CW3NE:A Genre-oriented Corpus for Nested Named Entity Recognition in Chinese Web Novels

Anonymous ACL submission

Abstract

Named entities are important to understand literary works, which emphasize characters, plots and environment. The research on named entity recognition (NER), especially nested named entity recognition in literary domain is still insufficient partly due to lack of enough annotated data. To address this issue, we construct the first Genre-oriented Corpus for **N**ested **N**amed **E**ntity Recognition in **C**hinese **W**eb **N**ovels, namely **CW3NE**, comprising 400 chapters totaling 1,214,283 tokens under two genres, XuanHuan (Eastern Fantasy) and History. Based on the corpus, we make a deep analysis of the distribution of different types of entities, including person, location and organization. We also make comparison of nesting patterns of nested entities between CW3NE and the English corpus LitBank. Even both belong to literary domain, entities in different genres share few overlaps, making genre adaptation of NER a hard problem. We provide several baseline NER methods and experimental results show that large language model based methods perform poorer than well designed small language model based method. Performance drops sharply on nested NER for all baseline methods, indicating the great challenge posed by the nested named entities. Genre adaptation also results in great performance drop especially on location and organization entities. We will release our corpus to promote research on literary NER. [1]

1 Introduction

Computational literature, an interdisciplinary field combining natural language processing (NLP) and literary studies, aims to leverage structured literary information for answering queries about entities within literature (Jia et al., 2020b). A critical component of literary analysis is the extraction of entities from texts. Named entity recognition (NER) (Sang and De Meulder, 2003), a fundamental task in information extraction (Cowie and Lehnert, 1996), identifies entities within sentences such as persons, locations, and organizations. This task serves as a cornerstone for various downstream NLP applications, including relation extraction (Zhou et al., 2005), event extraction (Hogenboom et al., 2011), and coreference resolution (Sukthanker et al., 2020).

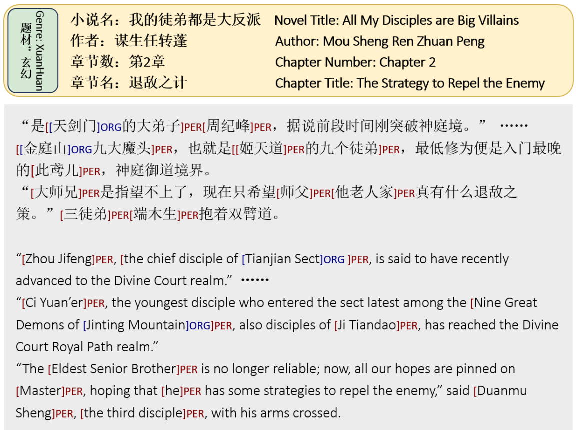

Figure 1: Dataset Examples: Corresponding Metadata Above, Novel Texts Below.

In Chinese literary research, the focus is increasingly shifting towards analyzing content-rich literary works. However, the critical aspect of nested entities and their influence has been overlooked. Within the task of NER, the predominant emphasis has been on general domains (Bamman et al., 2019), both in methodological approaches and dataset composition. While this focus has been comprehensive for general applications, it falls short in addressing literary analysis.

Specifically, models trained on datasets from general domains exhibit suboptimal performance when applied to the literary domain(Augenstein et al., 2017), highlighting a mismatch between general NER models and the unique requirements of literary texts. Even within the literary domain, different genres pose significant challenges for entity

[1] https://anonymous.4open.science/r/CW3NE/

Dataset	Language	Genre	Nested	Genre-oriented	Entity Types
SanWen (2017)	Chinese	Essay	✗	✗	7
LitBank (2019)	English	Novels, Stories	✓	✗	6
Books (2020a)	Chinese	XuanHuan	✗	✗	6
JinYong (2021)	Chinese	Martial	✗	✗	4
QiDian (2023)	Chinese	Multi-genres	✗	✓	3
CW3NE	Chinese	XuanHuan, History	✓	✓	3

Table 1: Statistics of Literary NER Datasets. "Nested" indicates nested annotation. Unless otherwise specified, the term "Genre" refers to the category of novels.

recognition. Moreover,the existing NER datasets in the literary domain do not sufficiently address the distinct characteristics of Chinese web novels, revealing a gap in this specific genre.

To address these challenges, we introduce **CW3NE**: A Genre-oriented Corpus for **N**ested **N**amed **E**ntity Recognition in **C**hinese **W**eb **N**ovels. Our comprehensive dataset tackles the critical issues in Chinese web novel NER, particularly the scarcity of extensive web novel data and analysis of the complexity of nested entity structures. To facilitate a deeper understanding of character dynamics, we include metadata annotations pertinent to character analysis. The organization and details of our dataset are depicted in Figure 1.

The contributions of this paper can be summarized as follows:

- **We annotate a genre-oriented corpus for nested named entity recognition in Chinese web novels.** We develop a genre-oriented dataset for nested NER derived from 400 chapters of Chinese web novels, containing over 1.2 million tokens. This dataset fills a critical gap in Chinese literary resources.

- **We unveil the nesting patterns of the nested entities and entity differences between genres.** We perform a comprehensive analysis, highlighting the distinct challenges and characteristics of nested entities in our dataset compared to existing ones, emphasizing the primary difficulties in entity recognition, especially genre adaptation.

- **Experiments reveal the challenges of nested entity recognition and cross-genre difficulties.** Our extensive experiments reveal challenges in NER for Chinese web novels. Current methods, including large language models, need further refinement for effective nested entity recognition. Cross-genre experiments highlight the complexity, emphasizing the need for more effective approaches.

2 Related Work

Named Entity Recognition (Sang and De Meulder, 2003) stands as a foundational task in information extraction, aiming to identify entities such as persons, locations, and organizations within the text. This process is crucial for downstream tasks like automated question answering, machine translation, and text analysis. However, the application of NER to the literary domain remains a challenging endeavor (Jia et al., 2021). The primary challenge stems from the scarcity of annotated corpora, with scholars such as Santana et al. (2023) emphasizing the pivotal role and intricacy of the data annotation process during the training phase. Despite the efforts of researchers (Bamman et al., 2019; Jia et al., 2021, 2020a; Zhao et al., 2023) in annotating novel datasets, a deficiency persists in high-quality Chinese web novel data for facilitating character recognition.

In our study, we provide a comprehensive overview of existing NER datasets within the literature. Table 1 presents detailed information, encompassing language, data source, entity categories, nested entity structures and other relevant characteristics.

In response to the scarcity of datasets in the literary domain, Bamman et al. (2019) curated a collection of 100 English novels from Project Gutenberg[2], annotating nested entity information specifically tailored to the literary context. Additionally, their cross-domain experiments with ACE(Augenstein et al., 2017) revealed pronounced disparities in en-

[2] https://gutenberg.org/

tity distribution between the literary and news domains. Moreover, a comparative analysis of gender across both domains served to underscore the literary domain's particular emphasis on the distinctive traits of people.

In the Chinese context, Xu et al. (2017) targeted the difficulties faced in Chinese literary works and conducted detailed entity and relationship annotation on 726 articles, to some extent addressing the problem of dataset scarcity. Jia et al. (2021), starting with Jin Yong's novels, annotated named entities in over 1.8 million words across two novels, totaling more than 50,000 annotations for 4 entity categories. Simultaneously, they conducted thorough analysis and experiments on the dataset, providing a paradigm for subsequent literary research.

The Books(Jia et al., 2020a) dataset is sourced from Chinese web[3] novels. The entity types encompass PER (person), LOC (location) and ORG (organization), WEP (weapons), TIT (titles), and KUN (kung fu). Additionally, Zhao et al. (2023) established the largest Chinese multi-genre literary NER corpus, which includes 260 Chinese novels across 13 different genres.

3 Corpus Construction

3.1 Data Collection and Preprocessing

We conduct web scraping activities targeting the largest Chinese web novel platform, QiDian Chinese Website [4], to collect a dataset of 40 popular web novels. Each chosen work comprises its initial 10 chapters and falls within the XuanHuan (XuanHuan blends Chinese folklore, mythology, and martial arts, whereas Western fantasy typically draws from European myths and medieval elements.) or History genre, including those adapted into cartoons or TV dramas. Furthermore, we extract metadata such as novel names, chapter titles, genre information, and author names to facilitate future literary analyses. It is important to note that all the collected data is openly accessible for research purposes.

3.2 Annotation Principles

In our study, we conduct annotations on the acquired dataset to identify nested entities, classifying them into distinct types including person (PER), location (LOC), and organization (ORG). Below, we provide a comprehensive overview of the annotation specifications.

3.2.1 Person Entities

Person entities refer to characters depicted within novels, holding a paramount role within the narrative. Our annotation process specifically focuses on entities that represent individual characters or groups of characters, excluding personal pronouns. Entities identified during the annotation process that signify characters are annotated, encompassing specific types such as:

Person names “乐正东(Le Zhengdong)”

Ordinary nouns or character relationships “师兄(the Senior Brother)”, “兄弟姐妹(Siblings)”

Descriptive nouns indicate person entities “一个身穿绿色长裙的女人(A woman wearing a green dress)”

Non-human creature but capable of independent thought or dialogue “冰蚕(Ice silkworm)”, “兽王(king of beasts)”

In the Books dataset (Jia et al., 2020a), the "TIT" label for character titles has been relabeled as "PER" for consistency in our study. For instance, “温师兄(Senior Brother Wen)” is now labeled as "PER".

3.2.2 Location Entities

Location entities play a crucial role in novels by indicating the settings where the story takes place. They serve as auxiliary elements alongside person entities.

Physical location We identify and label physical locations or settings referenced in the text, excluding prepositions from annotation.

Entity indicates where the storyline takes place Additionally, certain narratives frequently unfold within buildings, which may be designated as FAC (facilities) in other annotation schemes. However, as these entities contribute to establishing the story's setting, we annotate them as location entities.

3.2.3 Organization Entities

In the context of web literature, identifying organizational entities proves challenging due to their often sparse and ambiguous. To precisely identify organizational entities, we exclusively label those with explicit and distinguishable hierarchical relationships.

[3] https://babelnovel.com/
[4] https://www.qidian.com/

Entity type	Examples
PER	乐正东,局长,父亲,大师兄, [[父亲]的兄弟姐妹]$_{PER}$, [[弯刀盟]首领] Le Zhengdong, the Director, the Father, the Senior Brother, [[the Father] 's Siblings], [the Leader of [the Curved Blade Alliance]]
LOC	中云市,综合大楼, [[拜月国]皇城], [[普利兹港]白玫瑰区] Zhongyun City, Comprehensive Building, [Imperial City of [the Baiyue Kingdom]], [White Rose District of [Puli Port]]
ORG	城建局,司礼监,红山学院, [[慕容风]的军队],[[天玄城]四大世家] Urban Construction Bureau, Ceremonial Directorate, Hongshan College, [[Murong Feng] 's army],[[Tianxuan City] Four Great Families]

Table 2: Entity Examples: To distinguish nested entities, we use different colors to represent various categories and separate them using brackets [ ]. Light red is used for PER (person), light green for LOC (location), and light purple for ORG (organization).

Organizational entities are inherently intertwined with personal entities. For instance, in the context of a family, the removal of personal entities renders the family structure incomplete in terms of personal representation. This concept extends to other organization entities such as factions and nations.

We conduct a comprehensive analysis of frequently occurring entities across various types. As demonstrated in Table 2.

3.3 Annotation Process

We annotate our dataset using the open-source tool Label-Studio(Tkachenko et al., 2020-2022), employing five expert web novel readers for labeling. The process lasts three months.

Manual Annotation The team leader first annotates a small subset to create guidelines. The annotation team then applied these guidelines uniformly. Cross-validation identifies discrepancies, and secondary reviews ensure precision and recall. If the F1 score is below 70, further annotation and review are conducted.

Verification After annotations, we analyze entity frequency in each chapter. Errors, such as incorrect entity boundaries (e.g., labeling “张三” as “张三道”), are manually corrected to ensure accuracy.

3.4 Inner-annotator Agreement

To ensure dataset quality, we assess annotation consistency by comparing multiple annotations from different annotators. Using the latest annotations as the reference, we calculate the F1 score, resulting in an overall consistency rate of 94.91%. This high consistency underscores the dataset's quality, with detailed rates for each entity type shown in Table 3.

	PER	LOC	ORG	All
IAA	0.9623	0.8765	0.9200	0.9491

Table 3: The results of Inter-annotator Agreement (IAA) for the dataset.

4 Corpus Analysis

4.1 Corpus Statistics

In the analysis of the dataset, it is evident that PER (person) is notably more prevalent, especially in the context of web novels. This prominence is discernible from Figure 2, where person entities account for a significant portion of the dataset, underscoring their pivotal role in narratives.

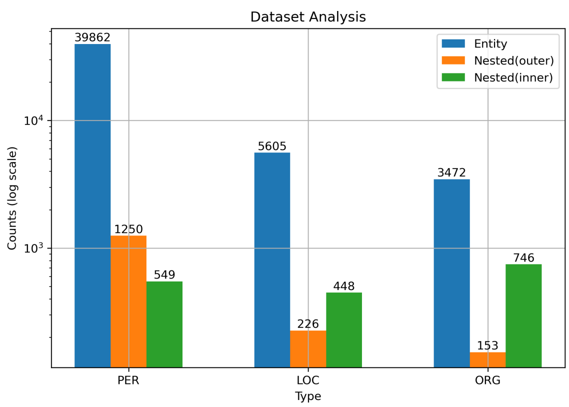

Figure 2: Statistical Data of Different Entity Types

Location entities, which frequently accompany person entities, delineate the background against which the stories take place, thereby enriching the narrative by setting the stage for the unfolding events. Organization entities, while less common, play a critical role in illustrating the affiliations and social structures within the narrative, often driving the plot forward.

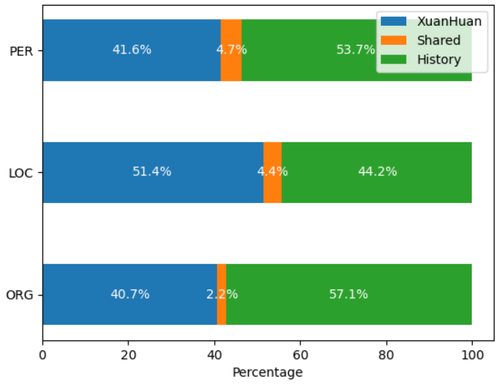

Figure 3: Distribution of entities across two different genres, with 'shared' indicating entities in both genres.

Further nested entities within web novels reveal a notable trend: person entities can most easily form nested structures, serving as the outer layer in these configurations. In contrast, locations and organizations forming nested entities are comparatively rarer, constituting about 20% of such structures, yet they are often nested in person entities. This pattern highlights the dominance of person entities in web novels, where they not only predominate in frequency but also in their ability to integrate other entity types within their nested structures.

In addition, we analyze the overlap of unique entities across different genres. Figure 3 demonstrates significant differences between the genres, with minimal shared entities, highlighting the substantial challenges in cross-genre recognition.

4.2 Analysis of Entity Nesting Patterns

We examine the distribution of nested entities in the dataset, focusing primarily on internal and external nested entities. The results are presented in Figure 4, where the vertical axis represents external nested entities and the horizontal axis represents internal entities.

In our dataset, individuals are often associated with their organizational affiliations. For instance, the title "大魏皇帝(Emperor of Da Wei)" signifies a person's role within the "大魏(Da Wei)" state organization. This results in numerous instances of organizational entities nested within person entities

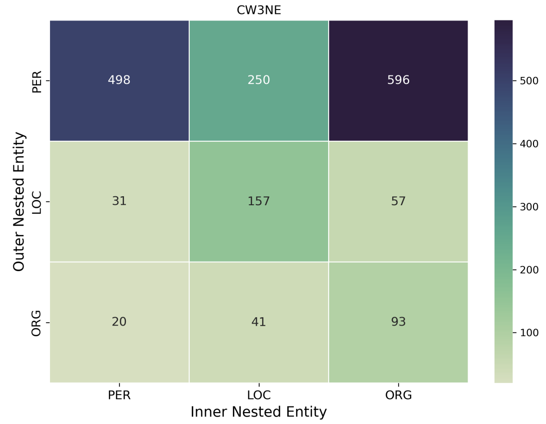

Figure 4: Distribution of Nested Entities.

in the CW3NE dataset, demonstrating the complexity of nested entity annotations in literary texts.

We conduct a detailed statistical analysis to investigate the origins and patterns of nested entities in both LitBank and CW3NE. Our findings show that LitBank, an English dataset, predominantly features nested entities with gender-specific markers such as "Mr." and "Mrs.," which highlight character gender identities. Detailed analysis figures are provided in the Appendix Figure 7.

In contrast, the Chinese dataset from CW3NE exhibits a higher incidence of nesting involving person and organization entities, reflecting inter-entity relationships. This discrepancy is attributed to linguistic differences between English and Chinese, presenting challenges for entity recognition in computational models.

5 Experiments

In the experimental section, we evaluate the dataset's quality and conduct a series of baseline models for comparison. Including **(1)** The state-of-the-art method, DiffusionNER(Shen et al., 2023), on commonly used datasets (ACE(Doddington et al., 2004; Walker et al., 2006), GENIA(Kim et al., 2003)) **(2)** and generative large language models, like ChatGPT[5], Baichuan2-7B(Baichuan, 2023).

5.1 Experimental Settings

5.1.1 Dataset Split

In our study, we follow the data partitioning methodology as delineated by Bamman et al. (2019). The segmentation of the dataset is based on novels, adopting an 8:1:1 split ratio.

[5]https://chat.openai.com/

model	F1-score			Overall		
	PER	LOC	ORG	P	R	F1
DiffusionNER	75.34	61.19	61.78	71.81	72.40	**72.10**
Baichuan2-7B(0-shot)	64.35	48.91	33.71	69.81	52.76	60.10
Baichuan2-7B(3-shot)	65.40	49.95	28.74	67.60	55.39	60.89
ChatGPT(0-shot)	19.19	33.58	18.23	68.08	13.05	21.90
ChatGPT(3-shot)	56.83	40.93	19.80	59.17	45.18	51.24

Table 4: Results of Named Entity Recognition. 0-shot and 3-shot represent the number of examples in the prompt.

	PER	LOC	ORG	#Sentences
Train	33,274	4,671	2,962	19,762
Valid	3,446	308	224	2,222
Test	3,142	626	286	1,822

Table 5: Distribution of the Partitioned Dataset.

This distribution allocates 32 novels to the training set, with 4 novels each dedicated to the validation and test sets. A comprehensive breakdown of this distribution, including detailed statistics, is presented in Table 5.

5.1.2 Baselines Details

Below are the settings for the experimental model.Detailed parameters are provided in the Appendix 13.

DiffusionNER DiffusionNER(Shen et al., 2023) represents a boundary-denoising model for NER, which uses BERT(Devlin et al., 2018) as the base model and demonstrates state-of-the-art performance across diverse datasets in general domains. We extend its application to the domain of literature. The experimental setup follows the configuration, with 30 epochs, a learning rate of 2e-05, and a batch size of 8.

Baichuan2 Baichuan 2 (Baichuan, 2023), the large language model from Baichuan Intelligence, is trained on a diverse corpus of 2.6 trillion high-quality tokens. We fine-tune Baichuan2-7B-Base using LoRA (Hu et al., 2021) with llama-factory (Zheng et al., 2024). The fine-tuning parameters are: batch size of 4, 3 epochs, and a rank of 8. We perform fine-tuning under two scenarios: 0-shot and 3-shot entity recognition.

ChatGPT Our research leveraged OpenAI's API to conduct experiments.[6] Specifically, we engaged in 0-shot experimentation utilizing a prompt structure composed of task definition, problem statement, annotation guidelines, and desired output format. The entity annotation standards we adopted are detailed in Section 3 .

[6]All experimental work was carried out using the API version available in March and April

5.2 Entity Recognition Results

In our experiments on the dataset (see Table 4), we observe that LLMs, including ChatGPT and Baichuan, perform approximately 10 percentage points lower in recognition accuracy compared to the fine-tuned state-of-the-art (SOTA) model.

A detailed analysis indicates that the main deficiency of LLMs relative to the SOTA model is their recall rate. However, LLMs show significant potential for improvement, especially when examples are incorporated into the prompt template, which notably enhances their recall rate. This enhancement is particularly pronounced in ChatGPT, which has not been fine-tuned.

Examining the overall F1 score, the 3-shot fine-tuning of Baichuan did not yield a significant improvement over the 0-shot approach, despite an increase in Recall by 2.63 points. Conversely, for ChatGPT, the inclusion of example prompt templates resulted in a substantial performance boost, with the F1 score rising from 21.9% to 51.24%. We conclude that including example prompts can significantly aid LLMs in understanding prompt information, thereby greatly enhancing their performance, especially for models lacking fine-tuning.

5.3 Cross-genre Novel Recognition Results

We conduct an analysis using two genres of novels by partitioning the dataset based on genre. As shown in Table 8, the results indicate a significant drop in model performance when the training and test sets belong to different genres, with a particularly notable 40 percentage point decrease in recognizing organization and location entities. This

Test→	XuanHuan				History			
Train↓	PER	LOC	ORG	micro-F1	PER	LOC	ORG	micro-F1
XuanHuan	75.09	52.99	47.10	70.85	59.60	15.95	7.66	52.24
History	57.19	17.28	13.33	50.19	68.00	59.41	58.18	65.20

Table 6: Cross-Genre Recognition Results.

mark performance decline highlights the model's limitations in handling cross-genre data.

Overall, our findings underscore the necessity for models to possess enhanced generalization capabilities and robustness to effectively manage the variations across multiple genres.

6 Analysis

6.1 Recognition Results of Each Novel

Figure 5 presents the recognition results for each novel in the test set. We analyze four novels: "完美世界(Perfect World)" and "将夜(Jiang Ye)" from the Xuanhuan genre, and "庆余年(Qing Yu Nian)" and "赘婿(Zhui Xu)" from the history genre.

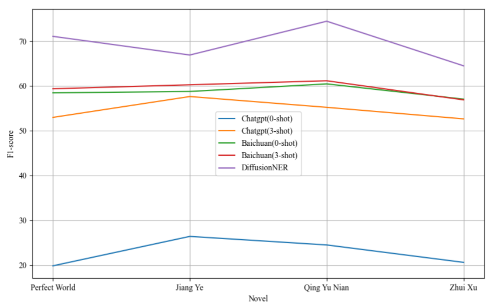

Figure 5: Recognition results of four novels in the test dataset.

Baichuan display stable recognition performance with minimal variation across different novels. In contrast, the 0-shot version of ChatGPT show significant sensitivity to novel differences, which is reduced in the 3-shot version with the inclusion of examples. DiffusionNER, despite achieving the best overall results, exhibit considerable fluctuations in character recognition across different novels, likely due to its model characteristics.

Consistent trends are observed in the results from the same model across different experimental settings. We hypothesize that this phenomenon is caused by the varying difficulty levels of the different novels.

6.2 Recognition Results of Nested and Non-nested Entities

To evaluate the accuracy of nested entity recognition, we analyze the recognition results for two types of entities. A nested entity is considered correctly identified only when both the inner and outer entities are accurately recognized by the model.

As shown in Figure 6, the accuracy of nested entity recognition in novels is significantly lower than that of non-nested entities. Often, the model only partially identifies nested entities, such as recognizing the outer entity while missing the inner one. This discrepancy is partly due to the complexity of nested entities and the relatively sparse data compared to non-nested entities. This highlights the need for models to improve their ability to recognize sparse and challenging data in the context of web novels.

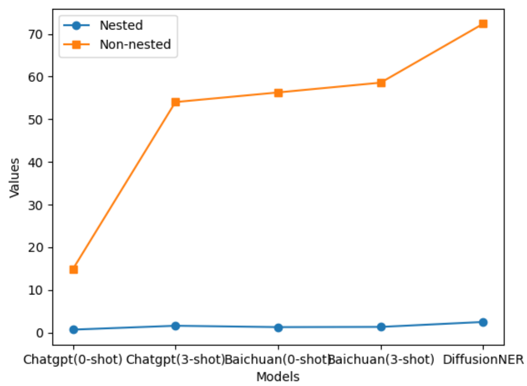

Figure 6: Results of Nested and Non-Nested Entities.

6.3 Recognition Results of IV and OOV Entities

For the cross-genre experiments, we observed a significant decline in performance. This is primarily due to the substantial differences between entities in the two genres.

As shown in Section 4.1, few overlapping entities exist between the genres. Our analysis of

Case1	“族长，我们已经有些日子没有进山了。”就在这时，一个雄壮的成年男子走进院中，他是狩猎队伍的头领，也将是石村的下任族长。 "Patriarch, we have not gone into the mountains for some days now."Just then, a robust adult male entered the courtyard. He is the leader of the hunting team and will also be the next Patriarch of the Stone Village.
Ground Truth	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, 一个雄壮的成年男子 **LOC:** 院中 **ORG:** 狩猎队伍, 石村 **PER:** Patriarch, the next Patriarch of the Stone Village, the leader of the hunting team, a robust adult male **LOC:** the courtyard **ORG:** the hunting team, the Stone Village
DiffusionNER	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, 一个雄壮的成年男子 **LOC:** 院中, 石村× **ORG:** 狩猎队伍, ○
ChatGPT(0-shot)	**PER:** 族长, ○, ○, 成年男子× **LOC:** 山×, 院中, 石村× **ORG:** 狩猎队伍, ○
ChatGPT(3-shot)	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, ○ **LOC:** ○ **ORG:** ○, ○
Baichuan2(0-shot)	**PER:** 族长, ○, ○ 成年男子× **LOC:** ○ **ORG:** ○, 石村
Baichuan2(3-shot)	**PER:** 族长, 石村的下任族长, 狩猎队伍的头领, ○ **LOC:** ○ **ORG:** ○, ○

Table 7: Case study. × indicates recognition errors, ○ indicates unrecognized entities.Light red highlights misclassifications, light green indicates inaccuracies of entity boundaries, and light blue marks non-entities.

	IV	OOV
Xuanhuan→History	68.38	39.78
History→Xuanhuan	58.46	31.21

Table 8: Analysis of IV and OOV entities: For the XuanHuan, the IV:OOV ratio is 117:606, while for the History, it is 65:410. The statistics are based on deduplicated entities.

in-vocabulary (IV) and out-of-vocabulary (OOV) entities revealed that the poor recognition of OOV entities is the main factor. Since OOV entities constitute a larger proportion compared to IV entities, this results in the overall poor performance of cross-genre recognition.

6.4 Case Study

To analyze the model's recognition outcomes and identify the dataset's vulnerabilities, two instances of model recognition have been meticulously chosen. Errors are categorized into these distinct types: 1) misclassifications regarding entity types, 2) inaccuracies in identifying entity boundaries, 3)misidentification of non-entities, and 4)unrecognized entities in the ground truth.

In the example, both the 0-shot ChatGPT and Baichuan incorrectly recognized the entity “石村的下任族长(the next Patriarch of the Stone Village)” as “族长(Patriarch)”. However, the 3-shot models accurately identified the entity, suggesting that providing examples can significantly enhance the comprehension capabilities of LLMs.

However, it is also important to acknowledge that examples can sometimes introduce interference. For instance, while the 0-shot Baichuan correctly identified the organizational entity “石村(the Stone Village)”, the 3-shot models failed to do so after examples were provided. This highlights the potential trade-offs in model performance when using example-based prompting.

7 Conclusion

We present the largest Chinese nested entity annotation dataset in literary domain, comprising 1.2 million tokens across 400 chapters in XuanHuan and History genres. Our analysis reveals the distribution and origin of nested entities in web novels. A series of methods are implemented to assess the quality of the CW3NE. Experimental results reveal the need for further enhancement of existing methods within the literary domain and cross-genres recognition.

In the future, our work will involve continued annotation for coreference resolution and entity relationships on this corpus, facilitating a more comprehensive analysis of literature. Furthermore, we aim to incorporate additional literary elements to augment the model's effectiveness.

Limitations

To begin with, due to time and cost constraints, our annotated dataset is limited to two genres: History and Xuanhuan. It does not provide a comprehensive coverage of the various categories within online novels. Additionally, our annotations only involve three basic entities. However, given the diverse nature of entity types across different novel genres, a more comprehensive and detailed analysis is required to design a dataset that includes a broader range of entities.

Furthermore, our dataset is not free from noise. While multiple rounds of iterative annotation have improved data quality, it is undeniable that some annotation errors may exist due to personal biases in understanding. We aim to further optimize the dataset in future work.

Ethics Statement

The entirety of the work presented in this paper adheres to ethical standards.

References

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

Jim Cowie and Wendy Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020a. Entity enhanced bert pre-training for chinese ner. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6384–6396.

Yuxiang Jia, Rui Chao, Hongying Zan, Huayi Dou, Shuai Cao, and Shuo Xu. 2021. Document-level literary named entity recognition. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 600–611.

Yuxiang Jia, Lu Wang, Pengcheng Liu, Qian Wang, Yue Zhang, and Hongying Zan. 2020b. Distributed representation of fictional characters and its applications. *Journal of Chinese Information Science*, 34(12):92–99.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Diffusionner: Boundary diffusion for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3875–3890.

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

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*.

Hanjie Zhao, Jinge Xie, Yuchen Yan, Yuxiang Jia, Yawen Ye, and Hongying Zan. 2023. A corpus for named entity recognition in Chinese novels with multi-genres. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 398–405, Hong Kong, China. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 427–434.

A Appendix

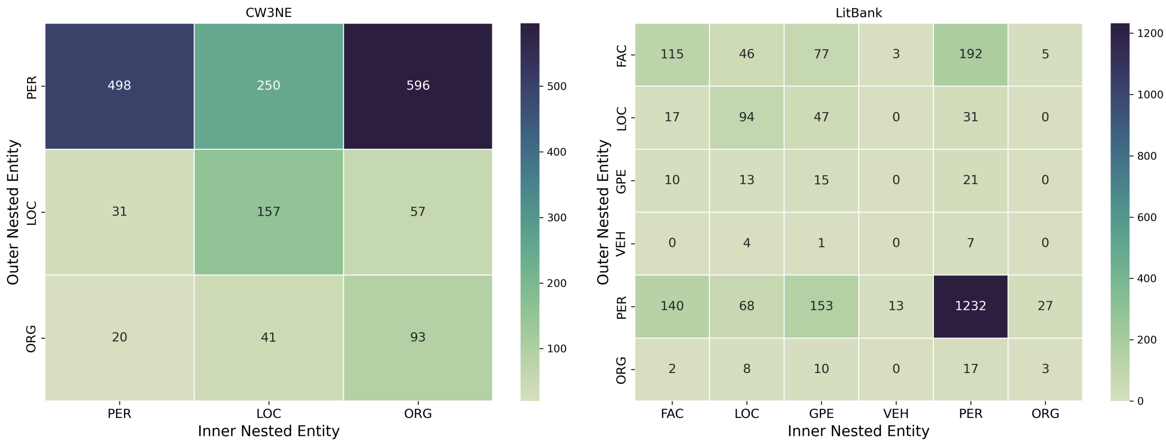

Figure 7: Distribution of Nested Entities and Comparison with LitBank.

Novel Name	Genre	Tokens	PER	LOC	ORG
斗罗大陆II绝世唐门	玄幻/XuanHuan	75124	1797	299	273
斗罗大陆IV终极斗罗	玄幻/XuanHuan	24743	509	61	63
斗罗大陆	玄幻/XuanHuan	19348	484	82	73
斗罗大陆III龙王传说	玄幻/XuanHuan	21143	574	61	44
斗破苍穹	玄幻/XuanHuan	28567	925	44	54
诡秘之主	玄幻/XuanHuan	34557	722	127	83
神墓	玄幻/XuanHuan	67893	2253	266	58
我的徒弟都是大反派	玄幻/XuanHuan	23673	1089	65	36
武动乾坤	玄幻/XuanHuan	28472	770	125	60
武神	玄幻/XuanHuan	31232	817	153	24
雪鹰领主	玄幻/XuanHuan	27854	1092	186	103
一世之尊	玄幻/XuanHuan	35780	1294	148	238
圣墟	玄幻/XuanHuan	29551	438	354	5
天道图书馆	玄幻/XuanHuan	28305	1014	65	100
天域苍穹	玄幻/XuanHuan	32548	705	164	39
完美世界	玄幻/XuanHuan	24735	558	152	9
万界天尊	玄幻/XuanHuan	18082	585	219	66
大主宰	玄幻/XuanHuan	31213	942	200	238
将夜	玄幻/XuanHuan	30622	593	207	89
牧神记	玄幻/XuanHuan	29038	904	92	3
秦吏	历史/History	27586	1076	207	112
庆余年	历史/History	23824	877	135	28
赘婿	历史/History	43363	1114	132	160
北宋大丈夫	历史/History	22000	984	125	59
回到明朝当王爷	历史/History	37735	1314	76	63
大汉帝国风云录	历史/History	35486	1425	158	93
大明最后一个狠人	历史/History	25279	1228	172	157
大魏宫廷	历史/History	37738	1756	135	351
带着仓库到大明	历史/History	22196	1005	140	71
汉乡	历史/History	29778	779	62	31
极品家丁	历史/History	22358	938	91	62
明朝败家子	历史/History	23905	990	75	90
明天下	历史/History	33053	1054	177	30
权柄	历史/History	24990	1027	72	109
如意小郎君	历史/History	29904	935	137	32
神话版三国	历史/History	21662	1064	88	55
时光之心	历史/History	23996	785	227	154
唐砖	历史/History	31166	971	149	38
小阁老	历史/History	23463	1053	57	38
医统江山	历史/History	32321	1422	120	81

Table 9: Detailed Statistics for Each Novel

0-shot	"对于给定的中文小说文本，任务的目标是识别并抽取出与小说相关的实体，并将他们归类到预先定义好的类别。小说文本命名实体划分为三大类，包括：人物(per)，地点(loc)，组织(org)。 命名实体标注的基本原则是： 1.人物(per)：指代单个人物或者人物集合的实体（代词你，我，他不标注）。 2.地点(loc)：指代存在的物理地点或者故事情节发生地（介词不标注）。 3.组织(org)：指示组织机构的实体（组织内要有明确的可区分的上下级关系）。 4.实体指向原则：只有在名词或者名词短语确定指向一个实体时标注，否则不标注。 5.实体边界原则：指向特定实体的最短名词或者名词短语。 输出格式如下，输出实体类型对应的实体使用'、'隔开，若无对应实体则输出无： 人物(per): 地点(loc): 组织(org): 下面是小说文本:"
3-shot	"对于给定的中文小说文本，任务的目标是识别并抽取出与小说相关的实体，并将他们归类到预先定义好的类别。小说文本命名实体划分为三大类，包括：人物(per)，地点(loc)，组织(org)。 命名实体标注的基本原则是： 1.人物(per)：指代单个人物或者人物集合的实体（代词你，我，他不标注）。 2.地点(loc)：指代存在的物理地点或者故事情节发生地（介词不标注）。 3.组织(org)：指示组织机构的实体（组织内要有明确的可区分的上下级关系）。 4.实体指向原则：只有在名词或者名词短语确定指向一个实体时标注，否则不标注。 5.实体边界原则：指向特定实体的最短名词或者名词短语。 实体识别的样例如下： 例1： 文本："父亲的兄弟姐妹" 实体： 人物(per):父亲、父亲的兄弟姐妹 地点(loc): 组织(org): 例2： 文本："城建局的局长乐正东" 实体： 人物(per):城建局的局长、乐正东 地点(loc): 组织(org):城建局 例3： 文本："小心翼翼的在树林中前行；办公室内很安静" 实体： 人物(per): 地点(loc):树林、办公室 组织(org): 输出格式如下，输出实体类型对应的实体使用'、'隔开，若无对应实体则输出无： 人物(per): 地点(loc): 组织(org): 下面是小说文本:"

Table 10: Prompt Detail.

Parameter	Value
base model	Chinese-bert-wwm-ext
batch_size	8
epochs	30
lr	2e-05
lr_warmup	0.1
weight_decay	0.01
max_grad_norm	1.0

Table 11: DiffusionNER Parameters

Parameter	Value
base model	gpt-3.5-turbo-16k
max_tokens	5000
temperature	1.0

Table 12: ChatGPT Parameters

Parameter	Value
base model	Baichuan2-7B-Base
finetuning_type	lora
batch_size	4
epochs	3
lora rank	8
warmup	0.1

Table 13: Baichuan Parameters

Case	按照宁毅之前的计划，原本是打算在外面跑一圈之后直接去豫山书院的...那是见过了几面的秦老家的小妾。 Following Ning Yi's initial plan, he intended to proceed directly to Yushan Academy after a brief excursion outside. This was Qin Lao's concubine, whom he had encountered on several occasions.
Ground Truth	**PER:** 宁毅, 秦老, 秦老家的小妾 **LOC:** **ORG:** 豫山书院 **PER:** Ning Yi, Qin Lao, Qin Lao's concubine **LOC:** **ORG:** Yushan Academy
DiffusionNER	**PER:** 宁毅, ◯, 小妾× **LOC:** **ORG:** 豫山书院
ChatGPT(0-shot)	**PER:** 宁毅, ◯, 秦老家的小妾 **LOC:** 外面× 豫山书院× **ORG:** ◯
ChatGPT(3-shot)	**PER:** 宁毅, ◯, 秦老家的小妾 **LOC:** 豫山书院× **ORG:** ◯
Baichuan2(0-shot)	**PER:** 宁毅, ◯, 秦老家的小妾 **LOC:** 豫山书院× **ORG:** ◯
Baichuan2(3-shot)	**PER:** 宁毅, ◯, 秦老家的小妾 **LOC:** **ORG:** 豫山书院

Table 14: Case study. × indicates recognition errors, ◯ indicates unrecognized entities.Light red highlights misclassifications, light green indicates inaccuracies of entity boundaries, and light blue marks non-entities.