ChapterCR: A Large-Scale Chapter-Level Coreference Resolution Benchmark

Anonymous ACL submission

Abstract

Coreference Resolution aims to identify mentions that refer to one another in documents. Existing coreference resolution datasets are either small in size or short in coreference chains. To address the issue, we propose ChapterCR, a large-scale chapter-level coreference resolution dataset. In ChapterCR, the coreference chains are longer and there are more distractors between the mention and the right entity, which makes it more challenging. Experiments on ChapterCR show that there is still a large gap between the state-of-art baselines and human beings. Even ChatGPT does not perform very well in ChapterCR, with the F1 score of 74.0% in ChapterCR-en and 58.8% in ChapterCR-zh, showing that ChapterCR is still an open problem.

1 Introduction

Coreference resolution (CR) aims to link textual mentions and the entities they refer to in documents. For instance, given the sentence *Recently, Apple sued Qualcomm, suing it for failing to cooperate in accordance with contracts*, CR needs to distinguish that *it* here refers to Qualcomm instead of Apple. CR plays an important role in evaluating the commonsense reasoning ability of large language models (Zhou et al., 2019), and is essential for many downstream tasks such as machine reading comprehension (Wu et al., 2020), information extraction (Zelenko et al., 2004), and multi-round dialogue system (Yu et al., 2022).

Existing datasets for CR have deficiencies in the following aspects: the small scale of data and the short and easy-resolved coreference chains. ACE2004 (Doddington et al., 2004) consists of only 451 documents and 158k words. STM-coref (Brack et al., 2021) and LongtoNotes (Shridhar et al., 2022) contain 110 and 2415 documents with less than 1000 and 6000 coreference annotations. LitBank (Bamman et al., 2020), MUC-6 (muc, 1995), MUC-7 (Hirschman, 1997) and WikiCoref (Ghaddar and Langlais, 2016) are even smaller, with only 100, 60, 50, and 30 documents respectively. All of the above five CR datasets are quite limited in data scale and can not fairly evaluate modern neural networks. WSC (Levesque et al., 2012) and GAP (Webster et al., 2018a) annotate coreference resolution within twin sentences, and the length of most coreference chains in CoNLL2012 (Weischedel et al., 2011) does not exceed 5. Short coreference chains in the three datasets lead to fewer distractors between mentions and entities, making them not challenging enough to test the limits of current CR models.

In the paper, we present ChapterCR to develop a large-scale CR dataset in longer texts to accelerate the research of coreference resolution. Figure 1 illustrates an example of ChapterCR. ChapterCR aims to resolve coreference chains across entire chapters of a novel. For example, given the entity *Quila* (highlighted in green), ChapterCR needs to find all references *the visitor, she and the man'sister* in Chapter 1 that refer to *Quila*.

We highlight the following three contributions of ChapterCR: (1) Large-scale. ChapterCR contains a total of 29k chapters with 55k coreference chains, far exceeding the scale of existing CR datasets. The large scale and high quality allow ChapterCR to fairly evaluate modern neural network models. (2) Long Coreference Chain. ChapterCR detects coreferences at the chapter level. Compared with previous datasets that detect coreferences at the sentence level or cross-sentence level, the length of the coreference chain in ChapterCR is longer, with an average length of 8.1 (see Table 1 for detail), which poses a greater challenge to the semantic understanding ability of existing CR models. (3) Bilingual Language. ChapterCR annotates both English novels (ChapterCR-en) and Chinese novels (ChapterCR-zh), which can promote the development of coreference resolution in the two languages.

Jerebai Quila Quil

Chaper 1

Hearing the voice of the visitor, the lady on the ground finally moved. Her cracked lips quivered, asking,"Quila, how's Quil?
Perhaps it was because she hadn't spoken for such a long time, but her voice sounded extremely hoarse, like the grinding of gravel on the floor.
Qulla frowned, with ever-growing abhorrence in her eyes. "Haaa--? My brother?" She hooked her lips into a smile full of ridicule and derision, "Jerebai are you still expecting him to come and save you? Do you know what day it is today? Today is the day that he marries my new sister-in-law! He is in love - do you really expect that you, a murderous demoness would even cross his mind?!" The man's sister cried.
He actually...
Jerebai heart felt as though it had been stabbed by a needle - and it wasn't an acute unbearable type of pain, but the type of pain that reverberates and lingers, even eking out traces of blood ever so slowly.
She should have known. After all, that person had not come to save her after such a long time...
Jerebai unconsciously held her abdomen. She once carried a child belonging to her and that man.

Figure 1: An example of ChapterCR. Mentions referring to the same entity are labeled in the same color. The coreference chain in ChapterCR is very long: 15 for entity Jerebai (highlighted in yellow), 8 for entity Quil (highlighted in blue), and 5 for entity Quila (highlighted in green), which makes ChapterCR more challenging.

于修逸很想知道什么意思，可秦亦封却处理起文件什么都不说了，气得他直跳脚。

算了算了，要是不想说，谁也拿他没办法，于修逸长叹了口气，觉得这一刻的秦亦封很可怕。任何人在他面前都只是蝼蚁，这次那个叫白净的恐怕要倒大霉了

Entity: 秦亦封
Normal Pronoun:他
Zero Pronoun:

Figure 2: An example of zero pronouns in Chinese.

In addition, as shown in Figure 2, we introduce zero pronoun resolution in ChapterCR-zh to further increase the difficulty of the proposed dataset.

We implement 8 state-of-the-art baselines along with the human evaluation to assess ChapterCR. Various experiments show that there is still a large gap between the SOTA baselines and human beings, showing the difficulty of ChapterCR.

2 Related Work

In recent years, coreference resolution has attracted widespread interest (Elango, 2005; Sukthanker et al., 2020; Lata et al., 2022; Liu et al., 2023), and a number of high-quality datasets and superior models have been proposed to promote the development of the field of coreference resolution.

2.1 Coreference Resolution Datasets

Muc-6 (muc, 1995) and MUC-7 (Hirschman, 1997) are the first two coreference resolution datasets, which contain only 60 and 50 documents with 30k and 25k words, which is too few to train a modern neural network model. After that, ACE2004 (Doddington et al., 2004) is developed by the Linguistic Data Consortium (LDC), which is annotated from a variety of sources including newswire, broadcast programming and weblogs, with only 451 documents and 158k words. CoNLL2012 (Weischedel et al., 2011) is annotated based on the Ontonotes corpus, a commonly used dataset in coreference resolution. CoNLL2012 has three languages, including English, Chinese and Arabic. CoNLL2012-en and CoNLL2012-zh contain only 3493 and 2280 documents with 12811 and 6727 coreferences. WikiCoref (Ghaddar and Langlais, 2016) is labeled from English wiki articles, containing only 7955 mentions in 30 documents. GUM (Zeldes, 2017) and ARRAU (Uryupina et al., 2016) are two anaphora-resolution datasets with less than 300 documents.

MASKEDWIKI (Kocijan et al., 2019b) and WikiCREM (Kocijan et al., 2019a) are relatively large datasets, but they are generated by unsupervised methods (replacing masked nouns with a pronoun in Wikipedia), rather than crowdsourced labeling, which cannot guarantee the quality of the data.

There are also domain-specific coreference res-

Datasets	#Doc.	#Sent.	#Tok.	#Mention	#Coref.	#ChainLen.
ACE2004	451	18530	158k	22550	-	-
MUC-6	60	3750	30k	-	-	-
WikiCoref+	30	2292	60k	7955	1255	6.34
WSC+	-	803	20k	2409	803	2
GAP+	-	8908	317k	26724	8908	2
STM-coref+	110	1480	26k	2577	908	2.84
CoNLL2012+	3493	112941	1.6M	56371	12811	4.4
LongtoNotes	2415	112941	1.6M	38640	5925	6.5
LitBank	100	108K	13M	57514	29103	1.98
ChapterCR-en(ours)	10235	53k	7.2M	136k	17k	8.1

Table 1: Statistics of coreference resolution datasets in English. Doc.: the number of documents, Sent.: the number of sentences, Entity: the number of entities, Mention: the number of mentions, Coref.: the number of coreferences, ChainLen.: the average length of the coreference chains

Datasets	#Doc.	#Sent.	#Tok.	#Mention	#Coref.	#ChainLen.
ACE2004	646	14233	154K	28135	-	-
CoNLL2012 +	2280	83763	950k	15136	6727	2.25
CLUEWSC2020 +	-	1648	276K	4944	1648	2
ChapterCR-zh(ours)	19288	81k	21M	310k	38k	8.17

Table 2: Statistics of coreference resolution datasets in Chinese.

olution datasets, such as MEDSTRACT (Pustejovsky et al., 2002), DrugNerAR (Segura-Bedmar et al., 2010), BioNLP-ST COREF (Nguyen et al., 2011) and CRAFT-CR (Cohen et al., 2017). These datasets are limited to a specific domain, and the coreference types are not rich enough.

Winograd Schema Challenge (WSC) (Levesque et al., 2012) is proposed by Hector Levesque in 2011 and named after Terry Winograd, professor of computer science at Stanford University, consisting of a total of 803 coreferences. WSCR (Rahman and Ng, 2012), PDP (Davis et al., 2017), WNLI (Wang et al., 2018), WINOBIAS (Zhao et al., 2018) and WinoGrande (Sakaguchi et al., 2021) are datasets derived from WSC. GAP (Webster et al., 2018a) is a gender-balanced dataset containing 8,908 coreferences of ambiguous pronouns and antecedent names, sampled from Wikipedia and released by Google AI Language. All of the above 7 datasets aim to resolve coreference within twin sentences, where there are few interference items between the mention and the entity, making these datasets less challenging. PreCo (Chen et al., 2018) proposes a larger dataset with 38k documents and 124M words, but it mainly involves preschool vocabulary and annotates massed singleton mentions, which reduces the difficulty of understanding the coreference chains.

In summary, previous coreference resolution datasets either suffer from small data size, low quality, limited domain or short and less challenging coreference chains. Therefore, we propose ChapterCR, a manually-annotated, large-scale coreference resolution dataset with longer coreference chains to make up for these deficiencies.

2.2 Coreference Resolution Models

There are four main kinds of coreference resolution models, including rule-based models, mention-pair models, mention-ranking models, and clustering-based models.

Rule-based models, such as Hobbs Algorithm (Hobbs, 1978), RAP (Lappin and Leass, 1994) and PRR (Lee et al., 2013), design syntactic constraints, gender agreement constraints, and grammar rules to resolve coreferences. Mention-pair models (Soon et al., 2001; Bengtson and Roth, 2008; Park et al., 2016) train a binary classifier that decides whether an active mention refers to a candidate antecedent. Mention-ranking models (Clark, 2015; Lee et al., 2017, 2018; Joshi et al., 2019a) employ feature systems, CNN, LSTM, and attention-based methods for mention pair score calculation and then choose the one with the highest score as the final answer. Clustering-based models (Cardie and Wagstaff, 1999; Yang et al., 2004; Clark and Manning, 2016; Zhang et al., 2018) start with a singleton cluster to each mention, and then in each step, it merges a pair of clusters if it predicts they are representing the same entity.

3 Data Construction

In this section, we illustrate the process of constructing ChapterCR. As shown in Figure 3, the process can be divided into three steps: chapter selection, entity & mention pre-annotation, and crowdsourced labeling. Chapter selection aims to screen high-quality chapters from online websites. Entity & mention pre-annotation aims to identify possible entities and references. Crowdsourced labeling aims to determine pairwise coreference between entities and mentions.

3.1 Chapter Selection

We choose novels as the data source, which have a more coherent narrative and are more likely to have long coreference chains. Following Chen et al. (2018), we crawl hundreds of popular English and Chinese novels from online reading site WUXI-AWORLD [1]. All crawled novels are open source and freely available to readers. The novel genres on this site are very diverse, including comprehension novels, fantasy novels, comedy novels, suspense novels, romance novels, science fiction novels, etc. Finally, we collect a total of 1000 novels for Chapter-en and 2000 novels for Chapter-zh.

We filter out articles with low entity density to ensure a sufficient number of annotations. Specifically, we first employ named entity recognition tools stanfordNLP (for English) and LTP (for Chinses) to extract all named entities in the collected chapters, and then we calculate entity density by dividing named entities by the total number of words in the chapter, and filter out chapters with entity density lower than 0.2. To improve the quality of the chapters, we also filter out chapters with less than 256 words and more than 8192 words to balance the lengths of the chapters.

Finally, we select 10k chapters with 7.2M words for ChapterCR-en and 19k chapters with 21M words for ChapterCR-zh.

3.2 Entity & Mention Pre-Annotation

Due to the large size and long text of the selected chapters, it is time-consuming to manually find candidate entities and mentions. Therefore, we pre-label entities and mentions to speed up the labeling process.

[1]https://www.wuxiaworld.com/

3.2.1 Entity Pre-Labeling

For English entity pre-labeling, we employ the NER tool from Stanford CoreNLP [2] to pre-label entities. For Chinese entity pre-labeling, we leverage the NER tool in the LTP platform[3] to pre-label entities. In total, we pre-label 34k and 80k candidate entities for ChapterCR-en and ChapterCR-zh respectively. To assess entity quality, we invite three students to conduct human evaluations. The average F1 is 96%, demonstrating the effectiveness of the named entity tools.

3.2.2 Mention Pre-Labeling

For mention pre-annotation, we divide two cases: Chinese zero mentions and other mentions. For Chinese zero mentions, we additionally train a sequence labeling model. The training data of the sequence labeling model comes from the OntoNotes corpus (Weischedel et al., 2011). During training, the sequence labeling model adopts BERT as the backbone and tags the token preceding the zero mentions to identify zero mentions. For instance, given the sentence "She poured water into the cup until *it* was full", where *it* is omitted in Chinese, the output of the sequence labeling model is "She poured water into the cup until *[Zero Pronoun]* was full".

For other mentions, we employ ChatGPT (Ouyang et al., 2022) for pre-annotation. ChatGPT is an artificial intelligence chatbot developed by OpenAI and trained to follow instructions in a prompt and provide a detailed response. We design multiple prompts to ask ChatGPT questions and adopt their answers as the candidate mentions in the articles. Mainly used prompt is *Please find all possible mentions in the article*. More prompts can be found in Table 3.

Table 3: Prompts for Mention Pre-labeling.

Prompts
List all possible mentions in the chapter
Tell me all the mentions that might refer to entities
As a semantic analyst, find all pronouns

To evaluate the performance of pre-annotated mentions with ChatGPT, we invite three students to do manual evaluations and employ the rule-based method Hobbs algorithm (Hobbs, 1978) as our baseline. Results are shown in Table 4.

[2]https://github.com/stanfordnlp/CoreNLP
[3]https://www.ltp-cloud.com/intro_en

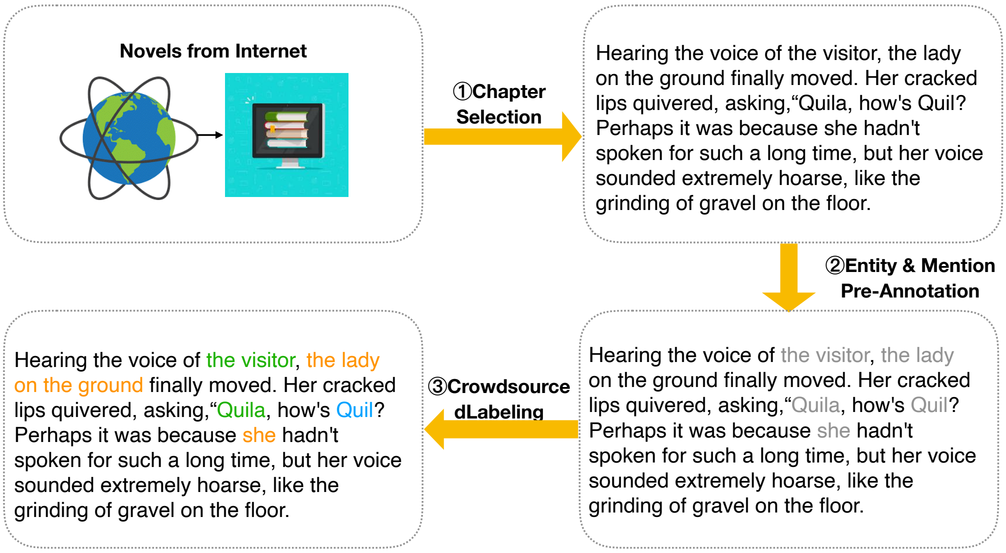

Figure 3: Labeling Process of ChapterCR

Table 4: ChatGPT Performance in Mention Pre-labeling (%).

	P	R	F
Rule-based	27	89	42
ChatGPT	74	90	81

As shown in Table 4, the F1 of ChatGPT is 81%, and ChatGPT outperforms the ruled-based baseline by 39% in F1, suggesting that ChatGPT is a very powerful tool for pre-labeled mentions.

3.3 Crowdsourced Labeling

In this section, we illustrate the process of crowdsourced labeling. Formally, given the selected chapter C and the pre-labeled mention/entity candidates m/e, our goal is to find all possible coreferences between any two of them.

To ensure the quality of crowdsourced labeling, the annotators of ChapterCR-en are either native English speakers or English-major students with TOEFL higher than 100 or IELTS higher than 7.5. The annotators of ChapterCR-zh are native Chinese speakers. Due to the heavy workload, we invited a total of 136 college students to participate in our crowdsourcing annotation through social platforms.

The annotation guideline is illustrated in Appendix A. As shown in the guideline, both ChapterCR-en and ChapterCR-zh have two stages of labeling: *boundary tuning* and *coreference pair matching*. Boundary tuning aims to re-edit the boundary of mentions and entities obtained in Section 3.2 to fix errors in the pre-annotation process. Coreference pair matching aims to determine whether there is a coreference relationship between any two entities and mentions. We respectively introduce the two stages of labeling.

In the stage of boundary tuning, each mention or entity is guaranteed to be labeled by three different annotators. The annotators are required to confirm, delete and re-edit the range of the span (For Chinese zero pronoun resolution, only confirm and delete options are available). If two of the three annotators edit the boundary in the same way, we will accept the revision, otherwise, we will keep the original boundaries as our final result. In addition, annotators will be given a bonus if they find new candidate entities or mentions.

In the stage of coreference pair matching, the annotation process is as follows: for each mention m in the chapter, we consider all entities in the same chapter as answer candidates, from which the annotator needs to select the correct entity referenced by the mention m. Each coreference pair will be labeled by three different annotators and we take the majority vote as the final result. If the three annotators can not agree with each other, we will employ another experienced annotator (accuracy higher than 95%) to make the final decision.

3.3.1 Annotation Quality & Remuneration

Following Artstein and Poesio (2008); McHugh (2012), we use Cohen's kappa coefficient to measure the inter-annotator agreement (IAA) of crowdsourced labeling. The IAA scores are respectively 96% and 92% for boundary tuning and coreference pair matching, indicating very high labeling agreement.

We pay 0.1$ per data per annotator in boundary tuning and 0.3$ per data per annotator in coreference pair matching. According to our standards, the hourly wage of annotators is not less than 10 US dollars per hour, which exceeds the US minimum hourly wage of 7.25 US dollars per hour.

4 Data Analysis

4.1 Overall Statistic

In total, ChapterCR-en labels 10k chapters, 136k mentions and 17k coreferences. ChapterCR-zh labels 19k chapters, 310k mentions and 38k coreferences. The longest length of coreference chains is 31, and the shortest length of coreference chains is 2.

We compare ChapterCR to various representative event extraction datasets in Table 1 and Table 2, including ACE, MUC-6, MUC-7, WikiCoref, CoNLL-2012, WSC, etc.

As shown in Table 1 and Table 2, the data scale of ChapterCR is much larger than existing datasets in many aspects, including the number of mentions and the number of coreferences. Besides, the average length of coreference chains in ChapterCR is 8.1, longer than existing datasets, which poses a great challenge to the long text reading comprehension capability of CR models. Although coreference chains in WikiCoref are also relatively long (6.34 VS 8.1(ours)), the data scale of WikiCoref is quite small and not sufficient for training modern deep learning models.

4.2 Detailed Statistic

We randomly sample 200 chapters with 2,724 mention annotations from ChapterCR-en for more detailed statistical analysis.

We start by analyzing the distribution of the length of the coreference chains in ChapterCR. As shown in Figure 4, 26.6% of the coreference chains have a length less than 5, 53.6% of the coreference chains have a length more than 5 and less than 10, 12.8% of the coreference chains have a length more than 10 and less than 15, and 6.9% coreference chains have a length more than 15.

Then, we analyze gender bias in ChapterCR. Following (Karimi et al., 2016; Webster et al., 2018b),

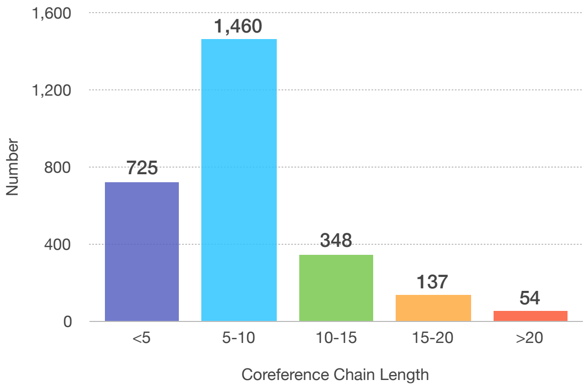

Figure 4: Statistics of Coreference Chain Lengths

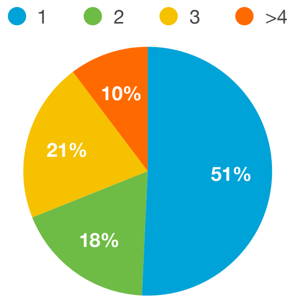

Figure 5: Statistics of Mention Lengths

we use the Gender Guesser library4 [4] to determine the gender of the mentions. According to the statistics, 46.3% of mentions belong to “male” or “mostly male” names, 32.9% of mentions belong to “female” or “mostly female” names, and 20.8% were classified as "unknown". The ratio between female and male candidates is estimated to be 0.58, with male candidates predominating.

Finally, we analyze the length of the mention in ChapterCR. According to the statistics in Figure 5, 51% of the mentions have 1 word, and most of them are personal pronouns, such as she and her. 49% mentions are constituted by more than 2 words, most of them are the description of named entities, such as *that person*, *the beloved woman in front of me* and *the wonderland that I have dreamed of many times in my dreams*.

5 Experiment

In this section, we conduct a variety of experiments to validate the quality and challenges of the proposed dataset. We first introduce the experimental setup and then report the experimental results of the baseline models on our dataset.

[4]https://pypi.org/project/gender-guesser/

5.1 Benchmark Settings

We split ChapterCR(ours) into the training set, validation set, and test set by the ratio of 8: 1: 1. Table 5 shows the data split results.

Method	ChapterCR-en			ChapterCR-zh		
	Train	Dev	Test	Train	Dev	Test
#Doc.	7k	1.5k	1.5k	15k	2k	2k
#Men.	104k	15k	15k	247k	31k	32k
#Coref.	12k	2k	2k	30k	4k	4k

Table 5: Data Split in ChapterCR

5.2 Hyperparameters

For ChatGPT, we use the official ChatGPT interface [5] provided by OpenAI to call it. The ChatGPT version is GPT-3.5-turbo-0301 and the maximum input length is 16k. All the baseline models are trained on 8 A100 GPUs with 80G memory. We report the average result of five rounds as the final result. For human evaluation, we randomly select 200 chapters from English and Chinese novels respectively, and invite three students to make annotations. The final result is the average of their annotation accuracy.

Following Joshi et al. (2019b), we utilize precision, recall, and F1 score to evaluate the performance of the baselines on our dataset. All the metrics are calculated in the B3 manner (Bagga and Baldwin, 1998), which treats each mention cluster (a set of mentions pointing to the same entity) as a class, and then calculates precision, recall, and macro-average F1 score via multi-classification.

5.3 Baseline

We introduce the following baselines to evaluate ChapterCR, including: **e2e-coref** (Lee et al., 2017) is an end-to-end coreference resolution model, which considers all spans in a document as potential mentions and learns the probabilities of possible antecedents for each mention. **c2f-coref** (Lee et al., 2018) introduces a coarse-to-fine approach that allows for more aggressive span pruning without compromising accuracy to accelerate coreference resolution. **CR-BERT** (Joshi et al., 2019b) applies BERT to coreference resolution, achieving strong improvements on the CoNLL2012 and GAP benchmarks. **SpanBERT** (Joshi et al., 2019a) upgrades BERT from word-level pre-training to span-level pre-training via geometric masking to better cope with span-level task coreference resolution. **WL-COREF** (Dobrovolskii, 2021) finds coreferences between words rather than word spans, and then reconstructs the word spans to reduce the complexity of the coreference model. **Link-Append** (Bohnet et al., 2022) uses the seq2seq paradigm and transition matrix to jointly predict mentions and entities, which formulate coreference resolution as a generation task. **Fast-COREF** (Otmazgin et al., 2022) is a substantially faster model based on the LingMess architecture, providing state-of-the-art coreference accuracy. **ChatGPT** is a chatbot developed by OpenAI, which has gained widespread popularity and media attention (Leiter et al., 2023). We introduce ChatGPT as our baseline to answer whether SOTA pre-trained models can perform well on chapter-level coreference resolution. We obtain the answer by asking ChatGPT "which entity is the <mention> in <sentence> referring to", where <mention> and <sentence> will be replaced with specific phrases in actual usage.

[5]https://openai.com/blog/introducing-chatgpt-and-whisper-apis

5.4 Overall Performance

Table 6 shows the experimental results of ChapterCR-en and ChapterCR-zh, from which we have the following observations.

(1) Human beings have achieved good performance on ChapterCR, with an average F1 score of 91.3 on the English corpus and 90.4 on the Chinese corpus, which shows the high quality of ChapterCR. (2) There is still a gap between the performance of SOTA coreference resolution models and human beings, indicating that ChapterCR is an open issue. Humans are good at connecting key information and thus can understand long text semantics more coherently, while current deep learning CR models suffer from catastrophic forgetting, which leads to inferior performance on long-chain coreference resolution. (3) Even the powerful ChatGPT does not achieve satisfactory performance on ChapterCR, with the F1 score of 74.0% in ChapterCR-en and 58.8% in ChapterCR-zh. One possible reason is that ChatGPT is trained by next token prediction, which does not help much for fine-grained coreference resolution. For example, in the sentence *Jack hits Bill, but he apologized later.*, whether we rewrite *he* with *Bill* or *Jack*, the probability of the next token prediction is not much different. (4) There is a performance degradation from ChapterCR-en to ChapterCR-zh. There are multiple zero pronoun resolutions in ChapterCR-zh. Due

Table 6: Overall Performance on ChapterCR (%).

Methods	ChapterCR-en			ChapterCR-zh		
	P	R	F	P	R	F
e2e-coref	62.4	58.3	60.3	53.2	62.3	57.4
c2f-coref	69.3	68.4	68.8	58.3	68.8	63.1
CR-BERT	75.6	70.5	73.0	62.7	70.8	66.5
SpanBERT	73.2	71.7	72.4	68.1	67.4	67.7
WL-COREF	71.8	72.9	72.3	60.7	63.3	62.0
Link-Append	68.6	64.1	66.3	58.9	67.2	62.8
Fast-COREF	74.3	77.6	75.9	67.9	68.1	68.0
ChatGPT	77.2	71.0	74.0	57.3	60.3	58.8
Human	93.6	89.1	91.3	96.3	85.1	90.4

Table 7: Error Analysis in ChapterCR.

Error Types	**Examples**
Closest Selection	Jerebai are you still expecting him to save you? Today is the day that he gets married! He is in love – do you really expect that **you** would even cross his mind?!" Quila cried. Predict: Quila Golden: Jerebai
Gender Confusion	Dad, you should mind your own business, she said. Don't say that to father, a little boy said. See what a sweet daughter you've got, the man's wife said. Predict: a little boy Golden: a sweet daughter
Multiple Entities	Emma said "I am not the killer, and I think it was James that killed Mason". "I didn't do that. I saw Oliver last night. It must be him". "No you are lying. Oliver does not hate Mason, and we all know that.", Ava said. Predict: Mason Golden: James

to the lack of mentions, existing models have little evidence to rely on during the resolution process, resulting in poor performance.

5.5 Error Analysis

In this section, we analyze common errors in ChapterCR, and propose several future research directions to improve coreference resolution.

A common error in ChapterCR is nearest selection. Existing CR models often simply and rudely believe that a mention refers to its closest entity. For instance, in the first example in Table 7, existing CR models do not take context into account and mistakenly assume that the mention *you* refers to the closer entity *Quila*, rather than the farther but correct entity *Jerebai*.

Another common error in ChapterCR is that existing CR models lack the commonsense to discern the gender of the mention. For instance, in the second example in Table 7, existing CR models fail to understand that the pronoun of *she* should be a female rather than a male, which leads to the model incorrectly resolving *she* to *a little boy* instead of *a sweet daughter*.

The third common error in ChapterCR is that existing CR models will be very confused if there are too many entities surrounding the mention in the text. For instance, in the third example in Table 7, there are lots of entities in the text, including *Emma, James, Mason, Oliver, Ava*. Faced with so many choices, it is difficult for existing CR models to understand that *you* here refers to *James*.

We believe the following directions are worthy of attention: (1) More diversity of data sources. Since we only annotate coreferences from novels, future datasets may include more types of data sources. (2) Injecting ontology and commonsense knowledge. With the help of external knowledge, existing CR models can be constrained by gender concordance, which can effectively reduce gender errors. (2) Focusing on entity-level information. By using entities as bridges, existing CR models can more coherently integrate information in longer texts, which helps to address the challenge of long-distance coreference resolution.

6 Conclusion

In this paper, we propose ChapterCR, a large-scale chapter-level coreference resolution dataset. ChapterCR greatly expands the data scale, with a total of 446k mentions and 55k coreference chains, and increases the length of the coreference chain, with an average coreference chain length of 8.1. Experiments demonstrate that the performance of SOTA models cannot catch up with human beings, showing that ChapterCR is an open issue.

References

1995. *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

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

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Eric Bengtson and Dan Roth. 2008. Understanding the value of features for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2022. Coreference resolution through a seq2seq transition-based system.

Arthur Brack, Daniel Uwe Müller, Anett Hoppe, and Ralph Ewerth. 2021. Coreference resolution in research papers from multiple domains. *CoRR*, abs/2101.00884.

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Hong Chen, Zhenhua Fan, Hao Lu, Alan L Yuille, and Shu Rong. 2018. Preco: A large-scale dataset in preschool vocabulary for coreference resolution. *arXiv preprint arXiv:1810.09807*.

Kevin Clark. 2015. Neural coreference resolution.

Kevin Clark and Christopher D Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC bioinformatics*, 18(1):1–14.

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

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *Lrec*, volume 2, pages 837–840. Lisbon.

Pradheep Elango. 2005. Coreference resolution: A survey. *University of Wisconsin, Madison, WI*, page 12.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142.

Lynette Hirschman. 1997. Muc-7 coreference task definition, version 3.0. *Proceedings of MUC-7, 1997.*

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. 2019a. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529.

Mandar Joshi, Omer Levy, Daniel S. Weld, and Luke Zettlemoyer. 2019b. Bert for coreference resolution: Baselines and analysis.

Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, pages 53–54.

Vid Kocijan, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. 2019a. Wikicrem: A large unsupervised corpus for coreference resolution.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019b. A surprisingly robust trick for winograd schema challenge. *arXiv preprint arXiv:1905.06290.*

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Kusum Lata, Pardeep Singh, and Kamlesh Dutta. 2022. Mention detection in coreference resolution: survey. *Applied Intelligence*, pages 1–45.

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

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. Chatgpt: A meta-analysis after 2.5 months.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, pages 1–43.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Overview of the protein coreference task in bionlp shared task 2011. In *Proceedings of the BioNLP shared task 2011 workshop*, pages 74–82.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Cheoneum Park, Kyoung-Ho Choi, Changki Lee, and Soojong Lim. 2016. Korean coreference resolution with guided mention pair model using deep learning. *Etri Journal*, 38(6):1207–1217.

James Pustejovsky, José Castano, Roser Sauri, Jason Zhang, and Wei Luo. 2002. Medstract: creating large-scale information servers from biomedical texts. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 85–92.

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*, pages 777–789.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Isabel Segura-Bedmar, Mario Crespo, César de Pablo-Sánchez, and Paloma Martínez. 2010. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. In *BMC bioinformatics*, volume 11, pages 1–9. BioMed Central.

Kumar Shridhar, Nicholas Monath, Raghuveer Thirukovalluru, Alessandro Stolfo, Manzil Zaheer, Andrew McCallum, and Mrinmaya Sachan. 2022. Longtonotes: Ontonotes with longer coreference chains. *arXiv preprint arXiv:2210.03650*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Kepa Rodriguez, and Massimo Poesio. 2016. Arrau: Linguistically-motivated annotation of anaphoric descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2058–2062.

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. *arXiv preprint arXiv:1804.07461*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018a. Mind the gap: A balanced corpus of gendered ambiguou. In *Transactions of the ACL*, page to appear.

Kim Webster, Kristin Diemer, Nikki Honey, Samantha Mannix, Justine Mickle, Jenny Morgan, Alexandra Parkes, Violeta Politoff, Anastasia Powell, Julie Stubbs, et al. 2018b. *Australians' attitudes to violence against women and gender equality*. Australia's National Research Organisation for Women's Safety.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Mingzhu Wu, Nafise Sadat Moosavi, Dan Roth, and Iryna Gurevych. 2020. Coreference reasoning in machine reading comprehension. *CoRR*, abs/2012.15573.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. An np-cluster based approach to coreference resolution. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 226–232.

Xintong Yu, Hongming Zhang, Ruixin Hong, Yangqiu Song, and Changshui Zhang. 2022. Vd-pcr: Improving visual dialog with pronoun coreference resolution. *Pattern Recognition*, 125:108540.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2004. Coreference resolution for information extraction. In *Proceedings of the Conference on Reference Resolution and Its Applications*, pages 24–31, Barcelona, Spain. Association for Computational Linguistics.

Rui Zhang, Cicero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. *arXiv preprint arXiv:1805.04893*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2019. Evaluating commonsense in pretrained language models. *CoRR*, abs/1911.11931.

A Annotation interface and instructions

The annotations interface is implemented based on label-studio. The annotations consist of two tasks: Boundary Tuning and Mention Pair Matching, and their details are shown in this section.

A.1 Boundary Tuning

As shown in Figure 6, the interface requires annotators to decide whether to modify the predefined boundary. The following passage is the instruction used during annotation.

The boundary tuning task aims to correct wrong spans pre-labeled. For example, in the sentence *the sad man is looking for his wife.*, *man* is labeled as a mention, but it is incorrect. The entire mention should be *the sad man*, which means that the annotators should identify the maximal extent of the string that represents the mention. Click the mention to highlight it and then click the modify button. The mention span can be modified, and click the save button after modification. Please stay unchanged if no mistakes are found. The annotations will be used for research purposes.

A.2 Mention Pair Matching

As shown in Figure 7, in mention pair matching, annotators should find the entity that best matches a mention. The instruction is as follows.

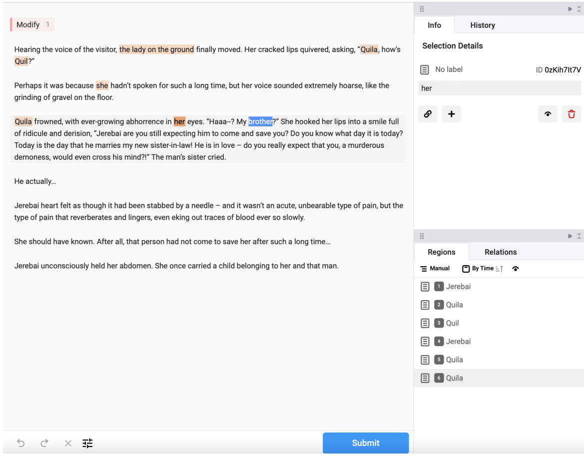

Figure 6: boundary tuning

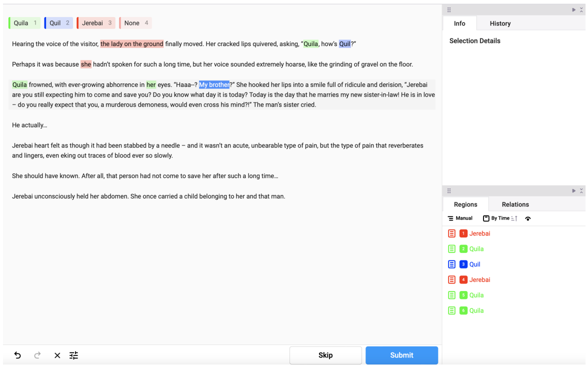

Figure 7: mention pair matching

Mentions are highlighted and the entities are listed above the text. Please choose the correct entity in the menu and then click the mention. If no correct entity is shown in the list, please click the None button and then click the mention. The numbers of total mentions and unannotated mentions are shown at the bottom of the page. Only after finishing all the annotations on one page, the results can be saved and annotators can get paid. The annotations will be used for research purposes.