A Discourse-Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text

Jingjing Xu*, Ji Wen*, Xu Sun, Qi Su
MOE Key Laboratory of Computational Linguistics, Peking University
School of Electronics Engineering and Computer Science, Peking University
{jingjingxu,wenjics,xusun,sukia}@pku.edu.cn

Abstract

Named Entity Recognition and Relation Extraction for Chinese literature text is regarded as the highly difficult problem, partially because of the lack of tagging sets. In this paper, we build a discourse-level dataset from hundreds of Chinese literature articles for improving this task. To build a high quality dataset, we propose two tagging methods to solve the problem of data inconsistency, including a heuristic tagging method and a machine auxiliary tagging method. Based on this corpus, we also introduce several widely used models to conduct experiments. Experimental results not only show the usefulness of the proposed dataset, but also provide baselines for further research. The dataset is available at https://github.com/lancopku/Chinese-Literature-NER-RE-Dataset.

Keywords: Chinese Literature Text, Named Entity Recognition, Relation Extraction

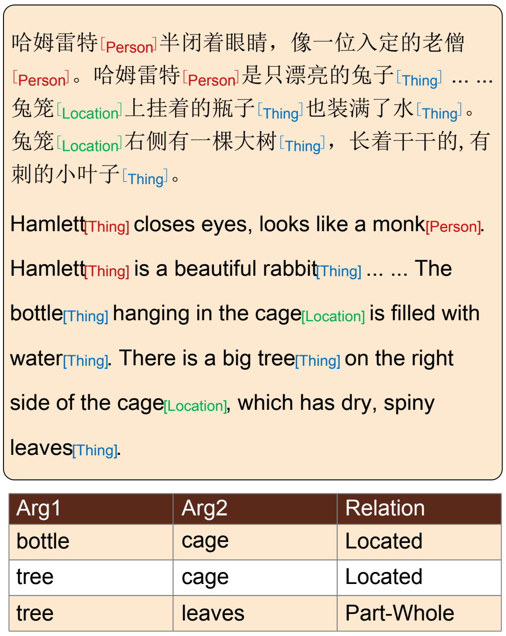
1. Introduction

Recent researches on Named Entity Recognition (NER) (Lin and Wu, 2009; Collobert et al., 2011; Huang et al., 2015) and Relation Extraction (RE) (Kambhatla, 2004; Zeng et al., 2014; Nguyen and Grishman, 2015) focused on news articles and achieved the promising performance. However, for a complex but important work, Chinese literature, this task becomes more difficult due to the lack of datasets. Thus, in this paper, we build a NER and RE dataset from hundreds of Chinese literature articles. Unlike previous sentence-level datasets where sentences are independent with each other, we build a discourse-level dataset where sentences from the same passage provide the additional context information.

However, tagging entities and relations in Chinese literature text is more difficult than that in traditional datasets which have simple entity classes and explicit relationships. Various rhetorical devices pose great challenges for building a high-consistency dataset. A simple example of personification is shown in Figure 1. “Hamlett” is a person name but refers to a rabbit. Some annotators label it with a “Person” tag and another annotators label it with a “Thing” tag. Thus, the major difficulty lies in how to handle massive ambiguous cases to ensure data consistency.

In this paper, we propose two methods to solve this problem. On one hand, we define several generic disambiguating rules to deal with the most common cases. On the other hand, since these heuristic rules are too generic to handle all ambiguous cases, we also introduce a machine auxiliary tagging method which uses the annotation standards learned from the subset of the corpus to predict labels on the rest data. Annotators just care about the cases where the predicted labels are different with the gold labels, which significantly reduces annotators’ efforts.

Overall speaking, we manually annotate 726 articles, 29,096 sentences and over 100,000 characters in total, which is accomplished in 300 person-hours spread across 5 people and three months.



哈姆雷特[Person]半闭着眼睛，像一位入定的老僧[Person]。哈姆雷特[Person]是只漂亮的兔子[Thing] ... ... 兔笼[Location]上挂着的瓶子[Thing]也装满了水[Thing]。兔笼[Location]右侧有一棵大树[Thing]，长着干干的, 有刺的小叶子[Thing]。

Hamlett[Thing] closes eyes, looks like a monk[Person].

Hamlett[Thing] is a beautiful rabbit[Thing] ... ... The bottle[Thing] hanging in the cage[Location] is filled with water[Thing]. There is a big tree[Thing] on the right side of the cage[Location], which has dry, spiny leaves[Thing].

Arg1	Arg2	Relation
bottle	cage	Located
tree	cage	Located
tree	leaves	Part-Whole

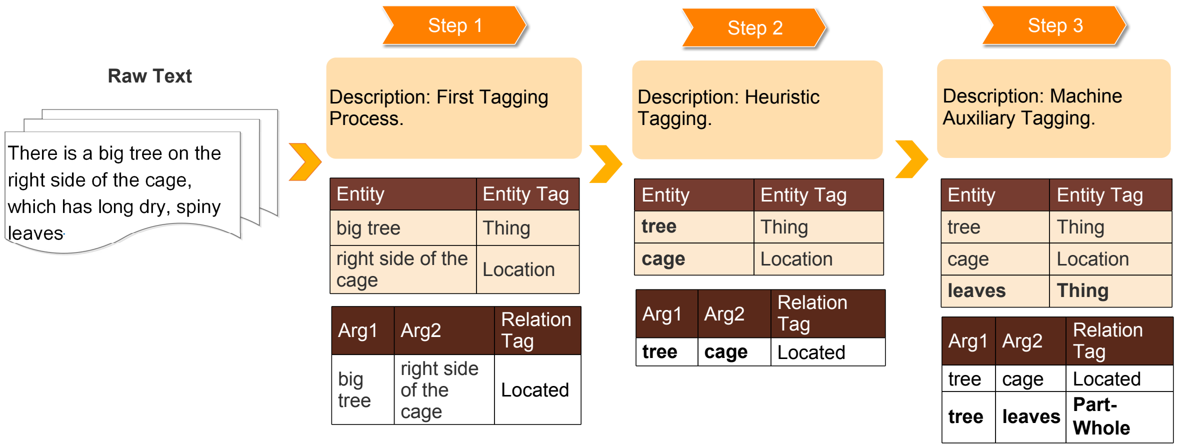
Figure 1: A tagging example. The top table describes the raw text and tagged entities which are shown in different color. The bottom table shows the tagged relations among these entities.

Based on this corpus, we also introduce some widely used models to conduct experiments. Experimental results not only show the usefulness of the proposed dataset, but also provide baselines for further research.

Our contributions are listed as follows:

- We provide a new dataset for joint learning of Named Entity Recognition and Relation Extraction for Chinese literature text.
- Unlike previous sentence-level datasets, the proposed dataset is based on the discourse level which provides additional context information.
- Based on this corpus, we introduce some widely used models to conduct experiments which can be used as baselines for further works.

[0]* Equal Contribution.



There is a big tree on the right side of the cage, which has long dry, spiny leaves

Description: First Tagging Process.

Entity	Entity Tag
big tree	Thing
right side of the cage	Location

Arg1	Arg2	Relation Tag
big tree	right side of the cage	Located

Description: Heuristic Tagging.

Entity	Entity Tag
tree	Thing
cage	Location

Arg1	Arg2	Relation Tag
tree	**cage**	Located

Description: Machine Auxiliary Tagging.

Entity	Entity Tag
tree	Thing
cage	Location
leaves	**Thing**

Arg1	Arg2	Relation Tag
tree	cage	Located
tree	**leaves**	**Part-Whole**

Figure 2: Illustration of the annotation process.

2. Related Work

Our work is related to recent works on Named Entity Recognition and Relation Extraction, which are briefly introduced as follows.

2.1. Named Entity Recognition

Named Entity Recognition has a long history in the field of natural language processing. One standard approach to NER is to regard the problem as a sequence labelling problem, where each word is assigned a tag, indicating whether the word belongs to part of any named entity or appears outside of all entities. Previous approaches used sequence labelling models, such as hidden Markov models (HMMs) (Zhou and Su, 2002), maximum entropy Markov models (MEMMs) (McCallum et al., 2000), structured perceptrons (Sun et al., 2009; Sun, 2015), as well as conditional random fields (Sun et al., 2014; Sun, 2014). While most research efforts exploited standard word-level features (Ratinov and Roth, 2009), more sophisticated features can also be used. Ling and Weld (2012) showed that using syntactic-level features from dependency structures in a CRFs-based model can lead to improved NER performance. Such dependency structures were also used in the work by Liu, Huang, and Zhu (2010).

More recently, neural networks have achieved promising results (Collobert et al., 2011; Huang et al., 2015; He and Sun, 2016; He and Sun, 2017). Collobert et al. (2011) used a CNN over a sequence of word embeddings with a CRF layer on top. Huang et al. (2015) presented a new CRF-LSTM models but using hand-crafted spelling features.

2.2. Relation Extraction

Relation Extraction plays an important role in NLP. Traditional methods (Kambhatla, 2004; Hendrickx et al., 2010) are usually feature-based models and their performance strongly depends on the quality of the extracted features. In kernel based methods (Bunescu and Mooney, 2005; Wang, 2008; Plank and Moschitti, 2013), similarity between two data samples is measured without explicit feature representation.

Recently, deep neural networks are widely used in relation classification. works are widely used in relation classification. Zeng et al. (2014) exploit a convolutional deep neural network to extract lexical and sentence level features. Zhang et al. (2015) used bidirectional long short-term memory networksto model the sentence with sequential information. Miwa et al. (2016) present an end-to-end neural model to extract entities and relations between them. Both word sequence and dependency tree information can be captured by stacking tree-structured LSTM-RNNs on sequential LSTM-RNNs. Cai et al. (2016) propose BRCNN to model the SDP, which can pick up bidirectional information with a combination of LSTM and CNN.

Datasets	Characters	Sentences	Articles
Training	1044966	24165	604
Validation	86454	1895	50
Testing	119647	2836	72

Table 1: Details of the proposed corpus.

3. Data Collections

We first obtain over 1,000 Chinese literature articles from the website and then filter, extract 726 articles. Too short or too noise articles are not included. Due to the difficulty of tagging Chinese literature text, we divide the annotation process into three steps. The detailed tagging process are shown in Figure 2

Step 1: First Tagging Process. We first attempt to annotate the raw articles based on defined entity and relation tags. In the tagging process, we find a problem of data inconsistency. To solve this problem, we design the next two steps.

Step 2: Heuristic Tagging Based on Generic disambiguating Rules. We design several generic disambiguation rules to ensure the consistency of annotation guidelines. For example, remove all adjective words and only tag "entity header" when tagging entities (e.g., change "a girl in red cloth" to "girl"). In this stage, we re-annotate all articles and correct all inconsistency entities and relations based on the heuristic rules.

Step 3: Machine Auxiliary Tagging. Even though the heuristic tagging process significantly improves dataset

Tags	Descriptions	Examples	%
Thing	Thing	苹果(apple)	36.45
Person	Person	李秋(Qiu Li)	33.10
Location	Location, country or city name	巴黎(France)	17.35
Time	Time related words	一天(one day)	7.31
Metric	Measurement related words	一升(1L)	3.73
Organization	Organization name	信息学报(Journal of Information Processing)	2.05

Table 2: The set of entity tags. The last column indicates each tag's relative frequency in the full annotated data.

Tags	Descriptions	Examples	%
Located	Be located in	幽兰(orchid)-山谷(valley)	37.43
Part-Whole	Be a part of	花(flower)-仙人掌(cactus)	23.76
Family	Family relationship	母亲(mother)-奶奶(grandmother)	10.25
General-Special	Generalization relationship	鱼(fish)-鲫鱼(carp)	6.99
Social	Be socially related	母亲(mother)-邻里(neighbour)	6.02
Ownership	Occupation relationship	村民(villager)-旧屋(house)	5.10
Use	Do something with	爷爷(grandfather)-毛笔(brush)	4.76
Create	Bring about something	男人(man)-陶器(pottery)	2.93
Near	A short distance away	山(hill)-县城(town)	2.76

Table 3: The set of relation tags. The last column indicates each tag's relative frequency in the full annotated data.

quality, it is too hard to handle all inconsistency cases based on limited heuristic rules. Thus, we introduce a machine auxiliary tagging method. The core idea is to train a model to learn annotation guidelines on the subset of the corpus and produce predicted tags on the rest data. The predicted tags are used to be compared with the gold tags to discovery inconsistent entities and relations, which largely reduce annotators' efforts. Specifically, we divide the corpus into 10 parts, and make predictions on each part of the corpus based on the model trained on the rest of the corpus. The model we used in this paper is CRF with a simple bigram feature template.

After all annotation steps, we also check all entities and relations to ensure the correctness of the dataset.

4. Data Properties

In this paper, we provide the standard splits which will be easier for benchmarking the related methods. The statistic of the proposed corpus is shown shown in Table 1. Next, we will describe the tagging set and annotation format in detail.

4.1. Tagging Set

We define 6 entity tags and 9 relation tags based on several available NER and RE datasets but with some additional categories specific to Chinese literature text. Details of the tags are shown in Table 2 and 3.

We add three new entity tags specific for understanding literature text, including "Thing", "Time" and "Metric". "Thing" is for capturing objects which articles mainly describe, such as "flower", "tree" and so on. "Time" is for capturing the time-line of a story, such as "one day", "one month" and so on. "Metric" is for capturing the measurement related words, such as "1L", "1mm" and so on.

As for relation tags, we set 9 different classes for better understanding the connection between entities, including "Located", "Near", "Part-Whole", "Family", "Social", "Create", "Use", "Ownership", "General-Special". For building the relations between people in literature articles, we use the "Social" tag, which is not quite common in other corpora.

4.2. Annotation Format

Each entity is identified by "T" tag, which takes several attributes.

- *Id*: a unique number identifying the entity within the document. It starts at 0, and is incremented every time a new entity is identified within the same document.
- *Type*: one of the entity tags.
- *Begin Index*: the begin index of an entity. It starts at 0, and is incremented every character.
- *End Index*: the end index of an entity. It starts at 0, and is incremented every character.
- *Value*: words being referred to an identifiable object.

Each relation is identified by "R" tag, which can take several attributes:

- *Id*: a unique number identifying the relation within the document. It starts at 0, and is incremented every time a new relation is identified within the same document.
- *Arg1 and Arg2*: two entities associated with a relation.
- *Type*: one of the relation tags.

5. Experiments

We introduce several baselines to conduct experiments. In this section we will describe experiment settings, baselines and experiment results in detail.

Models		Thing	Person	Location	Organization	Time	Metric	All
Bi-LSTM	P	67.07	80.30	58.09	Nan	64.47	46.15	70.52
	R	62.37	78.50	46.79	Nan	45.51	22.18	62.36
	F	64.63	79.39	51.83	Nan	53.36	29.96	66.19
CRF	P	75.72	87.92	68.41	46.69	76.20	70.50	77.72
	R	65.42	82.27	50.98	45.26	60.93	38.42	65.91
	F	70.19	85.00	58.42	45.96	67.72	49.74	71.33

Table 4: Results of Named Entity Recognition on the proposed corpus.

Models	Information	F_1
SVM (Hendrickx et al., 2010)	Word embeddings, NER, WordNet, HowNet, POS, dependency parse, Google n-gram	48.9
RNN	Word embeddings	48.3
(Socher et al., 2011)	+ POS, NER, WordNet	49.1
CNN	Word embeddings	47.6
(Zeng et al., 2014)	+ word position embeddings, NER, WordNet	52.4
CR-CNN	Word embeddings	52.7
(Santos et al., 2015)	+ word position embeddings	54.1
SDP-LSTM	Word embeddings	54.9
(Xu et al., 2015)	+ POS + NER + WordNet	55.3
DepNN (Lin and Wu, 2009)	Word embeddings, WordNet	55.2
BRCNN	Word embeddings	55.0
(Cai et al., 2016)	+ POS, NER, WordNet	55.6

Table 5: Results of Relation Extraction on the proposed corpus.

5.1. Settings

Experiments are performed on a commodity 64-bit Dell Precision T5810 workstation with one 3.0 GHz 16-core CPU and 64GB RAM. The performance of NER and RE models are evaluated by F_1-score. For training, we use mini-batch stochastic gradient descent to minimize negative log-likelihood. Training is performed with shuffled mini-batches of size 32.

5.2. Named Entity Recognition

We introduce LSTM (Hochreiter and Schmidhuber, 1997) and CRF (Lafferty, 2001) as our baselines, which are described as follows.

LSTM We consider bi-directional LSTM as one of models. Both the character embedding dimension and the hidden dimension are set to be 100.

CRF CRF is a statistical modeling method which often applied in pattern recognition and machine learning. Our features template includes unigram and bigram features.

The results are shown in Table 4. It can be clearly seen that CRF achieves the better performance than Bi-LSTM on all tags, which probably be attributed to the feature template.

All models perform better on "Person", "Thing" and "Time" tags than on "Location", "Organization" and "Metric" tags. It shows that "Person", "Thing" and "Time" tags are the most easily identifiable entities. The problem of data sparsity makes it hard for capturing "Location", "Organization" and "Metric" tags.

The higher accuracies show that the entities predicted by the model probably are the right tags, which reflects the data consistency between the training and testing sets. The lower recalls indicate that there is still a lot of unknown entities on the testing set. How to handle these unknown entities is a urgent problem for further research.

5.3. Relation Extraction

Table 5 compares several state-of-the-art methods on the proposed corpus. The first entry in the table presents the highest performance achieved by traditional feature-based methods. Hendrickx et al. (2010) feeds a variety of handcrafted features to the SVM classifier and achieves an F_1-score of 48.9.

Recent performance improvements on the task of relation classification are mostly achieved with the help of neural networks. Socher et al. (2011) builds a recursive neural network on the constituency tree and achieves a comparable performance with Hendrickx et al. (2010). Xu et al. (2015) introduces a type of gated recurrent neural network which could raise the F_1 score to 55.3. By diminishing the impact of the other classes, Santos et al. (2015) achieves an F_1-score of 54.1. Along the line of CNNs, Liu et al. (2009) achieves an F_1-score of 55.2.

6. Conclusions

We build a discourse-level Named Entity Recognition and Relation Extraction dataset for Chinese literature text. To solve the problem of data inconsistency in tagging process, we propose two methods in this paper, one is a heuristic tagging method and another is a machine auxiliary tagging method. Based on this corpus, we introduce several widely-used models to conduct experiments, which provides baselines for further research.

7. References

Bunescu, R. and Mooney, R. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Cai, R., Zhang, X., and Wang, H. (2016). Bidirectional recurrent convolutional neural network for relation classification. In *ACL (1)*.

Collobert, Weston, Bottou, aKrlen, Kavukcuoglu, and Kuksa. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

He, H. and Sun, X. (2016). F-score driven max margin neural network for named entity recognition in chinese social media. *CoRR*, abs/1611.04234.

He, H. and Sun, X. (2017). A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *Proceedings of the Thirty-First Conference on Artificial Intelligence*, pages 3216–3222.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Séaghdha, D. O., Padó, S., Pennacchiotti, M., Romano, L., and Szpakowicz, S. (2010). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 33–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *CoRR*, abs/1508.01991.

Kambhatla, N. (2004). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lafferty, J. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289. Morgan Kaufmann.

Lin, D. and Wu, X. (2009). Phrase clustering for discriminative learning. In *ACL/IJCNLP*, pages 1030–1038. The Association for Computer Linguistics.

Ling, X. and Weld, D. S. (2012). Fine-grained entity recognition. In *AAAI*.

Liu, J., Huang, M., and Zhu, X. (2010). Recognizing biomedical named entities using skip-chain conditional random fields. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pages 10–18. Association for Computational Linguistics.

Ma, S. and Sun, X. (2017). A generic online parallel learning framework for large margin models. *CoRR*, abs/1703.00786.

McCallum, A., Freitag, D., and Pereira, F. C. (2000). Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

Miwa, M. and Bansal, M. (2016). End-to-end relation extraction using lstms on sequences and tree structures.

Nguyen, T. H. and Grishman, R. (2015). Combining neural networks and log-linear models to improve relation extraction.

Plank, B. and Moschitti, A. (2013). Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *ACL (1)*, pages 1498–1507.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Santos, C. N. d., Xiang, B., and Zhou, B. (2015). Classifying relations by ranking with convolutional neural networks. pages 626–634.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics.

Sun, X., Matsuzaki, T., Okanohara, D., and Tsujii, J. (2009). Latent variable perceptron algorithm for structured classification. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pages 1236–1242.

Sun, X., Li, W., Wang, H., and Lu, Q. (2014). Feature-frequency-adaptive on-line training for fast and accurate natural language processing. *Computational Linguistics*, 40(3):563–586.

Sun, X. (2014). Structure regularization for structured prediction. In *Advances in Neural Information Processing Systems 27*, pages 2402–2410.

Sun, X. (2015). Towards shockingly easy structured classification: A search-based probabilistic online learning framework. *CoRR*, abs/1503.08381.

Wang, M. (2008). A re-examination of dependency path kernels for relation extraction. In *IJCNLP*, pages 841–846.

Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., and Jin, Z. (2015). Classifying relations via long short term memory networks along shortest dependency paths. In *EMNLP*, pages 1785–1794.

Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., et al. (2014). Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.

Zhang, S., Zheng, D., Hu, X., and Yang, M. (2015). Bidirectional long short-term memory networks for relation classification. In *PACLIC*.

Zhou, G. and Su, J. (2002). Named entity recognition using an hmm-based chunk tagger. In *ACL*, pages 473–480.