality, WinoGrande applied a bias reduction algorithm to filter out examples that may contain annotation bias. Experimental results prove that WinoGrande is much more challenging than the original WSC because the SOTA models on WSC only achieve 51% accuracy on WinoGrande, which is similar to the random guess.

4. **KnowRef**: KnowRef (Emami et al., 2019), similar to WinoGrande, also aimed at creating a larger scale WSC dataset but with a different approach. Instead of using crowd-sourcing + adversarial filtering framework, KnowRef tried to extract WSC-like questions from raw sentences. As a result, KnowRef collected eight thousand WSC-like questions.

3.2 Methods

In this subsection, we introduce existing approaches for the hard PCR task. As the majority of the methods are evaluated based on WSC, all the discussion and analysis are based on their performance on WSC.

3.2.1 Reasoning with Structured Knowledge

At first, people tried to leverage different commonsense knowledge resources to solve WSC questions in an explainable way. For example, Liu et al. (2016) first leveraged the commonsense triplets from ConceptNet (Liu and Singh, 2004) to train the word embeddings and then applied the embeddings to solve the WSC task. Knowledge hunter (Emami et al., 2018) proposed to leverage search engines (e.g., Google) to acquire needed commonsense knowledge. It first searched WSC questions in search engines and then used the returned searching results to solve WSC questions. SP-10K (Zhang et al., 2019a) conducted experiments to show that selectional preference (SP) knowledge such as human beings are more likely to eat 'food' rather than 'rock' can also be helpful for solving WSC questions. Last but not least, ASER (Zhang et al., 2020) tried to use knowledge about eventualities (e.g., 'being hungry' can cause 'eat food') to solve WSC questions. In general, structured commonsense knowledge can help solve one-third of the WSC questions, but their overall performance is limited due to their low coverage. There are mainly two reasons: (1) coverage of existing commonsense resources are not large enough; (2) lack of a principled way to use structured knowledge for NLP tasks. Current methods (Emami et al., 2018; Zhang et al., 2019a, 2020) mostly rely on string match. However, for many WSC questions, it is hard to find supportive knowledge in such way.

3.2.2 Language Representation Models

Another approach is leveraging language models to solve WSC questions (Trinh and Le, 2018), where each WSC question is first converted into two sentences by replacing the target pronoun with the two candidates respectively and then the language models can be employed to compute the probability of both sentences. The sentence with a higher probability will be selected as the final prediction. As this method does not require any string match, it can make prediction for all WSC questions and achieve better overall performance. Recently, a more advanced transformer-based language model GPT-2 (Radford et al., 2019) achieved better performance due to its stronger language representation ability. The success of language models demonstrates that rich commonsense knowledge can be indeed encoded within language models implicitly. Another interesting finding about these language model based approaches is that they proposed two settings to predict the probability: (1) Full: use the probability of the whole sentence as the final prediction; (2) Partial: only consider the probability of the partial sentence after the target pronoun. Experiments show that the partial model always outperforms the full model. One explanation is that the influence of the imbalanced distribution of candidate words is relieved by only considering the sentence probability after them. Such observation also explains why GPT-2 can outperform unsuper-

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

	Methods	Correct	Wrong	NA	A_p	A_o
Unsupervised	Random Guess	137	136	0	50.2%	50.2%
	Knowledge Hunting (Emami et al., 2018)	119	79	75	60.1%	57.3%
	SP (Human) (Zhang et al., 2019a)	15	0	258	**100%**	52.7%
	SP (PP) (Zhang et al., 2019a)	50	26	197	65.8%	54.4%
	ASER (String Match) (Zhang et al., 2020)	63	27	183	70.0%	56.6%
	LM (Single) (Trinh and Le, 2018)	149	124	0	54.5%	54.5%
	LM (Ensemble) (Trinh and Le, 2018)	168	105	0	61.5%	61.5%
	GPT-2 (Radford et al., 2019)	193	80	0	70.7%	70.7%
Finetuning	BERT (Devlin et al., 2019) +ASER (Zhang et al., 2020)	177	96	0	64.5%	64.5%
	BERT (Devlin et al., 2019) +DPR (Rahman and Ng, 2012)	195	78	0	71.4%	71.4%
	BERT (Devlin et al., 2019) +WinoGrande (Sakaguchi et al., 2020)	210	63	0	76.9%	76.9%
	RoBERTa (Liu et al., 2019) +DRP (Rahman and Ng, 2012)	227	46	0	83.1%	83.1%
	RoBERTa (Liu et al., 2019) +WinoGrande (Sakaguchi et al., 2020)	246	27	0	90.1%	**90.1%**
Human Beings	Original (Levesque et al., 2012)	252	21	0	92.1%	92.1%
	Recent (Sakaguchi et al., 2020)	264	9	0	96.5%	96.5%

Table 5: Performances of different models on the 273-question version WSC. *NA* means that the model cannot give a prediction, A_p means the accuracy of predict examples without *NA* examples. And A_o the overall accuracy of all examples (i.e., Correct, Wrong, and *NA* examples)

vised BERT on WSC because models based on BERT, which relies on predicting the probability of candidate words, cannot get rid of such noise.

3.2.3 Fine-tuning Representation Models

Last but not least, we introduce current best-performing models on the WSC task, which fine-tunes pre-trained language representation models (e.g., BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019)) with a similar dataset (e.g., DPR (Rahman and Ng, 2012) or WinoGrande (Sakaguchi et al., 2020)). CorefBERT (Ye et al., 2020), in addition, introduced a new pre-training task that requires models to predict mention references. This idea of fine-tuning was originally proposed by (Kocijan et al., 2019), which first converts the original WSC task into a token prediction task and then selects the candidate with higher probability as the final prediction. In general, the stronger the language models and the larger the fine-tuning datasets are, the better the model can perform on the WSC task.

3.3 Performances and Analysis

To clearly understand the progress we have made on solving hard PCR problems, we show the performance of all models on Winograd Schema challenge in Table 5. From the results, we can make the following observations:

1. Even though methods that leverage structured knowledge can provide explainable solutions to WSC questions, their performance is typically limited by their low coverage.
2. Different from them, language model based methods represent knowledge contained in human language with an implicit approach, and thus do not have the matching issue and achieve better overall performance.
3. In general, fine-tuning pre-trained language representation models (e.g., BERT and RoBERTa) with similar datasets (e.g., DPR and WinoGrande) achieve the current SOTA performances and two observations can be made: (1) The stronger the pre-trained model, the better the performance. This observation shows that current language representation models can indeed cover commonsense knowledge and along with the increase of their representation ability (e.g., deeper model or larger pre-training corpus like RoBERTa), more commonsense knowledge can be effectively represented. (2) The larger the fine-tuning dataset, the better the performance. This is probably because the knowledge about some WSC questions is only covered by WinoGrande but not in DPR.

To investigate the reason behind WinoGrande's success, we divide WinoGrande into subsets based on the instances' relevance towards WSC. Assuming that the instance set of WinoGrande and WSC are $\mathcal{I}_{WG}$ and $\mathcal{I}_{WSC}$ respectively, for each instance $i \in \mathcal{I}_{WG}$, we design its relevance score as follows:

$$R_{WSC}(i) = Max(\frac{O^2(i,i')}{L(i) \cdot L(i')}, i' \in \mathcal{I}_{WSC}), \quad (1)$$

L.R. \ Rel.	High (13,466)	Medium (13,466)	Low (13,466)	Overall (40,398)
1e-6	87.81%	85.63%	84.95%	**88.89%**
2e-6	87.46%	87.81%	50.53%	**90.32%**
5e-6	87.10%	86.74%	50.17%	**91.76%**
1e-5 (default)	**87.81%**	85.66%	84.94%	87.46%
2e-5	53.04%	51.25%	52.33%	**84.58%**
5e-5	51.97%	50.09%	51.97%	**55.56%**
1e-4	**53.75%**	53.05%	52.69%	51.06%
Average	72.71%	71.46%	61.08%	**78.51%**

Table 6: Performance of fine-tuning RoBERTa with different learning rates and three subsets of WinoGrande split by their instances' relevance towards the original WSC. L.R. means learning rate and Rel. means relevance to WSC data. Numbers of instances are shown in brackets. Best performed datasets for each learning rate is indicated with the **bold** font.

where $O(i, i')$ is the unigram co-occurrence of i and i' and $L()$ is the instance length. We use the released code and dataset to conduct the experiments and follow all hyper-parameters as the original paper (Sakaguchi et al., 2020) except the batch size[10].

From the results in Table 6, we can observe that: (1) The most relevant instances contribute the most to the success. In some learning rate settings, it performs similar to or even better than the overall set; (2) Less relevant instances also help, which shows that current fine-tuning approach is not just fitting the data but also learning some underneath knowledge about solving the task from the data; (3) The model can be sensitive to the hyper-parameters (i.e., learning rate). Different subsets have different best hyper-parameters and the learning process can easily fail with a bad hyper-parameter choice. To achieve a good performance on a fixed dataset like WSC, we can tune the hyper-parameters. But to create a reliable PCR system we can rely on in real life, we probably need a more robust model.

4 Other PCR Tasks

Besides the ordinary and hard PCR tasks, PCR is also an important research topic for many special purposes (e.g., gender bias) or in some special settings (e.g., Visual-aware PCR). In this section, we briefly introduce these tasks:

1. **PCR in the Medical Domain**: I2b2 (Uzuner et al., 2012) is a dataset that focuses on identifying coreference relations in electronic medical records. As reported in (Zhang et al., 2019c), the training set of I2b2 contains 2,024 third personal pronouns, 685 possessive pronouns, and 270 demonstrative pronouns. Its test set contains 1,244 third personal pronouns, 367 possessive pronouns, and 166 demonstrative pronouns. As a dataset in a relatively narrow domain, the usage of domain knowledge becomes important. As shown in (Zhang et al., 2019c), i2b2 can be used as an additional dataset to evaluate models' cross-domain abilities.

2. **PCR for Machine Translation**: ParCor (Guillou et al., 2014) and ParCorFull (Lapshinova-Koltunski et al., 2018) are datasets focusing on PCR in parallel multi-lingual datasets, which can be used in downstream machine translation tasks. Different from other PCR works, it focuses on how to leverage the PCR results for better translation rather than how to solve the PCR problem.

3. **PCR for Chatbots**: CIC (Chen and Choi, 2016) is a dataset focusing 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.

4. **PCR for Studying Gender Bias**: Nowadays, gender bias has been a hot research topic in the NLP community (Rudinger et al., 2018; Zhao et al., 2018). WinoGender (Rudinger et al., 2018) is among the most popular works. The setting of WinoGender is similar to the setting of WSC (Levesque et al., 2012), where each sentence contains one target pronoun and two candidate noun phrases and the models are required to select the correct antecedent from the two candidates. But the purpose is different. WSC aims at evaluating models' abilities to understand commonsense knowledge, while WinoGender aims at evaluating how well models can predict without the influence of gender bias. The experiments show that some gender bias (e.g., 'he' is more likely to be predicted to be the doctor rather than the nurse by the machine) indeed exists in pre-trained language representation models. Such observation is astonishing and motivates the community to think about how to minimize the influence of such gender bias.

[10]The original batch size is 16 and our batch size is 4 due to the GPU memory limitation, so the experimental result is slightly different from the one reported in the original paper.

5. **Visual-aware PCR**: Recently, a visual-aware PCR dataset (Yu et al., 2019), which evaluates how well models can ground pronouns to visual objects, was proposed. Similar to CIC (Chen and Choi, 2016), visual-aware PCR also focuses on pronouns in daily dialogue, where the language usage is informal and a lot of background knowledge can be missing. For example, if one speaker refers to something both speakers can see, they may directly use a pronoun rather than introduce it first. In such a case, a pronoun may refer to not mentioned objects in the conversation. As analyzed in the original paper, 15% of pronouns in conversations refer to not mentioned objects and for them, leveraging the visual context information becomes crucial. As shown in (Kottur et al., 2018), grounding pronouns to the visual objects can significantly help the model to better understand the dialog and generate the better response, which further proves that visual-aware PCR is an important research topic to explore.

5 Conclusion

In this paper, we survey the progress on the pronoun coreference resolution (PCR) task, and analyze the improvement and limitations of existing approaches. Experiments and analysis on both the ordinary and hard PCR tasks demonstrate that even though we have made great progress according to the main evaluation metric, the PCR task is still far away from being solved. For example, all best-performing ordinary PCR models struggle on the cross-domain setting as well as infrequent objects. Also, even though fine-tuning pre-trained language representation models can achieve near-human performance on WSC, it can be sensitive to the hyper-parameters. All codes will be released to encourage the research on the PCR task.

Acknowledgements

This paper was supported by Early Career Scheme (ECS, No. 26206717), General Research Fund (GRF, No. 16211520), and Research Impact Fund (RIF, No. R6020-19) from the Research Grants Council (RGC) of Hong Kong.

References

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP 2008*, pages 294–303.

Kai-Wei Chang, Rajhans Samdani, and Dan Roth. 2013. A constrained latent variable model for coreference resolution. In *Proceedings of EMNLP 2013*, pages 601–612.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of SIGDIAL 2016*, pages 90–100.

Nancy A Chinchor. 1998. Overview of muc-7/met-2. In *the Seventh Message Understanding Conference(MUC7)*.

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

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of EMNLP 2016*, pages 2256–2262.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of LREC 2004*.

Ali Emami, Noelia De La Cruz, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2018. A knowledge hunting framework for common sense reasoning. In *Proceedings of EMNLP 2018*, pages 1949–1958.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of ACL 2019*, pages 3952–3961.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of LREC 2016*.

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

Liane Guillou, Christian Hardmeier, Aaron Smith, Jörg Tiedemann, and Bonnie L. Webber. 2014. Parcor 1.0: A parallel pronoun-coreference corpus to support statistical MT. In *Proceedings of LREC 2014*, pages 3191–3198.

Christian Hardmeier, Luca Bevacqua, Sharid Loáiciga, and Hannah Rohde. 2018. Forms of anaphoric reference to organisational named entities: Hoping to

widen appeal, they diversified. In *Proceedings of the Seventh Named Entities Workshop*, pages 36–40.

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

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of ACL 2019*, pages 673–677.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of ACL 2019*, pages 4837–4842.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Survey: Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of ECCV 2018*, pages 160–178.

Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of LREC 2018*.

Heeyoung Lee, Angel X. 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 EMNLP 2017*, pages 188–197.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of NAACL-HLT 2018*, pages 687–692.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceeedings of KRR 2012*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Quan Liu, Hui Jiang, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. 2016. Commonsense knowledge enhanced embeddings for solving pronoun disambiguation problems in winograd schema challenge. *arXiv preprint arXiv:1611.04146*.

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

Ruslan Mitkov et al. 1995. Anaphora resolution in machine translation. In *TMMT*.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pages 2227–2237.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of NAACL-HLT 2019*, pages 1778–1789.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of NAACL-HLT 2006*, pages 192–199.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proceedings of CoNLL 2012*, pages 1–40.

Sameer Pradhan, Lance A. Ramshaw, Mitchell P. Marcus, Martha Palmer, Ralph M. Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of CoNLL 2011*, pages 1–27.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher D. Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP 2010*.

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

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINOGRANDE: an adversarial winograd schema challenge at scale. In *Proceedings of AAAI 2020*.

Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Jevzek. 2007. Two uses of anaphora resolution in summarization. *Information Processing & Management*.

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

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

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

Yannick Versley, Massimo Poesio, and Simone Ponzetto. 2016. *Using Lexical and Encyclopedic Knowledge*, pages 393–429. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of NAACL-HLT 2016*, pages 994–1004.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *Proceedings of EMNLP 2020*.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5122–5131.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019a. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of ACL 2019*, pages 722–731.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *Proceedings of WWW 2020*, pages 201–211.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019b. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of NAACL-HLT 2019*, pages 872–881.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. 2019c. Knowledge-aware pronoun coreference resolution. In *Proceedings of ACL 2019*, pages 867–876.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT 2018*, pages 15–20.