NovelCR: A Large-Scale Bilingual Dataset Tailored for Long-Span Coreference Resolution

Anonymous ACL submission

Abstract

Coreference resolution (CR) endeavors to match pronouns, noun phrases, etc. with their referent entities, acting as an important step for deep text understanding. Presently available CR datasets are either small in scale or restrict coreference resolution to a limited text span. In this paper, we present NovelCR, a large-scale bilingual benchmark trailer for long-span coreference resolution. NovelCR not only contains extensive mentions and coreferences annotations (148k mentions and 128k coreferences in NovelCR-en, 311k mentions and 273k coreferences in NovelCR-zh), but also contains numerous long-span coreferences. Specifically, 74% of the coreferences in NovelCR-en and 83% of the coreferences in NovelCR-zh span over three or more sentences, which is significantly higher than the proportion of long-span coreferences in existing datasets. Experiments on NovelCR reveal a large gap between state-of-the-art baselines and human performance, highlighting that NovelCR remains an open issue.

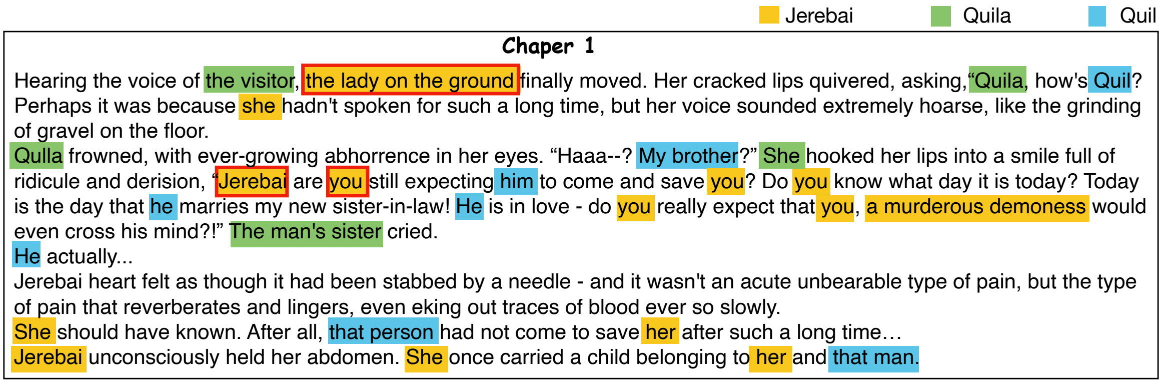
1 Introduction

Coreference resolution (CR) aims to identify mentions and their referent entities from text. For instance, given the sentence *Recently, Apple sued Qualcomm, suing it for failing to cooperate by contracts*, coreference resolution needs to distinguish that mention *it* here refers to entity *Qualcomm* instead of *Apple*. Coreference resolution is a core task in deep text analysis and acts as a prerequisite for multiple advanced natural language processing applications such as machine reading comprehension (Wu et al., 2020), information extraction (Zelenko et al., 2004), and multi-round dialogue construction (Yu et al., 2022).

However, existing coreference resolution datasets either suffer from small data scales or restrict coreference resolution within a limited text span. ACE2004 (Doddington et al., 2004) annotates coreferences from merely 451 documents. The data scales of WikiCoref (Ghaddar and Langlais, 2016), MUC-6 (muc, 1995), MUC-7 (Hirschman, 1997), STM-coref (Brack et al., 2021) are even smaller, comprising 30, 60, 50, and 110 documents, respectively. LongtoNotes (Shridhar et al., 2022) encompasses a larger but still restricted 2415 documents. Given their small scale, none of these coreference datasets can fairly assess the performance of modern neural coreference resolution models. Besides, WSC (Levesque et al., 2012), LitBank (Bamman et al., 2020), GAP (Webster et al., 2018a) and CLUEWSC2020 (Xu et al., 2020) limits the scope of the coreference resolution in a single sentence, and most of the coreferences in CoNLL2012 (Weischedel et al., 2011), ECB+ (Cybulska and Vossen, 2014), and DWIE (Zaporojets et al., 2020) scatter in three sentences. The prevalence of short-span coreferences leads to little interference between the mention and the referred entity, making these datasets less challenging.

It is necessary to focus on long-span coreference resolution. Long spans mean more complex relationships between entities and references, such as distant mentions, ambiguous pronouns, and intervening references, which can encourage the development of CR models that can handle more complex linguistic phenomena. Taking Figure 1 as an example, it is easy to recognize that *you* refers to *Jerebal*. However, understanding that *the lady on the ground* also refers to *Jerebal* is much more challenging, needing to unravel the correspondences between speakers and participants in the conversation and requiring a deep analysis of the text.

In this paper, we introduce NovelCR, a large-scale, high-quality benchmark to address long-span coreference resolution. Specifically, we focus on resolving the coreferences of novel characters to enable NovelCR to contain abundant long-span coreferences. The underlined reason is that due to the strong narrative coherence of novels, novel



🟨 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.

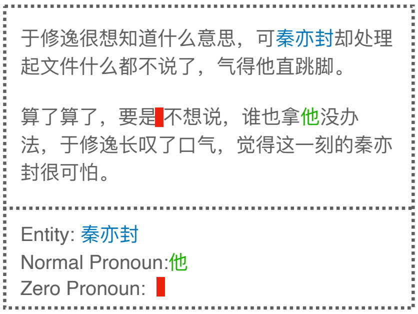
Figure 1: An example of NovelCR. NovelCR focuses on coreference resolution of novel characters such as Jerebal, Quila, and Quil.*Jerebal* and *the lady on the ground* is a long-span coreference, 6 sentences apart. *Jerebal* and *you* is a short-span coreference, located in 1 sentence.

characters, such as *Jerebal*, *Quila*, and *Quil* in Figure 1, are highly likely to be referenced again after spanning multiple sentences.

The construction process of NovelCR is as follows: we first obtain English and Chinese novels from online websites. Then, we leverage NER tools and prompt learning to collect candidate entities and mentions from novel chapters. We cover a wide range of mentions in our dataset, including pronouns (e.g., *she* and *her*), proper and common noun phrases (e.g., *the visitor*, *the man's sister* and *a murderous demones*) to reduce the likelihood of missing labels. Afterward, we utilize crowdsourcing to remove improper mentions and re-edit mention boundaries to ensure that all mentions adhere to the maximum span principle. Finally, annotators are required to answer multiple-choice questions to match mentions to their referent entities.

We highlight the three contributions of NovelCR: (1) Large scale. NovelCR contains a total of 460k mentions and 128k coreferences, which is much larger than existing CR datasets. (2) Abundant long-span coreferences. NovelCR contains numerous long-span coreferences. The number of coreferences scattered over 3 or more sentences is 95,346 in NovelCR-en, which is significantly higher than the number of 12,104 in LongtoNotes (Shridhar et al., 2022), another dataset that specializes in long-span coreferences. (3) Bilingual. NovelCR annotates coreferences from both English and Chinese novels. Besides, we introduce zero pronoun resolution in NovelCR-zh (as shown in Figure 2), which brings additional challenges to the provided dataset.

We evaluate NovelCR against eight state-of-the-art CR baselines. Experiments show that there is still a large gap between the SOTA baselines and human beings, revealing that NovelCR remains an unresolved challenge. Detailed experiments further demonstrate that existing CR models have a significant performance degradation when coreference is scattered throughout a longer text, showing that abundant long-span coreferences make NovelCR very 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。

Entity: 秦亦封
Normal Pronoun:他
Zero Pronoun: 🟥

Figure 2: An example of zero pronoun resolution.

2 Related Work

Coreference resolution is a core task in natural language processing (Elango, 2005; Sukthanker et al., 2020; Lata et al., 2022; Liu et al., 2023). Numerous high-quality datasets have been proposed to promote the development of coreference resolution.

Muc-6 (muc, 1995) and MUC-7 (Hirschman, 1997) are the two earlier proposed coreference resolution datasets, and the data scale is relatively small, with 60 and 50 documents, 30k and 25k tokens respectively. WikiCoref (Ghaddar and Langlais, 2016) contains merely 30 documents and 7955 mentions. STM-coref (Brack et al., 2021) annotates coreferences from no more than 110 research papers. GUM (Zeldes, 2017) and ARRAU (Uryupina

Datasets	#Doc.	#Sent.	#Tok.	#Mention	#Coref.	#CorefDis.	#LongCorefRatio
ACE2004	451	18,530	158k	22550	-	-	-
MUC-6	60	3,750	30k	-	-	-	-
MUC-7	50	3,197	25k	-	-	-	-
WikiCoref	30	2,292	60k	7,955	6,700	3.5	0.46
WSC	-	803	20k	2,409	1,606	1.0	0.0
GAP	-	8,908	317k	26,724	17,816	1.0	0.0
STM-coref	110	1,480	26k	2,577	1,669	2.4	0.29
CoNLL2012	3,493	112,941	1.6M	56,371	43,560	2.9	0.34
LongtoNotes	2,415	112,941	1.6M	38,640	32,715	3.3	0.37
LitBank	100	108,000	13M	57,514	28,411	1.0	0.0
ECB+	502	9,171	221K	32297	12,930	3.1	0.41
DWIE	802	13,628	501k	43,373	20,243	2.8	0.35
NovelCR-en(ours)	9,462	54,820	8.1M	148,529	128,847	4.6	0.74

Table 1: Statistics of English coreference resolution datasets. Doc.: chapters, Sent.: sentences, Entity: entities, Mention: mentions, Coref: coreference pairs, CorefDis: the distance between mention and referring entity, measured in sentences, LongCorefRatio: ratio of coreferences spread over 3 or more sentences

Datasets	#Doc.	#Sent.	#Tok.	#Mention	#Coref.	#CorefDis.	#LongCorefRatio
ACE2004	646	14,233	154K	28,135	-	-	-
CoNLL2012	2,280	83,763	950k	15,136	8,859	3.1	0.45
CLUEWSC2020	-	1,648	276K	4,944	3,296	1.0	0.0
NovelCR-zh(ours)	19,288	80,872	21M	311,482	273,379	5.2	0.83

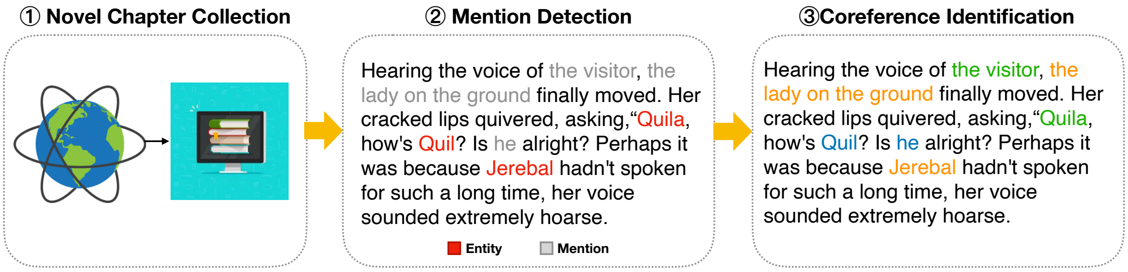
Table 2: Statistics of Chinese coreference resolution datasets

et al., 2016) solve anaphora resolution from open source multi-layer corpus with barely 300 documents. ACE2004 (Doddington et al., 2004) is a widely adopted CR dataset that covers multiple domains, including news communications, broadcast programs, and online blogs. However, it contains a relatively small amount of data, with just 451 documents and 22,550 mentions. Similarly, the data size of LongtoNotes (Shridhar et al., 2022) is also limited, with only 2,415 documents and 38,640 mentions. In contrast, the proposed dataset NovelCR features an extensive dataset, with 28k documents and 460k mentions, far exceeding existing CR datasets. Additionally, the proposed dataset NovelCR focuses on both English and Chinese coreference resolution, unlike PreCo (Chen et al., 2018), which is a single-language dataset.

Winograd Schema Challenge (WSC) (Levesque et al., 2012) is a well-known CR benchmark proposed by Hector Levesque, consisting of 803 coreferences. WSCR (Rahman and Ng, 2012), PDP (Davis et al., 2017), WINOBIAS (Zhao et al., 2018), and WinoGrande (Sakaguchi et al., 2021) are datasets evolved from the WSC. LitBank (Bamman et al., 2020), like our dataset, considers annotating coreferences from novels, and a total of 100 British novels are annotated. GAP (Webster et al., 2018a), sampled from Wikipedia, is a gender-balanced dataset and contains 8,908 coreferences of ambiguous pronouns and antecedent names. All the above seven datasets limit coreference resolution in a single sentence. Besides, most of the coreference pairs in CoNLL2012-en, CoNLL2012-zh (Weischedel et al., 2011), ECB+ (Cybulska and Vossen, 2014), and DWIE (Zaporojets et al., 2020) appear within the scope of three sentences. The prevalence of short-span coreferences in these datasets makes them less challenging. Unlike them, our proposed dataset NovelCR contains a significant number of long-span coreferences, and this abundance of long-span coreferences necessitates a more robust semantic understanding model to effectively handle NovelCR.

3 Dataset Construction

In this section, we illustrate the dataset construction process. As shown in Figure 3, We construct NovelCR in three steps: novel chapter collection, mention detection, and coreference identification. Novel chapter collection aims to gather chapters from a wide range of genres sourced from online novel websites. Mention detection leverages NER tools and prompt learning to mine potential entities and mentions from novel chapters. Coreference



Hearing the voice of the visitor, the lady on the ground finally moved. Her cracked lips quivered, asking,“Quila, how's Quil? Is he alright? Perhaps it was because Jerebal hadn't spoken for such a long time, her voice sounded extremely hoarse.

Hearing the voice of the visitor, the lady on the ground finally moved. Her cracked lips quivered, asking,“Quila, how's Quil? Is he alright? Perhaps it was because Jerebal hadn't spoken for such a long time, her voice sounded extremely hoarse.

Figure 3: Labeling Process of NovelCR

identification uses crowdsourcing to distinguish coreference pairs in chapters by converting coreference resolution into multiple-choice questions.

3.1 Novel Chapter Collection

We select online novels as our data source. The underlined reason is that novels, unlike news articles, exhibit strong narrative coherence and are more likely to include long-span coreferences. Specifically, we crawl hundreds of popular English and Chinese novels from the online reading site WUXIAWORLD [1], all of which are open source and free to access. The crawled novels encompass a wide range of genres such as cultivation, fantasy, comedy, suspense, romance, science fiction, etc. In total, we collected 1000 English novels for NovelCR-en and 2000 Chinese novels for NovelCR-zh. These novels were originally written in Chinese and translated into English by human experts. Due to the incomplete translation, the number of English novels is less than that of Chinese novels.

We filter out novel chapters with less than 256 tokens and more than 32,768 tokens to balance the document lengths. Additionally, we utilize NER tools (Stanford NLP for English and LTP for Chinese) to filter out chapters with less than 8 entities, ensuring abundant coreference annotations. After two rounds of filtering, we collect 9,462 novel chapters for NovelCR-en and 19,288 novel chapters for NovelCR-zh.

3.2 Mention Detection

This section aims to detect candidate mentions and entities from novel chapters. To reduce the burden on annotators, mention detection is divided into two steps. The first step is to use NER tools and prompt learning to mine candidate entities and mentions, and the second step is to employ annotators to do manual verification.

3.2.1 Candidates Collection

To detect candidate entities, we employ Stanford CoreNLP [2] and LTP NER tool[3] to recognize named entities from English and Chinese chapters, respectively. Finally, we detected 42,849 and 98,571 person entities for NovelCR-en and NovelCR-zh respectively. We involve three students to conduct human evaluations to assess the quality of annotations. The average recall rates of NER on NovelCR-en and NovelCR-zh are 99.1% and 98.9%, respectively, demonstrating the effectiveness of the named entity tools.

Previous datasets usually employ POS tagging to detect pronoun mentions and semantic parsers to detect noun phrase mentions, yet these mention detection methods are pattern-dependent and the mention recall rate is not that high. In this paper, we employ prompt-learning (Ouyang et al., 2022) to jointly identify pronoun and noun phrase mentions. Empowering by ChatGPT, prompt learning has strong text comprehension capabilities and can identify a wider variety of mentions. Finely designed prompts are shown in Appendix C. We take the union of annotations of different prompts as the final annotation result.

Datasets	NovelCR-en	NovelCR-zh
	Recall	
POS+Semantic Parser	91.3	90.7
Prompt-Learning(ours)	99.1	98.9

Table 3: Candidate Mention Detection Performance(%)

We engage three students to conduct human evaluations. As shown in Table 3, compared to the traditional method (POS tagging+Semantic Parser), our proposed method (Prompt-Learning) improves

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

[2] https://github.com/stanfordnlp/CoreNLP

[3] https://www.ltp-cloud.com/intro_en

the recall rate by 7.8% and 8.2% on NovelCR-en and Novel-zh, respectively, effectively reduce the risk of missing annotations.

We additionally train a sequence labeling model to handle Chinese zero pronoun resolution. We leverage OntoNotes (Weischedel et al., 2011) as our training corpus and adopt BERT as the model backbone. The training goal is to insert a special token before the zero pronoun. For instance, given the sentence "She poured water until *it* was full", where *it* is omitted in Chinese, the output of the sequence labeling model is "She poured water until *[Zero Pronoun]* was full". The average recall rate in human evaluation is 87.4% on Chinese zero pronoun resolution.

3.2.2 Manual Verification

In the section, we manually verify the entities and mentions obtained in Section 3.2. We invite a total of 136 Chinese college students to participate in our crowdsourcing annotation. The annotators of NovelCR-en are English-major students with TOEFL higher than 100 or IELTS higher than 7.5, and the annotators of NovelCR-zh are native Chinese speakers.

As shown in the guideline in Appendix A, annotators first remove invalid mentions, i.e., mentions that do not refer to a person entity, such as *the bank* and *this beautiful knife*. In particular, *her* in *her split lips* is also considered an invalid mention as it functions as a modifier of *lips* rather than an independent personal pronoun. Only mentions verified by at least two annotators will be retained. By removing invalid mentions, we ensure the quality of mentions in the proposed dataset, but it may also cause missing annotations, which we will discuss in the limitations section.

After that, the annotators are required to refine the boundary of the mention. We adopt the principle of maximum span. For example, given the mention *a little child*, if the original annotation is *child*, the annotator needs to adjust the boundary to *a little child*. If two of the three annotators edit the boundary in the same way, we will accept the revision, otherwise, we will ask one additional annotator to make the final decision.

3.3 Coreferences Identification

In this section, we leverage crowdsourcing to identify coreferences from the novel chapters, which is the core task of NovelCR.

We reframe coreference identification as a multiple-choice question. Specifically, we first collect the entity set E from the chapter and deduplicate it. Then, for each mention m in the chapter, We ask the annotators to determine which entity in E the mention m refers to. Taking Figure 3 as an example, the entity set in the novel chapter is $\{$*Quila*, *Quil*, *Jerebal*$\}$. Given the mention *the visitor*, annotators need to determine which entity *the visitor* refers to, *Quila*, *Quil* or *Jerebal*. We adopt the answer *Quila* as the final coreference annotation.

Each mention undergoes labeling by three individual annotators, with the final result determined by the majority vote. 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. The guideline is shown in Appendix B. We remove the singleton mentions after finishing the annotation.

3.4 Annotation Quality & Remuneration

We use Cohen's kappa coefficient (Artstein and Poesio, 2008; McHugh, 2012) to measure the inter-annotator agreement (IAA) of crowdsourced labeling. The IAA scores are respectively 96% and 92% for mention verification (Section 3.2.2) and coreference identification (Section 3.3) respectively, indicating very high labeling agreement.

We pay 0.1$ per data per annotator in mention verification and 0.3$ per data per annotator in coreference identification. 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. We release NovelCR under the Open Data Commons Open Database License (ODC-ODbL).

4 Data Analysis

4.1 Overall Statistic

We compare NovelCR-en and NovelCR-zh to existing representative English and Chinese coreference resolution datasets in Table 1 and Table 2 respectively.

From the tables, we can draw the following observations. First, our dataset is much larger than existing CR datasets. As shown in Table 1, NovelCR-en contains 9,462 documents, 54,820 sentences, 8.1M tokens, 148,529 mentions, and 128,837 coreference pairs. Even compared with the current large CR datasets CoNLL2012 and LongtoNotes, our dataset is still 2.9 and 2.6 times larger in terms of documents and 3.0 and 4.0 times larger in terms of the number of coreference pairs. This phenomenon is more pronounced in comparisons involving Chinese datasets. As shown in Table 2, NovelCR-zh contains 19,288 documents, 80,872 sentences, 21M tokens, 311,482 mentions, and 273,379 coreference pairs. The number of coreference pairs is 30.9 times that of CoNLL2012 and 82.9 times that of CLUEWSC2020.

In addition, our dataset contains abundant long-span coreferences. As shown in Table 1, the average distance between coreferences in NovelCR-en is 4.6 sentences, longer than that in longtoNote (3.3 sentences) and ECB+ (3.1 sentences), both of which also focus on long-span coreference resolution. NovelCR-en also has the largest proportion of coreferences spread over 3 or more sentences, reaching 74%, which is greater than longtoNote (37%) and ECB+ (41%). This is even more obvious in the comparison of Chinese data sets. The ratio of long-span coreferences in NovelCR-zh has reached 83% as shown in Table 2, far exceeding the existing datasets, indicating the complexity and challenges of our dataset.

4.2 Detailed Statistic

In the section, we present detailed statistics for Novel-en, including the distribution of coreference distance, mention length, document length, and gender balance.

First, we analyze the distribution of coreference distances to observe the proportion of long-span coreferences in our dataset. As shown in Figure 4, 26.5% of coreference pairs appear in less than three sentences. In addition, it can be seen that NovelCR contains a large number of coreferences with very long spans. For example, 16.6% of coreference pairs are scattered in three and four sentences. 17.4% of coreference pairs span five to seven sentences. Coreferences separated by eight to ten sentences account for 17.9%, and coreferences separated by more than ten sentences also account for a large proportion, reaching 21.6%.

Then, we analyze the distribution of the length of mentions. According to statistics, 54% mentions contain 1 word, most of which are entities and personal pronouns, such as *she* and *her*. 36% mentions consist of 2-5 tokens, and 10% mentions exceed 5 tokens, most of which were noun phrases of named entities, such as *that person*, and *the beloved woman in front of me*.

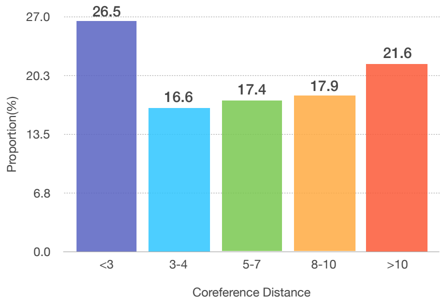


Figure 4: Coreference Distance Distribution

We also analyze the length of documents. Statistics reveal that 61% of documents consist of less than 10k tokens. 33% of documents are comprised of 10k-20k tokens, while 6% of documents extend beyond 20k tokens.

Lastly, we analyze gender bias within our dataset. Following (Karimi et al., 2016; Webster et al., 2018b), we use the Gender Guesser library4[4] to determine the gender of the mentions. According to the statistics, 45.1% of mentions belong to *male* or *mostly male* names, 34.2% of mentions belong to *female* or *mostly female* names, and 20.7% were classified as *unknown*. The ratio between female and male candidates is estimated to be 58%, with male candidates predominating.

5 Experiment

5.1 Benchmark Settings

We split NovelCR-en and NovelCR-zh into the training, validation, and test set by 8: 1: 1. Table 4 shows the statistical analysis of the dataset splits.

Method	NovelCR-en			NovelCR-zh		
	Train	Dev	Test	Train	Dev	Test
#Doc.	7k	1.5k	1.5k	15k	2k	2k
#Men.	118k	15k	15k	247k	31k	32k
#Coref.	14k	3k	3k	30k	4k	4k

Table 4: Data Split in NovelCR

5.2 Hyperparameters & Metrics

Our experiments are conducted on eight A100 GPUs with 80GB of memory. The baseline training time is less than half an hour. We utilize the default hyperparameters in the baseline release code. Long chapters are split into non-overlapping segments of up to 2048 word-piece tokens. For human evaluation, we invited three students to annotate 200

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

documents randomly selected from NovelCR-en and NovelCR-zh and report the average accuracy of the three students as the final results.

Following Cattan et al. (2021), we utilize precision, recall, and F1 to evaluate the performance of existing baselines on the proposed dataset. All the metrics are calculated in B3, MUC, CEAFe, and CoNLL to allow adequate comparison. We report the average result of five rounds.

5.3 Baseline

In this section, we introduce eight baselines to validate the challenges of the NovelCR, including: **e2e-coref** (Lee et al., 2017) is an end-to-end coreference resolution model, which considers all spans 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 to accelerate coreference resolution, which allows for more aggressive span pruning without compromising accuracy. **CR-BERT** (Joshi et al., 2019b) applies BERT to coreference resolution, achieving significant 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 coreference resolution. **WL-COREF** (Dobrovolskii, 2021) finds coreferences at the granularity of tokens 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. **Fastcoref** (Otmazgin et al., 2022) is a precise and user-friendly coreference resolution algorithm that is widely used. We employ LingMess implementation in our experiments. **GPT-4** is a powerful open-domain large language model developed by OpenAI [5]. We construct the prompt: *Which entity is the <mention> in <sentence> referring to?* to apply GPT-4 to coreference resolution.

5.4 Overall Performance

Table 5 and Table 6 show the experimental results of NovelCR-en and NovelCR-zh, from which we have the following observations.

(1) Human beings have achieved good performance on NovelCR, achieving an F1 score of 91.4% on NovelCR-en and 90.5% on NovelCR-zh using the CoNLL metric, demonstrating the high quality of NovelCR. (2) Current CR baselines still suffer from a performance gap compared to human beings, with the state-of-the-art model achieving 77.0% F1 score on NovelCR-en (Fastcoref) and 68.5% F1 score on NovelCR-zh (SpanBERT), about 20% lower than the scores of human evaluations. Also, the powerful GPT-4 does not achieve satisfactory performance on NovelCR, with an F1 score of 82.5% on NovelCR-en and 73.2% on NovelCR-zh, indicating that NovelCR remains an open issue. Humans can not only utilize extensive world knowledge to infer coreference relationships, but also possess strong logical reasoning abilities, capable of handling complex scenarios such as indirect references and implicit information. Therefore, humans achieve better results than current CR models in this regard.

5.4.1 Short-Span or Long-Span

In this section, we observe the performance differences of existing CR models when dealing with short-span and long-span coreference resolution. Specifically, we categorize the coreference pairs in NovelCR-en into three groups: coreference pairs appear in less than 3 sentences (*<3*), between 3-5 sentences (*3-5*), and beyond 5 sentences (*>5*). We adopt Fastcoref as our CR baseline.

Sent.	<3	3-5	>5
Fastcoref	82.6	75.3	61.0

Table 7: Short-Span VS Long-Span(%)

From Table 7, we have the following observation. As the distance between the coreference pairs increases, from <3 sentences, 3-5 sentences to >5 sentences, the existing state-of-the-art CR method (Fastcoref) suffers from significant performance degradation, from 82.6%, 75.3% to 61.0% in F1 score, indicating that long-span coreference resolution is indeed a challenging task. It is necessary to propose NovelCR to pave the way for better long-span coreference resolution models.

5.5 Error Analysis

In this section, we analyze common errors in NovelCR. One of the common errors is the 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

[5] https://chatgpt.com

Methods	B3			MUC			CEAFe			CoNLL
	P	R	F	P	R	F	P	R	F	F
e2e-coref	59.4	57.1	58.2	62.3	59.4	60.8	59.8	62.2	61.0	60.0
c2f-coref	64.7	66.5	65.6	67.2	65.9	66.5	65.3	68.7	67.0	66.4
CR-BERT	74.3	71.8	73.0	74.5	71.9	73.2	74.7	72.5	73.6	73.3
SpanBERT	68.4	72.2	70.2	71.6	69.4	70.5	73.4	71.2	72.3	71.0
WL-COREF	73.1	71.6	72.3	72.3	70.8	71.5	70.6	74.9	72.7	72.2
Link-Append	63.4	62.7	63.0	65.5	68.1	66.8	67.8	64.2	66.0	65.3
Fastcoref	76.8	78.3	77.5	77.3	74.6	76.0	78.6	76.5	77.5	77.0
GPT-4	82.4	83.9	83.1	84.7	82.6	83.6	81.5	79.6	80.7	82.5
Human	93.6	89.1	91.3	94.0	90.3	92.1	93.2	88.3	90.7	91.4

Table 5: Overall Performance on NovelCR-en (%).

Methods	B3			MUC			CEAFe			CoNLL
	P	R	F	P	R	F	P	R	F	F
e2e-coref	53.2	62.3	57.4	58.9	57.2	58.0	59.4	56.8	58.1	57.8
c2f-coref	58.3	68.8	63.1	60.3	66.9	63.4	67.3	64.8	66.0	64.2
CR-BERT	62.7	70.8	66.5	68.6	67.2	67.9	63.9	69.1	66.4	67.0
SpanBERT	68.1	67.4	67.7	72.4	65.8	69.0	67.4	70.2	68.8	68.5
WL-COREF	60.7	63.3	62.0	64.7	62.2	63.4	68.5	63.7	66.0	63.8
Link-Append	58.9	67.2	62.8	63.0	66.7	64.8	65.4	67.1	66.2	64.6
Fastcoref	67.9	68.1	68.0	69.5	67.3	68.4	68.3	64.7	66.5	67.6
GPT-4	74.1	72.3	73.2	73.0	75.2	74.1	71.8	72.6	72.2	73.2
Human	96.3	85.1	90.4	94.3	86.8	90.4	95.4	86.2	90.6	90.5

Table 6: Overall Performance on NovelCR-zh (%).

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

Table 8: Error Analysis in NovelCR

8, 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 NovelCR is that existing CR models lack the common sense to discern the gender of the mention. For instance, in the second example in Table 8, 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 NovelCR 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 8, there are numerous 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* rather than *Emma, Mason, Oliver, Ava*.

6 Conclusion

In this paper, we propose a large-scale bilingual dataset, NovelCR, focusing on long-span coreference resolution. NovelCR features a substantial dataset size, with a total of 148k mentions and 128k coreferences in NovelCR-en and 311k mentions and 273k coreferences in NovelCR-zh. Moreover, NovelCR contains a large number of lengthy coreferences. Extensive experiments on NovelCR demonstrate that the performance of the state-of-the-art baselines cannot catch up with human beings, showing that NovelCR remains an unresolved challenge.

6.1 Limitations

While we have made significant strides in constructing a high-quality CR dataset, it is important to acknowledge the limitations that may affect the interpretation and generalizability of our work.

Few Entity Types As outlined in the introduction, we concentrate on resolving coreferences of characters in the novel. This is a double-edged choice. On one side, it enables NovelCR to contain abundant long-span coreferences. On the other side, it restricts NovelCR's entity type exclusively to persons, omitting locations, organizations, times, events, and others. The restricted entity type compromises NovelCR's diversity and constrains NovelCR's applicability across diverse natural language understanding contexts. Future endeavors could explore extracting more long-span coreferences for additional entity types from varied data sources.

Missing Mention Annotation Referring back to the mention detection process in Section 3.2, we initially use tools and models to pre-label mentions, and then ask annotators to manually remove invalid mentions. This two-step annotation can ensure the quality of the candidate mentions but also results in overlooking certain mentions. To measure the magnitude of this problem, we manually evaluate a sample of 200 documents in NovelCR-en and in NovelCR-zh respectively, revealing a missing rate of 0.9% and 1.1%. Given the substantial size of our dataset, a minor degree of missing labeling is acceptable.

6.2 Ethics Statement

This work uses publicly available novels as dataset annotation sources, and we respect and abide by their licenses and agreements.

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.

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.

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.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. *arXiv preprint arXiv:2106.01210*.

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*.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

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*.

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.

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

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*.

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.

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.

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.

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*.

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.

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.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

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.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2020. DWIE: an entity-centric dataset for multi-task document-level information extraction. *CoRR*, abs/2009.12626.

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.

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*.

A Annotation Guideline of Mention Verification

In mention verification, the annotation instructions are outlined below.

Please read the novel chapter, and finish the two tasks: (1)delete invalid mentions, and (2)re-edit the mention boundaries. The second task can only be started after the first task is completed.

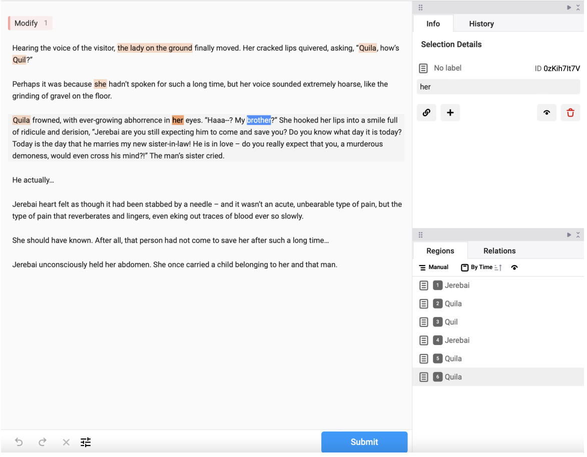
When deleting invalid mentions, you should remove mentions that do not refer to the person entities, such as *the bank* and *this beautiful knife*. Note that dependent personal pronouns should also be deleted. For instance, *her* in *her split lips* is also an invalid mention since it functions as a modifier of *lips*. To delete invalid mentions, click the mention to highlight it and then click the *Delete* button.

When re-editing the boundary of the mentions, we follow the maximum span principle. This means that you should identify the longest string representing the mention. For instance, in the sentence *the sad man is looking for his wife*, you should annotate the mention as *the sad man* rather than just *man*. If the mention does not meet the maximum span criteria, you should drag the gray border to correct the boundaries of the mention. Please do nothing if no mistakes are found. When you have completed all annotations on a page, remember to click the *Submit* button to store the annotation results. We assure you that all annotations will be utilized solely for research purposes.

B Annotation Guideline of Coreference Identification

As shown in Figure 6, annotates need to match entities and mentions. The annotation instructions are as follows.

Please read the novel chapter and match each mention to the entity it refers to. We recommend reading the entire chapter before starting any annotations, as coreference resolution relies on a broad context span understanding. We already highlight mentions in grey and list the entity options at the top of the chapter. All you need to do is click the mention and then the entity it refers to to match them. If the mention doesn't refer to any entities, you can simply click on the *None* option. When you have completed all annotations on a page, remember to click the *Submit* button to store the annotation results. We promise that all annotations will be used for research purposes.



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.

Quila 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.

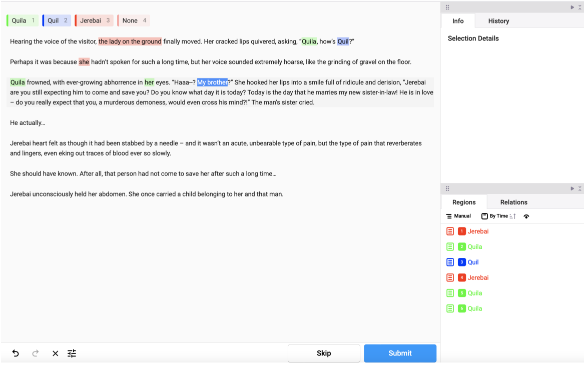
He actually...

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 5: Screenshot of Mention Verification



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.

Quila 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 6: Screenshot of Coreference Identification

C Prompt for Mention Detection.

We leverage direct prompting, chain-of-thought (CoT) prompting, and ReAct prompting, respectively, to detect mentions from the novel chapter.

Prompt 1 (direct prompting)

Question:Please find all words or phrases that may refer to a person in the following passage:[novel chapter].

Prompt 2 (CoT prompting)

Question: Please find all words or phrases that may refer to a person in the following passage:[novel chapter]

Thought:The possible candidates include pronouns, human names, and noun phrases. pronouns could be he, she, him, her, their, and them. Noun phrases could be nouns like man, woman, girl, and boy with their adjectives. Human names can be discovered using the rules of different languages.

Prompt 3 (ReAct prompting)

Tools:NER(p) takes a passage as parameter and returns Named Entities that belong to human beings. PosTag(p) takes a passage as parameter and returns all pronouns and nouns phrases.

Question: Please find all words or phrases that may refer to a person in the following passage: [novel chapter]

Thought:The possible candidates include pronouns, human names, and noun phrases. Human names can be found by NER first.

Action:NER

Observation: [entities]

Thought:Then noun phrases and pronouns can be found by PosTag.

Action:PosTag

Observation: [mentions]