Coreferential Reasoning Learning for Language Representation

Deming Ye[1,2], **Yankai Lin**[4], **Jiaju Du**[1,2], **Zhenghao Liu**[1,2], **Peng Li**[4], **Maosong Sun**[1,3], **Zhiyuan Liu**[1,3]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology
[2]State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China
[3]Beijing Academy of Artificial Intelligence
[4]Pattern Recognition Center, WeChat AI, Tencent Inc.
ydm18@mails.tsinghua.edu.cn

Abstract

Language representation models such as BERT could effectively capture contextual semantic information from plain text, and have been proved to achieve promising results in lots of downstream NLP tasks with appropriate fine-tuning. However, most existing language representation models cannot explicitly handle coreference, which is essential to the coherent understanding of the whole discourse. To address this issue, we present CorefBERT, a novel language representation model that can capture the coreferential relations in context. The experimental results show that, compared with existing baseline models, CorefBERT can achieve significant improvements consistently on various downstream NLP tasks that require coreferential reasoning, while maintaining comparable performance to previous models on other common NLP tasks. The source code and experiment details of this paper can be obtained from https://github.com/thunlp/CorefBERT.

1 Introduction

Recently, language representation models such as BERT (Devlin et al., 2019) have attracted considerable attention. These models usually conduct self-supervised pre-training tasks over large-scale corpus to obtain informative language representation, which could capture the contextual semantic of the input text. Benefiting from this, language representation models have made significant strides in many natural language understanding tasks including natural language inference (Zhang et al., 2020), sentiment classification (Sun et al., 2019b), question answering (Talmor and Berant, 2019), relation extraction (Peters et al., 2019), fact extraction and verification (Zhou et al., 2019), and coreference resolution (Joshi et al., 2019).

However, existing pre-training tasks, such as masked language modeling, usually only require models to collect local semantic and syntactic information to recover the masked tokens. Hence, language representation models may not well model the long-distance connections beyond sentence boundary in a text, such as coreference. Previous work has shown that the performance of these models is not as good as human performance on the tasks requiring coreferential reasoning (Paperno et al., 2016; Dasigi et al., 2019), and they can be further improved on long-text tasks with external coreference information (Cheng and Erk, 2020; Xu et al., 2020; Zhao et al., 2020). Coreference occurs when two or more expressions in a text refer to the same entity, which is an important element for a coherent understanding of the whole discourse. For example, for comprehending the whole context of "*Antoine published The Little Prince in 1943. The book follows a young prince who visits various planets in space.*", we must realize that *The book* refers to *The Little Prince*. Therefore, resolving coreference is an essential step for abundant higher-level NLP tasks requiring full-text understanding.

To improve the capability of coreferential reasoning for language representation models, a straightforward solution is to fine-tune these models on supervised coreference resolution data. Nevertheless, on the one hand, we find fine-tuning on existing small coreference datasets cannot improve the model performance on downstream tasks in our preliminary experiments. On the other hand, it is impractical to obtain a large-scale supervised coreference dataset.

To address this issue, we present CorefBERT, a language representation model designed to better capture and represent the coreference information. To learn coreferential reasoning ability from large-scale unlabeled corpus, CorefBERT introduces a novel pre-training task called *Mention Reference Prediction* (MRP). MRP leverages those repeated mentions (e.g., noun or noun phrase) that appear

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7170–7186,
November 16–20, 2020.

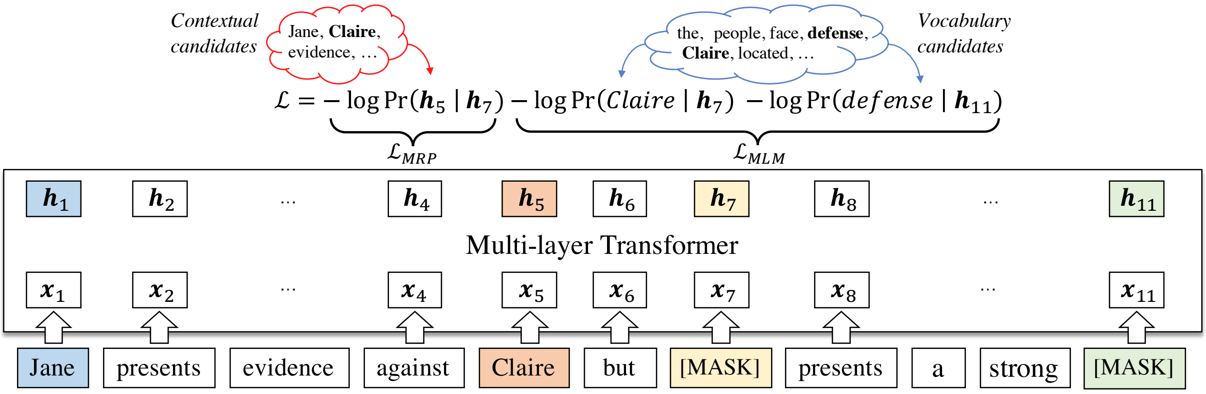

Figure 1: An illustration of CorefBERT's training process. In this example, the second *Claire* and a common word *defense* are masked. The overall loss of *Claire* consists of the loss of both Mention Reference Prediction (MRP) and Masked Language Modeling (MLM). MRP requires model to select contextual candidates to recover the masked tokens, while MLM asks model to choose from vocabulary candidates. In addition, we also sample some other tokens, such as *defense* in the figure, which is only trained with MLM loss.

multiple times in the passage to acquire abundant co-referring relations. Among the repeated mentions in a passage, MRP applies mention reference masking strategy to mask one or several mentions and requires model to predict the masked mention's corresponding referents. Figure 1 shows an example of the MRP task, we substitute one of the repeated mentions, *Claire*, with `[MASK]` and ask the model to find the proper contextual candidate for filling it. To explicitly model the coreference information, we further introduce a copy-based training objective to encourage the model to select words from context instead of the whole vocabulary. The internal logic of our method is essentially similar to that of coreference resolution, which aims to find out all the mentions that refer to the masked mentions in a text. Besides, rather than using a context-free word embedding matrix when predicting words from the vocabulary, copying from context encourages the model to generate more context-sensitive representations, which is more feasible to model coreferential reasoning.

We conduct experiments on a suite of downstream tasks which require coreferential reasoning in language understanding, including extractive question answering, relation extraction, fact extraction and verification, and coreference resolution. The results show that CorefBERT outperforms the vanilla BERT on almost all benchmarks and even strengthens the performance of the strong RoBERTa model. To verify the model's robustness, we also evaluate CorefBERT on other common NLP tasks where CorefBERT still achieves comparable results to BERT. It demonstrates that the introduction of the new pre-training task about coreferential reasoning would not impair BERT's ability in common language understanding.

2 Related Work

Pre-training language representation models aim to capture language information from the text, which facilitate various downstream NLP applications (Kim, 2014; Lin et al., 2016; Seo et al., 2017). Early works (Mikolov et al., 2013; Pennington et al., 2014) focus on learning static word embeddings from the unlabeled corpus, which have the limitation that they cannot handle the polysemy well. Recent years, contextual language representation models pre-trained on large-scale unlabeled corpora have attracted intensive attention and efforts from both academia and industry. SA-LSTM (Dai and Le, 2015) and ULMFiT (Howard and Ruder, 2018) pre-trains language models on unlabeled text and perform task-specific fine-tuning. ELMo (Peters et al., 2018) further employs a bidirectional LSTM-based language model to extract context-aware word embeddings. Moreover, OpenAI GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) learn pre-trained language representation with Transformer architecture (Vaswani et al., 2017), achieving state-of-the-art results on various NLP tasks. Beyond them, various improvements on pre-training language representation have been proposed more recently, including (1) designing new pre-trainning tasks or objectives such as SpanBERT (Joshi et al., 2020) with span-based learning, XLNet (Yang et al., 2019) considering masked positions dependency with auto-regressive loss, MASS (Song et al., 2019) and BART (Wang et al., 2019b) with sequence-to-sequence pre-training, ELECTRA (Clark et al., 2020) learning from replaced token detection with generative adversarial networks and InfoWord (Kong et al., 2020) with contrastive learning; (2) integrating external knowledge such as factual knowledge in knowledge graphs (Zhang et al., 2019; Peters et al., 2019; Liu et al., 2020a); and (3) exploring multilingual learning (Conneau and Lample, 2019; Tan and Bansal, 2019; Kondratyuk and Straka, 2019) or multimodal learning (Lu et al., 2019; Sun et al., 2019a; Su et al., 2020). Though existing language representation models have achieved a great success, their coreferential reasoning capability are still far less than that of human beings (Paperno et al., 2016; Dasigi et al., 2019). In this paper, we design a mention reference prediction task to enhance language representation models in terms of coreferential reasoning.

Our work, which acquires coreference resolution ability from an unlabeled corpus, can also be viewed as a special form of unsupervised coreference resolution. Formerly, researchers have made efforts to explore feature-based unsupervised coreference resolution methods (Bejan et al., 2009; Ma et al., 2016). After that, Word-LM (Trinh and Le, 2018) uncovers that it is natural to resolve pronouns in the sentence according to the probability of language models. Moreover, WikiCREM (Kocijan et al., 2019) builds sentence-level unsupervised coreference resolution dataset for learning coreference discriminator. However, these methods cannot be directly transferred to language representation models since their task-specific design could weaken the model's performance on other NLP tasks. To address this issue, we introduce a mention reference prediction objective, complementary to masked language modeling, which could make the obtained coreferential reasoning ability compatible with more downstream tasks.

3 Methodology

In this section, we present CorefBERT, a language representation model, which aims to better capture the coreference information of the text. As illustrated in Figure 1, CorefBERT adopts the deep bidirectional Transformer architecture (Vaswani et al., 2017) and utilizes two training tasks:

(1) **Mention Reference Prediction** (MRP) is a novel training task which is proposed to enhance coreferential reasoning ability. MRP utilizes the mention reference masking strategy to mask one of the repeated mentions and then employs a copy-based training objective to predict the masked tokens by copying from other tokens in the sequence.

(2) **Masked Language Modeling** (MLM)[1] is proposed from vanilla BERT (Devlin et al., 2019), aiming to learn the general language understanding. MLM is regarded as a kind of cloze tasks and aims to predict the missing tokens according to its final contextual representation. Except for MLM, Next Sentence Prediction (NSP) is also commonly used in BERT, but we train our model without the NSP objective since some previous works (Liu et al., 2019; Joshi et al., 2020) have revealed that NSP is not as helpful as expected.

Formally, given a sequence of tokens[2] $X = (x_1, x_2, \ldots, x_n)$, we first represent each token by aggregating the corresponding token and position embeddings, and then feeds the input representations into deep bidirectional Transformer to obtain the contextual representations, which is used to compute the loss for pre-training tasks. The overall loss of CorefBERT is composed of two training losses: the mention reference prediction loss L_{MRP} and the masked language modeling loss L_{MLM}, which can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{MRP}} + \mathcal{L}_{\text{MLM}}. \quad (1)$$

3.1 Mention Reference Masking

To better capture the coreference information in the text, we propose a novel masking strategy: mention reference masking, which masks tokens of the repeated mentions in the sequence instead of masking random tokens. We follow a distant supervision assumption: the repeated mentions in a sequence would refer to each other. Therefore, if we mask one of them, the masked tokens would be inferred through its context and unmasked references. Based on the above strategy and assumption, the CorefBERT model is expected to capture the coreference information in the text for filling the masked token.

In practice, we regard nouns in the text as mentions. We first use a part-of-speech tagging tool to extract all nouns in the given sequence. Then, we cluster the nouns into several groups where each group contains all mentions of the same noun. After that, we select the masked nouns from different groups uniformly. For example, when *Jane* occurs

[1]Details of MLM are in the appendix due to space limit.
[2]In this paper, tokens are at the subword level.

three times and *Claire* occurs two time in the text, all the mentions of *Jane* or *Claire* will be grouped. Then, we choose one of the groups, and then sample one mention of the selected group.

To maintain the universal language representation ability in CorefBERT, we utilize both the MLM (masking random word) and MRP (masking mention reference) in the training process. Empirically, the masked words for MLM and MRP are sampled on a ratio of 4:1. Similar to BERT, 15% of the tokens are sampled for both masking strategies mentioned above, where 80% of them are replaced with a special token `[MASK]`, 10% of them are replaced with random tokens, and 10% of them are unchanged. We also adopt whole word masking (WWM) (Joshi et al., 2020), which masks all the subwords belong to the masked words or mentions.

3.2 Copy-based Training Objective

In order to capture the coreference information of the text, CorefBERT models the correlation among words in the sequence. Inspired by copy mechanism (Gu et al., 2016; Cao et al., 2017) in sequence-to-sequence tasks, we introduce a copy-based training objective to require the model to predict missing tokens of the masked mention by copying the unmasked tokens in the context. Since the masked tokens would be copied from context, low-frequency tokens, such as proper nouns, could be well processed to some extent. Moreover, through copying mechanism, the CorefBERT model could explicitly capture the relations between the masked mention and its referring mentions, therefore, to obtain the coreference information in the context.

Formally, we first encode the given input sequence $X = (x_1, \ldots, x_n)$ into hidden states $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n)$ via multi-layer Transformer (Vaswani et al., 2017). The probability of recovering the masked token x_i by copying from x_j is defined as:

$$\Pr(x_j|x_i) = \frac{\exp((\boldsymbol{V} \odot \boldsymbol{h}_j)^T \boldsymbol{h}_i)}{\sum_{x_k \in X} \exp((\boldsymbol{V} \odot \boldsymbol{h}_k)^T \boldsymbol{h}_i)}, \quad (2)$$

where $\odot$ denotes element-wise product function and $\boldsymbol{V}$ is a trainable parameter to measure the importance of each dimension for token's similarity.

Moreover, since we split a word into several word pieces as BERT does and we adopt whole word masking strategy for MRP, we need to extend our copy-based objective into word-level. To this end, we apply the token-level copy-based training objective on both start and end tokens of the masked word, because the representations of these two tokens could typically cover the major information of the whole word (Lee et al., 2017; He et al., 2018). For a masked noun w_i consisting of a sequence of tokens $(x_s^{(i)}, \ldots, x_t^{(i)})$, we recover w_i by copying its referring context word, and define the probability of choosing word w_j as:

$$\Pr(w_j|w_i) = \Pr(x_s^{(j)}|x_s^{(i)}) \times \Pr(x_t^{(j)}|x_t^{(i)}). \quad (3)$$

A masked noun possibly has multiple referring words in the sequence, for which we collectively maximize the similarity of all referring words. It is an approach widely used in question answering (Kadlec et al., 2016; Swayamdipta et al., 2018; Clark and Gardner, 2018) designed to handle multiple answers. Finally, we define the loss of Mention Reference Prediction (MRP) as:

$$\mathcal{L}_{\text{MRP}} = -\sum_{w_i \in M} \log \sum_{w_j \in C_{w_i}} \Pr(w_j|w_i), \quad (4)$$

where M is the set of all masked mentions for mention reference masking, and C_{w_i} is the set of all corresponding words of word w_i.

4 Experiment

In this section, we first introduce the training details of CorefBERT. After that, we present the fine-tuning results on a comprehensive suite of tasks, including extractive question answering, document-level relation extraction, fact extraction and verification, coreference resolution, and eight tasks in the GLUE benchmark.

4.1 Training Details

Since training CorefBERT from scratch would be time-consuming, we initialize the parameters of CorefBERT with BERT released by Google[3], which is also used as our baselines on downstream tasks. Similar to previous language representation models (Devlin et al., 2019; Joshi et al., 2020), we adopt English Wikipeida[4] as our training corpus, which contains about 3,000M tokens. We employ spaCy[5] for part-of-speech-tagging on the corpus. We train CorefBERT with contiguous sequences of up to 512 tokens, and randomly shorten the input sequences with 10% probability in training. To verify the effectiveness of our method for the language

[3]https://github.com/google-research/bert
[4]https://en.wikipedia.org
[5]https://spacy.io

representation model trained with tremendous corpus, we also train CorefBERT initialized with RoBERTa[6], referred as CorefRoBERTa. Additionally, we follow the pre-training hyper-parameters used in BERT, and adopt Adam optimizer (Kingma and Ba, 2015) with batch size of 256. Learning rate of 5×10^{-5} is used for the base model and 1×10^{-5} is used for the large model. The optimization runs 33k steps, where the learning rate is warmed-up over the first 20% steps and then linearly decayed. The pre-training process consumes 1.5 days for base model and 11 days for large model with 8 RTX 2080 Ti GPUs in mixed precision. We search the ratio of MRP loss and MLM loss in 1:1, 1:2 and 2:1, and find the ratio of 1:1 achieves the best result. Beyond this, training details for downstream tasks are shown in the appendix.

4.2 Extractive Question Answering

Given a question and passage, the extractive question answering task aims to select spans in passage to answer the question. We first evaluate models on Questions Requiring Coreferential Reasoning dataset (QUOREF) (Dasigi et al., 2019). Compared to previous reading comprehension benchmarks, QUOREF is more challenging as 78% of the questions in QUOREF cannot be answered without coreference resolution. In this case, it can be an effective tool to examine the coreferential reasoning capability of question answering models.

We also adopt the MRQA, a dataset not specially designed for examining coreferential reasoning capability, which involves paragraphs from different sources and questions with manifold styles. Through MRQA, we hope to evaluate the performance of our model in various domains. We use six benchmarks of MRQA, including SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), SearchQA (Dunn et al., 2017), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), and Natural Questions (NaturalQA) (Kwiatkowski et al., 2019). Since MRQA does not provide a public test set, we randomly split the development set into two halves to generate new validation and test sets.

Baselines For QUOREF, we compare our CorefBERT with four baseline models: (1) **QANet** (Yu et al., 2018) combines self-attention mechanism with the convolutional neural network, which achieves the best performance to date without pre-training; (2) **QANet+BERT** adopts BERT representation as an additional input feature into QANet; (3) **BERT** (Devlin et al., 2019), simply fine-tunes BERT for extractive question answering. We further design two components accounting for coreferential reasoning and multiple answers, by which we obtain stronger BERT baselines; (4) **RoBERTa-MT** trains RoBERTa on CoLA, SST2, SQuAD datasets before on QUOREF. For MRQA, we compare CorefBERT to vanilla BERT with the same question answering framework.

[6] https://github.com/pytorch/fairseq

Model	Dev EM	Dev F1	Test EM	Test F1
QANet*	34.41	38.26	34.17	38.90
QANet+BERT$_{\text{BASE}}$ *	43.09	47.38	42.41	47.20
BERT$_{\text{BASE}}$ *	58.44	64.95	59.28	66.39
BERT$_{\text{BASE}}$	61.29	67.25	61.37	68.56
CorefBERT$_{\text{Base}}$	**66.87**	**72.27**	**66.22**	**72.96**
BERT$_{\text{LARGE}}$	67.91	73.82	67.24	74.00
CorefBERT$_{\text{LARGE}}$	**70.89**	**76.56**	**70.67**	**76.89**
RoBERTa-MT+	74.11	81.51	72.61	80.68
RoBERTa$_{\text{LARGE}}$	74.15	81.05	75.56	82.11
CorefRoBERTa$_{\text{LARGE}}$	**74.94**	**81.71**	**75.80**	**82.81**

Table 1: Results on QUOREF measured by exact match (EM) and F1. Results with *, + are from Dasigi et al. (2019) and official leaderboard respectively.

Implementation Details Following BERT's setting (Devlin et al., 2019), given the question $Q = (q_1, q_2, \ldots, q_m)$ and the passage $P = (p_1, p_2, \ldots, p_n)$, we represent them as a sequence $\boldsymbol{X} = ([\text{CLS}], q_1, q_2, \ldots, q_m, [\text{SEP}], p_1, p_2, \ldots, p_n, [\text{SEP}])$, feed the sequence $\boldsymbol{X}$ into the pre-trained encoder and train two classifiers on the top of it to seek answer's start and end positions simultaneously. For MRQA, CorefBERT maintains the same framework as BERT. For QUOREF, we further employ two extra components to process multiple mentions of the answers: (1) Spurred by the idea from MTMSN (Hu et al., 2019) in handling the problem of multiple answer spans, we utilize the representation of [CLS] to predict the number of answers. After that, we first selects the answer span of the current highest scores, then continues to choose that of the second-highest score with no overlap to previous spans, until reaching the predicted answer number. (2) When answering a question from QUOREF, the relevant mention could possibly be a pronoun, so we attach a reasoning Transformer layer for pronoun resolution before the span boundary classifier.

Model	SQuAD	NewsQA	TriviaQA	SearchQA	HotpotQA	NaturalQA	Average
$\text{BERT}_{\text{BASE}}$	88.4	66.9	68.8	78.5	74.2	75.6	75.4
$\text{CorefBERT}_{\text{BASE}}$	**89.0**	**69.5**	**70.7**	**79.6**	**76.3**	**77.7**	**77.1**
$\text{BERT}_{\text{LARGE}}$	91.0	69.7	73.1	81.2	77.7	79.1	78.6
$\text{CorefBERT}_{\text{LARGE}}$	**91.8**	**71.5**	**73.9**	**82.0**	**79.1**	**79.6**	**79.6**

Table 2: Performance (F1) on six MRQA extractive question answering benchmarks.

Results Table 1 shows the performance on QUOREF. Our adapted $\text{BERT}_{\text{BASE}}$ surpasses original BERT by about 2% in EM and F1 score, indicating the effectiveness of the added reasoning layer and multi-span prediction module. $\text{CorefBERT}_{\text{BASE}}$ and $\text{CorefBERT}_{\text{LARGE}}$ exceeds our adapted $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ by 4.4% and 2.9% F1 respectively. Leaderboard results are shown in the appendix. Based on the TASE framework (Efrat et al., 2020), the model with CorefRoBERTa achieves a new state-of-the-art with about 1% EM improvement compared to the model with RoBERTa. We also show four case studies in the appendix, which indicate that through reasoning over mentions, CorefBERT could aggregate information to answer the question requiring coreferential reasoning.

Table 2 further shows that the effectiveness of CorefBERT is consistent in six datasets of the MRQA shared task besides QUOREF. Though the MRQA shared task is not designed for coreferential reasoning, CorefBERT still achieves averagely over 1% F1 improvement on all of the six datasets, especially on NewsQA and HotpotQA. In NewsQA, 20.7% of the answers can only be inferred by synthesizing information distributed across multiple sentences. In HotpotQA, 63% of the answers need to be inferred through either bridge entities or checking multiple properties in different positions. It demonstrates that coreferential reasoning is an essential ability in question answering.

4.3 Relation Extraction

Relation extraction (RE) aims to extract the relationship between two entities in a given text. We evaluate our model on DocRED (Yao et al., 2019), a challenging document-level RE dataset which requires the model to extract relations between entities by synthesizing information from all the mentions of them after reading the whole document. DocRED requires a variety of reasoning types, where 17.6% of the relational facts need to be uncovered through coreferential reasoning.

Model	Dev		Test	
	IgnF1	F1	IgnF1	F1
CNN*	41.58	43.45	40.33	42.26
LSTM*	48.44	50.68	47.71	50.07
BiLSTM*	48.87	50.94	50.26	51.06
ContextAware*	48.94	51.09	48.40	50.70
$\text{BERT-TS}_{\text{BASE}}$ +	-	54.42	-	53.92
$\text{HINBERT}_{\text{BASE}}$ #	54.29	56.31	53.70	55.60
$\text{BERT}_{\text{BASE}}$	54.63	56.77	53.93	56.27
$\text{CorefBERT}_{\text{BASE}}$	**55.32**	**57.51**	**54.54**	**56.96**
$\text{BERT}_{\text{LARGE}}$	56.51	58.70	56.01	58.31
$\text{CorefBERT}_{\text{LARGE}}$	**56.82**	**59.01**	**56.40**	**58.83**
$\text{RoBERTa}_{\text{LARGE}}$	57.19	59.40	57.74	60.06
$\text{CorefRoBERTa}_{\text{LARGE}}$	**57.35**	**59.43**	**57.90**	**60.25**

Table 3: Results on DocRED measured by micro ignore F1 (IgnF1) and micro F1. IgnF1 metrics ignores the relational facts shared by the training and dev/test sets. Results with *, +, # are from Yao et al. (2019), Wang et al. (2019a), and Tang et al. (2020) respectively.

Baselines We compare our model with the following baselines for document-level relation extraction: (1) **CNN / LSTM / BiLSTM / BERT**. CNN (Zeng et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997), bidirectional LSTM (BiLSTM) (Cai et al., 2016), BERT (Devlin et al., 2019) are widely adopted as text encoders in relation extraction tasks. With these encoders, Yao et al. (2019) generates representations of entities for further predicting of the relationships between entities. (2) **ContextAware** (Sorokin and Gurevych, 2017) takes relations' interaction into account, which demonstrates that other relations in the context are beneficial for target relation prediction. (3) **BERT-TS** (Wang et al., 2019a) applies a two-step prediction to deal with the large number of irrelevant entities, which first predicts whether two entities have a relationship and then predicts the specific relation. (4) **HinBERT** (Tang et al., 2020) proposes a hierarchical inference network to aggregate the inference information with different granularity.

Results Table 3 shows the performance on DocRED. The $\text{BERT}_{\text{BASE}}$ model we implemented with mean-pooling entity representation and hyperparameter tuning[7] performed better than previous RE models with $\text{BERT}_{\text{BASE}}$ size, which provides a stronger baseline. $\text{CorefBERT}_{\text{BASE}}$ outperforms $\text{BERT}_{\text{BASE}}$ model by 0.7% F1. $\text{CorefBERT}_{\text{LARGE}}$ beats $\text{BERT}_{\text{LARGE}}$ by 0.5% F1. We also show a case study in the appendix, which further proves that considering coreference information of text is helpful for exacting relational facts from documents.

4.4 Fact Extraction and Verification

Fact extraction and verification aim to verify deliberately fabricated claims with trust-worthy corpora. We evaluate our model on a large-scale public fact verification dataset FEVER (Thorne et al., 2018). FEVER consists of $185,455$ annotated claims with all Wikipedia documents.

Baselines We compare our model with four BERT-based fact verification models: (1) **BERT Concat** (Zhou et al., 2019) concatenates all of the evidence pieces and the claim to predict the claim label; (2) **SR-MRS** (Nie et al., 2019) employs hierarchical BERT retrieval to improve the performance; (3) **GEAR** (Zhou et al., 2019) constructs an evidence graph and conducts a graph attention network for jointly reasoning over several evidence pieces; (4) **KGAT** (Liu et al., 2020b) conducts a fine-grained graph attention network with kernels.

Results Table 4 shows the performance on FEVER. KGAT with $\text{CorefBERT}_{\text{BASE}}$ outperforms KGAT with $\text{BERT}_{\text{BASE}}$ by 0.4% FEVER score. KGAT with $\text{CorefRoBERTa}_{\text{LARGE}}$ gains 1.9% FEVER score improvement compared to the model with $\text{RoBERTa}_{\text{LARGE}}$, and arrives at a new state-of-the-art on FEVER benchmark. It again demonstrates the effectiveness of our model. CorefBERT, which incorporates coreference information in distant-supervised pre-training, contributes to verify if the claim and evidence discuss about the same mentions, such as a person or an object.

4.5 Coreference Resolution

Coreference resolution aims to link referring expressions that evoke the same discourse entity. We examine models' coreference resolution ability under the setting that all mentions have been detected. We evaluate models on several widely-used datasets, including GAP (Webster et al., 2018), DPR (Rahman and Ng, 2012), WSC (Levesque, 2011), Winogender (Rudinger et al., 2018) and PDP (Davis et al., 2017). These datasets provide two sentences where the former has two or more mentions and the latter contains an ambiguous pronoun. It is required that the ambiguous pronoun should be connected to the right mention.

[7]Details are in the appendix due to space limit.

Model	LA	FEVER
BERT Concat*	71.01	65.64
GEAR*	71.60	67.10
SR-MRS+	72.56	67.26
KGAT ($\text{BERT}_{\text{BASE}}$) #	72.81	69.40
KGAT ($\text{CorefBERT}_{\text{BASE}}$)	**72.88**	**69.82**
KGAT ($\text{BERT}_{\text{LARGE}}$) #	73.61	70.24
KGAT ($\text{CorefBERT}_{\text{LARGE}}$)	**74.37**	**70.86**
KGAT ($\text{RoBERTa}_{\text{LARGE}}$) #	74.07	70.38
KGAT ($\text{CorefRoBERTa}_{\text{Large}}$)	**75.96**	**72.30**

Table 4: Results on FEVER test set measured by label accuracy (LA) and FEVER. The FEVER score evaluates the model performance and considers whether the golden evidence is provided. Results with *, +, # are from Zhou et al. (2019), Nie et al. (2019) and Liu et al. (2020b) respectively.

Model	GAP	DPR	WSC	WG	PDP
$\text{BERT-LM}_{\text{BASE}}$	75.3	75.4	61.2	68.3	76.7
$\text{CorefBERT}_{\text{BASE}}$	**75.7**	**76.4**	**64.1**	**70.8**	**80.0**
$\text{BERT-LM}_{\text{LARGE}}$ *	76.0	80.1	70.0	78.8	81.7
$\text{WikiCREM}_{\text{LARGE}}$ *	**78.0**	84.8	70.0	76.7	86.7
$\text{CorefBERT}_{\text{LARGE}}$	76.8	**85.1**	**71.4**	**80.8**	**90.0**
$\text{RoBERTa-LM}_{\text{LARGE}}$	**77.8**	90.6	**83.2**	77.1	93.3
$\text{CorefRoBERTa}_{\text{LARGE}}$	**77.8**	**92.2**	**83.2**	**77.9**	**95.0**

Table 5: Results on coreference resolution test sets. Performance on GAP is measured by F1, while scores on the others are given in accuracy. WG: Winogender. Results with * are from Kocijan et al. (2019).

Baselines We compare our model with two coreference resolution models: (1) **BERT-LM** (Trinh and Le, 2018) substitutes the pronoun with `[MASK]` and uses language model to compute the probability of recovering the mention candidates; (2) **WikiCREM** (Kocijan et al., 2019) generates GAP-like sentences automatically and trains BERT by minimizing the perplexity of correct mentions on these sentences. Finally, the model is fine-tuned on supervised datasets. Benefiting from the augmented data, WikiCREM achieves state-of-the-art in sentence-level coreference resolution. For BERT-LM and CorefBERT, we adopt the same data split and the same training method on supervised datasets as those of WikiCREM in order to make a fair comparison.

Model	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{CorefBERT}_{\text{BASE}}$	84.2/83.5	71.3	90.5	93.7	51.5	85.8	89.1	67.2	79.6
$\text{BERT}_{\text{LARGE}}$	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	81.9
$\text{CorefBERT}_{\text{LARGE}}$	86.9/85.7	71.7	92.9	94.7	62.0	86.3	89.3	70.0	82.2

Table 6: Test set performance metrics on GLUE benchmarks. Matched/mistached accuracies are reported for MNLI; F1 scores are reported for QQP and MRPC, Spearmanr correlation is reported for STS-B; Accuracy scores are reported for the other tasks.

Model	QUOREF	SQuAD	NewsQA	TriviaQA	SearchQA	HotpotQA	NaturalQA	DocRED
$\text{BERT}_{\text{BASE}}$	67.3	88.4	66.9	68.8	78.5	74.2	75.6	56.8
-NSP	70.6	88.7	67.5	68.9	79.4	75.2	75.4	56.7
-NSP, +WWM	70.1	88.3	69.2	70.5	**79.7**	75.5	75.2	57.1
-NSP, +MRM	70.0	88.5	69.2	70.2	78.6	75.8	74.8	57.1
$\text{CorefBERT}_{\text{BASE}}$	**72.3**	**89.0**	**69.5**	**70.7**	79.6	**76.3**	**77.7**	**57.5**

Table 7: Ablation study. Results are F1 scores on development set for QUOREF and DocRED, and on test set for others. $\text{CorefBERT}_{\text{BASE}}$ combines "-NSP, +MRM" scheme and copy-based training objective.

Results Table 5 shows the performance on the test set of the above coreference resolution dataset. Our CorefBERT model significantly outperforms BERT-LM, which demonstrates that the intrinsic coreference resolution ability of CorefBERT has been enhanced by involving the mention reference prediction training task. Moreover, it achieves comparable performance with state-of-the-art baseline WikiCREM. Note that, WikiCREM is specially designed for sentence-level coreference resolution and is not suitable for other NLP tasks. On the contrary, the coreferential reasoning capability of CorefBERT can be transferred to other NLP tasks.

4.6 GLUE

The Generalized Language Understanding Evaluation dataset (GLUE) (Wang et al., 2018) is designed to evaluate and analyze the performance of models across a diverse range of existing natural language understanding tasks. We evaluate CorefBERT on the main GLUE benchmark used in BERT.

Implementation Details Following BERT's setting, we add [CLS] token in front of the input sentences, and extract its representation on the top layer as the whole sentence or sentence pair's representation for classification or regression.

Results Table 6 shows the performance on GLUE. We notice that CorefBERT achieves comparable results to BERT. Though GLUE does not require much coreference resolution ability due to its attributes, the results prove that our masking strategy and auxiliary training objective would not weaken the performance on generalized language understanding tasks.

5 Ablation Study

In this section, we explore the effects of the Whole Word Masking (WWM), Mention Reference Masking (MRM), Next Sentence Prediction (NSP) and copy-based training objective using several benchmark datasets. We continue to train Google's released $\text{BERT}_{\text{BASE}}$ on the same Wikipedia corpus with different strategies. As shown in Table 7, we have the following observations: (1) Deleting NSP training task triggers a better performance on almost all tasks. The conclusion is consistent with that of RoBERTa (Liu et al., 2019); (2) MRM scheme usually achieves parity with WWM scheme except on SearchQA, and both of them outperform the original subword masking scheme especially on NewsQA (averagely +1.7% F1) and TriviaQA (averagely +1.5% F1); (3) On the basis of MRM scheme, our copy-based training objective explicitly requires model to look for mention's referents in the context, which could adequately consider the coreference information of the sequence. CorefBERT takes advantage of the objective and further improves the performance, with a substantial gain (+2.3% F1) on QUOREF.

6 Conclusion and Future Work

In this paper, we present a language representation model named CorefBERT, which is trained on a novel task, Mention Reference Prediction (MRP), for strengthening the coreferential reasoning ability of BERT. Experimental results on several downstream NLP tasks show that our CorefBERT significantly outperforms BERT by considering the coreference information within the text and even improve the performance of the strong RoBERTa model. In the future, there are several prospective research directions: (1) We introduce a distant supervision (DS) assumption in our MRP training task. However, the automatic labeling mechanism inevitably accompanies with the wrong labeling problem and it is still an open problem to mitigate the noise. (2) The DS assumption does not consider pronouns in the text, while pronouns play an important role in coreferential reasoning. Hence, it is worth developing a novel strategy such as self-supervised learning to further consider the pronoun.

Acknowledgement

This work is supported by the National Key R&D Program of China (2020AAA0105200), Beijing Academy of Artificial Intelligence (BAAI) and the NExT++ project from the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative.

References

Cosmin Adrian Bejan, Matthew Titsworth, Andrew Hickl, and Sanda M. Harabagiu. 2009. Nonparametric bayesian models for unsupervised event coreference resolution. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 73–81.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Ziqiang Cao, Chuwei Luo, Wenjie Li, and Sujian Li. 2017. Joint copying and restricted generation for paraphrase. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3152–3158.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 1–14.

Pengxiang Cheng and Katrin Erk. 2020. Attending to entities for better text understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7554–7561. AAAI Press.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 845–855.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5924–5931. Association for Computational Linguistics.

Ernest Davis, Leora Morgenstern, and Charles L. Ortiz Jr. 2017. The first winograd schema challenge at IJCAI-16. *AI Magazine*, 38(3):97–98.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases.

In *Proceedings of the Third International Workshop on Paraphrasing, IWP@IJCNLP 2005, Jeju Island, Korea, October 2005, 2005*.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

Avia Efrat, Elad Segal, and Mor Shoham. 2020. A simple and effective model for answering multi-span questions. *CoRR*, abs/1909.13375.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. pages 1–13.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL@ACL 2007 Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic, June 28-29, 2007*, pages 1–9.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 364–369.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 328–339.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. pages 1596–1606.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5802–5807. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Vid Kocijan, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. 2019. Wikicrem: A large unsupervised corpus for coreference resolution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4302–4311. Association for Computational Linguistics.

Daniel Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2779–2795. Association for Computational Linguistics.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *TACL*, 7:452–466.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 188–197.

Hector J. Levesque. 2011. The winograd schema challenge. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020a. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

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.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020b. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7342–7351. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 13–23.

Xuezhe Ma, Zhengzhong Liu, and Eduard H. Hovy. 2016. Unsupervised ranking model for entity coreference resolution. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 1012–1018.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2553–2566. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 777–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392.

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.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 5926–5936.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1784–1789.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019a. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7463–7472. IEEE.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019b. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 380–385.

Swabha Swayamdipta, Ankur P. Parikh, and Tom Kwiatkowski. 2018. Multi-mention learning for reading comprehension with neural cascades. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Alon Talmor and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4911–4921.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. HIN: hierarchical inference network for document-level relation extraction. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part I*, volume 12084 of *Lecture Notes in Computer Science*, pages 197–209. Springer.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819.

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

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and

Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 191–200.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019a. Fine-tune bert for docred with two-step process. *CoRR*, abs/1909.11898.

Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. 2019b. Denoising based sequence-to-sequence pre-training for text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4001–4013. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *TACL*, 7:625–641.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *TACL*, 6:605–617.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5021–5031. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 5754–5764.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2335–2344.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware BERT for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9628–9635. AAAI Press.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020.

Transformer-xh: Multi-evidence reasoning with extra hop attention. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 892–901.

Appendices

A Masked Language Modeling (MLM)

MLM is regarded as a kind of cloze tasks and aims to predict the missing tokens according to its contextual representation. In our work, 15% of the tokens in input sequence are sampled as the missing tokens. Among them, 80% are replaced with a special token `[MASK]`, 10% are replaced with random tokens and 10% are unchanged. The task aims to predict original tokens from corrupted input.

B Leaderboard Results on QUOREF

TASE (Efrat et al., 2020) converts the multi-span prediction problem as a sequence tagging problem, which substantially improves the model's ability in terms of handling multi-span answer. Though the study of TASE and our CorefBERT are conducted in the same period, we still run TASE with CorefRoBERTa encoder. As Table 8 shows, the performance of TASE with CorefRoBERTa encoder gains about 1% EM improvement compared to that with RoBERTa encoder, which demonstrates the effectiveness of CorefBERT for different question answering frameworks.

Model	EM	F1
XLNet (Dasigi et al., 2019)	61.88	71.51
RoBERTa-MT	72.61	80.68
CorefRoBERTa$_{\text{LARGE}}$	75.80	82.81
TASE (RoBERTa) (Efrat et al., 2020)	79.66	86.13
TASE (CorefRoBERTa)	**80.61**	**86.70**

Table 8: Leaderboard results on QUOREF test set.

C Case Study on QUOREF

Table 9 shows examples from QUOREF (Dasigi et al., 2019). For the first example, it is essential to obtain the fact that the asthmatic boy in question refers to Barry. After that, we should synthesize information from two Mr. Lee's mentions: (1) Mr. Lee trains Barray; (2) Mr. Lee is the uncle of

(1) Q: Whose uncle trains the asthmatic boy? Paragraph: [1] `Barry Gabrewski` is an asthmatic boy ... [2] `Barry` wants to learn the martial arts, but is rejected by the arrogant dojo owner Kelly Stone for being too weak. [3] Instead, `he` is taken on as a student by an old Chinese man called ***Mr. Lee***, **Noreen**'s sly uncle. [4] ***Mr. Lee*** finds creative ways to teach Barry to defend himself from his bullies.
(2) Q: Which composer produced String Quartet No. 2? Paragraph: [1] **Tippett**'s Fantasia on a Theme of Handel for piano and orchestra was performed at the Wigmore Hall in March 1942, with ***Sellick*** again the soloist, and the same venue saw the premiere of `the composer`'s String Quartet No. 2 a year later. ... [2] In 1942, Schott Music began to publish **Tippett**'s works, establishing an association that continued until the end of the `the composer`'s life.
(3) Q: What is the first name of the person who lost her beloved husband only six months earlier? Pargraph: [1] Robert and ***Cathy*** Wilson are a timid married couple in 1940 London. ... [2] Robert toughens up on sea duty and in time becomes a petty officer. [3] His hands are badly burned when his ship is sunk, but he stoically rows in the lifeboat for five days without complaint. [4] He recuperates in a hospital, tended by **Elena**, a beautiful nurse. [5] He is attracted to `her`, but `she` informs him that `she` lost her beloved husband only six months earlier, kisses him, and leaves.
(4) Q: Who would have been able to win the tournament with one more round? Paragraph: [1] At a jousting tournament in 14th-century Europe, young squires ***William*** Thatcher, Roland, and Wat discover that their master, Sir **Ector**, has died. [2] If `he` had completed one final pass `he` would have won the tournament. [3] Destitute, ***William*** wears **Ector**'s armour to impersonate him, winning the tournament and taking the prize.

Table 9: Examples from QUOREEF (Dasigi et al., 2019) that were correctly predicted by CorefBERT$_{\text{BASE}}$, but wrongly predicted by BERT$_{\text{BASE}}$. ***Answers from BERT$_{\text{BASE}}$***, **Answers from CorefBERT$_{\text{BASE}}$**, and `Clue` are colored respectively.

Eclipse (Meyer novel)
[1] ***Eclipse*** is the third novel in the **Twilight Saga** by **Stephenie Meyer**. It continues the story of Bella Swan and her vampire love, ***Edward Cullen***. [2] The novel explores Bella's compromise between her love for ***Edward*** and her friendship with shape-shifter ***Jacob Black***, ... [3] ***Eclipse*** is preceded by ***New Moon*** and followed by ***Breaking Dawn***. [4] The book was released on **August 7, 2007**, with an initial print run of one million copies, and sold more than 150,000 copies in the first 24 hours alone.
Subject: ***New Moon*** / ***Breaking Dawn*** **Object**: **Twilight Saga** **Relation**: `Part of the series`
Subject: ***Edward Cullen*** / ***Jacob Black*** **Object**: **Stephenie Meyer** **Relation**: `Creator`
Subject: ***Eclipse*** **Object**: **August 7, 2007** **Relation**: `Publication date`

Table 10: An example from DocRED (Yao et al., 2019). We show some relational facts detected by CorefBERT$_{\mathrm{BASE}}$ but missed by BERT$_{\mathrm{BASE}}$.

Noreen. Reasoning over the above information, we could know that Noreen's uncle trains the asthmatic boy. For the second example, it needs to infer that Tippett is a composer from the second sentence for obtaining the final answer from the first sentence. After training on the mention reference prediction task, CorefBERT has become capable of reasoning over these mentions, summarizing messages from mentions in different positions, and finally figuring out the correct answer. For the third and fourth examples, it is necessary to know *she* refers to Elena, and *he* refers to Ector by respective coreference resolution. Benefiting from a large amount of distant-supervised coreference resolution training data, CorefBERT successfully finds out the reference relationship and provides accurate answers.

D Case Study on DocRED

Table 10 shows an example from DocRED (Yao et al., 2019). We show some relational facts detected by CorefBERT$_{\mathrm{BASE}}$ but missed by BERT$_{\mathrm{BASE}}$. For the first relational fact, it is necessary to connect the first and the third sentences through the coreference of Eclipse for acquiring the fact that New Moon and Breaking Dawn are also the novel in the Twilight Saga. For the second and the third relational fact, the referring expressions *it*, *the novel*, and *the book* should be linked to Eclipse correctly to increase model's confidence to find out all the characters and the publication date of the novel from the context. CorefBERT considers coreference information of text, which helps to discover relation facts beyond sentence boundary.

Claim: ***Bob Ross*** created `ABC` drama **The Joy of Painting**.
[1] **[Bob Ross]** ***Robert Norman Ross*** was an American painter and television host. [2] **[Bob Ross]** ***He*** was the creator and host of **The Joy of Painting**, an instructional television program that aired from 1983 to 1994 on `PBS` in the United States, and also aired in Canada, ... [3] **[Bob Ross]** **The Joy of Painting** is an American half hour instructional television show hosted by painter ***Bob Ross*** which ran from January 11, 1983, until May 17, 1994. [4] **[The Joy of Painting]** In each episode, ***Ross*** taught techniques for landscape oil painting, completing a painting in each session. [5] **[The Joy of Painting]** The program followed the same format as its predecessor, The Magic of Oil Painting , hosted by ***Ross***'s mentor.
Label: REFUTES

Table 11: An example from FEVER (Thorne et al., 2018). Five pieces of evidence from article **[Bob Ross]** and **[The Joy of Painting]** are retrieved by the retriever.

E Case Study on FEVER

Table 11 shows an example from FEVER (Thorne et al., 2018). The given claim is fabricated since the drama "The Joy of Painting" was aired on PBS instead of ABC. With the CorefBERT encoder, KGAT (Liu et al., 2020b) could propagate and aggregate the entity information from evidence for refuting the wrong claim more accurately.

F Task-Specific Model Details

All the models are implemented based on Huggingface transformers[8]. We train models on down-

[8]https://github.com/huggingface/transformers

stream tasks with Adam optimizer (Kingma and Ba, 2015).

F.1 Question Answering (QA)

For QA models, we uses a batch size of 32 instances with a maximum sequence length of 512.

We adopt the official data split for QUOREF (Dasigi et al., 2019), where train / development / test set contains 19399 / 2418 / 2537 instances respectively. And we submit our model to the test sever[9] for online evaluation. We conduct a grid search on the learning rate (lr) in $[1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}]$ and epoch number in $[2, 4, 6]$. The best $\text{BERT}_{\text{BASE}}$ configuration on development set used $lr = 2 \times 10^{-5}$, 6 epochs. We adopt this configuration for the $\text{BERT}_{\text{LARGE}}$ and $\text{RoBERTa}_{\text{LARGE}}$ models. We regard MRQA (Fisch et al., 2019) as a testbed to examine whether models can answer questions well across various data distributions. For fair comparison, we keep $lr = 3 \times 10^{-5}$, 2 epochs for all of the MRQA experiments.

For TASE (Efrat et al., 2020) with CorefRoBERTa encoder, we keep the same configuration[10] as that of the original paper, which used a batch size of 12, learning rate of 5×10^{-6}, 35 epochs.

F.2 Document-level Relation Extraction

We modify the official code[11] to implement BERT-based models for DocRED (Yao et al., 2019). In our implementation, the representation of a mention, which consists of several words, is the average of representations of those words. Furthermore, the representation of an entity is defined as the mean of all mentions referring to it. Finally, two entities' representations are fed to a bi-linear layer to predict relations between them.

We use the official data split for DocRED, where train / development / test set consists of 3053 / 1000 / 1000 documents respectively. We adopt batch size of 32 instances with maximum sequence length of 512 and conduct a grid search on the learning rate in $[2 \times 10^{-5}, 3 \times 10^{-5}, 4 \times 10^{-5}, 5 \times 10^{-5}]$ and number epochs in $[100, 150, 200]$. We find the configuration used learning rate of 4×10^{-5}, 200 epochs is best for both the base and the large model. We evaluate models on development set every 5 epochs and save the checkpoint with the highest F1 score. After that, the test results of the best model are submitted to the evaluation server[12].

F.3 Fact Extraction and Verification

We apply the released code[13] of KGAT (Liu et al., 2020b) for evaluating CorefBERT. We use the official data split for FEVER (Thorne et al., 2018), where train / development / test set contains 145449 / 19998 / 19998 claims respectively. We adopt a batch size of 32, maximum length of 512 tokens and search the learning rate in $[2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}]$. We achieved the best performance with learning rate of 5×10^{-5} for the base model and 2×10^{-5} for the large model. All models are trained with a batch size of 32 instances for 3 epochs and evaluated on development set every 1000 steps. After that, we submit test results of our best model to evaluation server[14].

F.4 Coreference Resolution

We use the released code[15] of WikiCREM (Kocijan et al., 2019) for fine-tuning BERT-LM (Trinh and Le, 2018) and CorefBERT on supervised datasets. For a sentence S, which possesses a correct candidate $\mathbf{a}$ and an incorrect candidate $\mathbf{b}$, the loss consists of two parts: (1) the negative log-likelihood of the correct candidate; (2) a max-margin between the log-likelihood of the correct candidate and the incorrect candidate:

$$\begin{aligned} \mathcal{L} = &- \log \Pr(\mathbf{a}|S) \\ &+ \alpha \max\left(0, \log \Pr(\mathbf{b}|S) - \log \Pr(\mathbf{a}|S) + \beta\right), \end{aligned} \tag{5}$$

where α, β are hyperparameters. We follow the data split and fine-tuning setting of WikiCREM, which adopts a batch size of 64, a maximum sequence length of 128 and 10 epochs training. We search the learning rate $lr \in [3 \times 10^{-5}, 1 \times 10^{-5}, 5 \times 10^{-6}, 3 \times 10^{-6}]$, hyperparameters $\alpha \in [5, 10, 20]$, $\beta \in [0.1, 0.2, 0.4]$. The best performance of models with base size and $\text{CorefBERT}_{\text{LARGE}}$ on validation set were achieved with $lr = 3 \times 10^{-5}$, $\alpha = 10$, $\beta = 0.2$. We keep this configuration for the RoBERTa-based models.

[9]https://leaderboard.allenai.org/quoref/submissions/public

[10]https://github.com/eladsegal/tag-based-multi-span-extraction

[11]https://github.com/thunlp/DocRED

[12]https://competitions.codalab.org/competitions/20717

[13]https://github.com/thunlp/KernelGAT

[14]https://competitions.codalab.org/competitions/18814

[15]https://github.com/vid-koci/bert-commonsense

Model	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
$\text{CorefBERT}_{\text{BASE}}$	2×10^{-5}	4×10^{-5}	3×10^{-5}	3×10^{-5}	5×10^{-5}	4×10^{-5}	5×10^{-5}	4×10^{-5}
$\text{CorefBERT}_{\text{LARGE}}$	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	3×10^{-5}	5×10^{-5}	5×10^{-5}	3×10^{-5}

Table 12: Learning rate for CorefBERT on GLUE benchmarks.

Model	Parameters	Layers	Hidden	Embedding	Vocabulary
$\text{CorefBERT}_{\text{BASE}}$	110M	12	768	768	28,996
$\text{CorefBERT}_{\text{LARGE}}$	340M	24	1,024	1,024	28,996
$\text{CorefRoBERTa}_{\text{LARGE}}$	355M	24	1,024	1,024	50,265

Table 13: Parameter number and the configuration of CorefBERT.

Model	QUOREF	MRQA	DocRED	FEVER	GLUE	Coref.
$\text{CorefBERT}_{\text{BASE}}$	13.23	13.15	117.37	18.88	2.95	4.27
$\text{CorefBERT}_{\text{LARGE}}$	43.40	43.37	180.65	54.03	9.22	10.90

Table 14: Average inference runtime per example for CorefBERTs on different benchmarks. Inference is done on a RTX 2080ti GPU with a batch of 32 instances and inference time is measured in milliseconds. The input sequence length is 512 for QUOREF, MRQA, DocRED, FEVER, and 128 for others. Coref.: Coreference resolution.

F.5 Generalized Language Understanding (GLUE)

We evaluate CorefBERT on the main GLUE benchmark (Wang et al., 2018) used in BERT, including MNLI (Williams et al., 2018), QQP[16], QNLI (Rajpurkar et al., 2016), SST-2 (Socher et al., 2013), CoLA (Warstadt et al., 2019) , STS-B (Cer et al., 2017), MRPC (Dolan and Brockett, 2005) and RTE (Giampiccolo et al., 2007).

We use a batch size of 32, maximum sequence length of 128, fine-tune models for 3 epochs for all GLUE tasks and select the learning rate of Adam among $[2\times10^{-5}, 3\times10^{-5}, 4\times10^{-5}, 5\times10^{-5}]$ for the best performance on the development set. After that, we submit the result of our best model to the official evaluation server[17]. Table 12 shows the best learning rate for $\text{CorefBERT}_{\text{BASE}}$ and $\text{CorefBERT}_{\text{LARGE}}$.

F.6 Number of Parameters and Average Runtime

CorefBERT's architecture is a multi-layer bidirectional Transformer (Vaswani et al., 2017). Tables 13 shows the parameter number of CorefBERTs with different model size. Compared to BERT (Devlin et al., 2019), CorefBERT add a few parameters for computing the copy-based objective. Hence, CorefBERT keeps similar number of parameters as BERT with the same size.

Table 14 shows the task-specific average inference runtime per example for CorefBERT. The inferenece is done on a RTX 2080ti GPU with a batch of 32 instances. The inference time includes time on CPU and time on GPU. $\text{CorefRoBERTa}_{\text{LARGE}}$ consumes a similar time as $\text{CorefBERT}_{\text{LARGE}}$ since they both use a 24-layer Transformer architecture.

F.7 Resolving the Coreference in the Corpus

In our preliminary experiment, we resolve the coreference of training corpus via the StanfordNLP tool[18] and apply our copy-based objective on this training corpus. We find the obtained model performs better than the BERT model without NSP but worse than the current CorefBERT. We think that considering coreference such as pronoun in pre-training could also enhance model's coreferential reasoning ability, while how to deal with the noise from coreference tools remains a problem to be explored.

[16] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

[17] https://gluebenchmark.com

[18] https://stanfordnlp.github.io/CoreNLP