Topic Sentence Named Entity Recognition: A New Task with Its Dataset and Benchmarks

Anonymous ACL submission

Abstract

In this paper, we focus on a new type of named entity recognition (NER) task called topic sentence NER. A topic sentence means a short and compact sentence that acts as a summary of a long document. For example, a title can be seen as a topic sentence of its article. Topic sentence NER aims to extract named entities in a topic sentence given the corresponding unlabeled document as a reference. This task represents real-world scenarios where full-document NER is too expensive and obtaining the entities only in topic sentences is sufficient for downstream tasks. To achieve this, we construct a large-scale human-annotated Topic Sentence NER dataset (TSNER). The dataset contains 12,000 annotated sentences accompanied by their unlabeled document. Based on TSNER, we propose a family of representative and strong baseline models, which can utilize both single-sentence and document-level features. We will make the dataset public in the hope of advancing the research on the topic sentence NER task.

1 Introduction

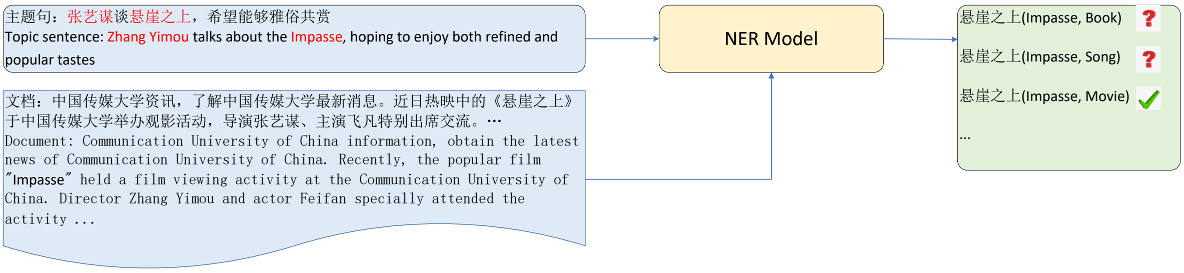
Named entity recognition is a fundamental Natural Language Processing task, which aims to label each word in sentences with predefined types, such as Person (PER), Organization (ORG), Location (LOC), etc. The results of NER play a crucial role in many downstream NLP tasks, e.g., relation extraction (Bunescu and Mooney, 2005), information retrieval (Chen et al., 2015), and question answering (Yao and Van Durme, 2014).

In this paper, we propose a new type of NER task named Topic Sentence NER, which attempts to recognize entities in topic sentences. A topic sentence is a key sentence for a document or a paragraph, which usually conveys the gist of them in a concise way. An example is shown in Figure 1. The task is defined to extract named entities like '悬崖之上(*Impasse*)' in the topic sentence. The significance of the topic sentence NER lies in two aspects. First, in many practical scenarios, it is not necessary to obtain all entities in a full-text document. Due to the time and cost of labeling and processing documents, topic sentence NER can be an effective alternative. Second, topic sentence NER is more challenging by nature and it requires new ways to incorporate the heterogeneous inputs. On the one hand, topic sentences are more informative but short in length, making the in-sentence context limited for NER. On the other hand, there are unlabeled documents that potentially enrich the context of the topic sentence, but it is unclear how to effectively utilize the information for NER.

Given the realistic necessity and challenges of topic sentence NER, in this paper, we focus on addressing this new kind of NER task. First, we construct a new dataset named TSNER. Specifically, we collect 12,000 online articles in Chinese, which are about 9 topics and contain entities of 16 types. For each article, we consider the title as the topic sentence and label the entities in the title.

Based on the proposed dataset, we establish a family of strong baseline models as benchmarks for topic sentence NER. We consider two categories of models: single-sentence NER models and document-enhanced NER models. The former only use the topic sentence as the input and consist of commonly used models that have achieved SOTA performance on many single-sentence NER datasets. The latter take both the topic sentence and its corresponding document into consideration. Two challenges have to be tackled for the document-enhanced NER model: 1) capturing long-term dependency in a computational efficiency way, and 2) distinguishing information helpful for NER from a large unrelated, noisy text. Based on the analysis, we adapt four lines of work for document-enhanced NER: distant supervision, document-level pre-trained language modeling, direct information fusion, and document gist fusion.

To the best of our knowledge, this paper is the



主题句：张艺谋谈悬崖之上，希望能够雅俗共赏
Topic sentence: Zhang Yimou talks about the Impasse, hoping to enjoy both refined and popular tastes

文档：中国传媒大学资讯，了解中国传媒大学最新消息。近日热映中的《悬崖之上》于中国传媒大学举办观影活动，导演张艺谋、主演飞凡特别出席交流。…
Document: Communication University of China information, obtain the latest news of Communication University of China. Recently, the popular film "Impasse" held a film viewing activity at the Communication University of China. Director Zhang Yimou and actor Feifan specially attended the activity ...

NER Model
悬崖之上(Impasse, Book) ?
悬崖之上(Impasse, Song) ?
悬崖之上(Impasse, Movie) ✓
...

Figure 1: A case of topic sentence NER. The topic sentence is brief and it alone provides limited context. With the help of document information, '悬崖之上(*Impasse*)' can be recognized as an entity of Movie type.

first to propose and address the topic sentence NER task. Our key contributions are as follows:

- We introduce topic sentence NER, a new NER task focusing on recognizing entities in topic sentences. This task is driven by real-world needs and is of particular research value.
- To better understand the topic sentence NER task, we propose the TSNER dataset, in which each annotated topic sentence is paired with an unlabeled document.
- Based on TSNER, we establish a family of benchmark models and conduct extensive experiments, revealing effective ways to leverage document information for this task.

2 Related work

2.1 Single-Sentence NER

Previous works mainly consider the NER task as a sentence-level task. Traditional methods try to manually construct features from single sentences and use the CRF model to learn dependency relations (Lafferty et al., 2001). With the advantages of eliminating feature engineering and significantly improving performance, neural network models become prevalent in NER research recently, e.g. FFN (Collobert et al., 2011), LSTM (Lample et al., 2016), CNN (Ma and Hovy, 2016), and pre-trained language models (Devlin et al., 2019). Single-sentence NER models work better when the entity has abundant context information, which is not the case for the topic sentence NER task.

2.2 Document-Level NER

Document-level NER extends single-sentence NER to recognize all entities in the whole document. Gui et al. (2020) introduce a two-stage label refinement approach to improve document-level label consistency. Luo et al. (2018) explore a new global attention layer on the top of BiLSTM layer to capture each word's related words in the whole document. Luoma and Pyysalo (2020) directly use BERT to obtain word representations in a cross-sentence context. Akbik et al. (2019); Luo et al. (2020) attempt to use a memory network to better address the long-term dependency problem in the document.

Aimed to recognize all entities in the document, previous document-level NER methods rely on full-document entity labels and treat each sentence in the document as equal importance. In contrast, our goal is to recognize entities in topic sentences. While some of the document-level methods can be adapted for our task, our task is still challenging due to the information gap between concisely-written topic sentences and regular sentences in documents. Recently, Wang et al. (2021) utilize a search engine to retrieve online documents as additional context for NER. Their work focuses on the scenario where relevant documents are absent, and it is difficult and time-consuming to request and process the noisy content of search engine. In contrast, the topic sentence NER task proposed by us is more realistic and practical.

2.3 Other Document-Level NLP models

Our work is also related to other document-level NLP tasks, such as document-level classification, question answering, and coreference resolution. Existing approaches to modeling document information can be summarized into three categories. The first is to chunk a document into smaller pieces of text, independently process them by single-sentence models, and then combine the results through a fusion method (Joshi et al., 2019). The second is to shorten the document by selecting only the informative parts as the input of the model (Clark and Gardner, 2018; Chen et al., 2017). The third is to develop new model architectures to efficiently accommodate the whole document (Beltagy et al., 2020; Gupta and Berant, 2020; Zaheer et al., 2020). Some of our benchmark models are derived from these three types of models.

3 Topic Sentence NER

In real-world situations, NER results are often used in downstream tasks like relation extraction, information retrieval, and question answering. In these applications, the requirement to recognize all entities in a full-text document is not always necessarily essential, and recognizing entities only in topic sentences is enough, especially when huge amounts of text have to be processed with a limit of time and cost. For example, the entities in the abstract of a scientific paper are sufficient for an up-to-date scholar search engine; the entities in a news title are enough for hot event detection and trend analysis. However, such a need for entity recognition on topic sentences has not been put forward and explored in previous NER research.

Compared with regular sentences or documents involved in previous NER tasks, topic sentences exhibit unique linguistic characteristics that makes NER more challenging. Specifically, topic sentences are often short in length but more informative in that it contains a higher density of entity words. Take the topic sentence shown in Figure 1 as an example. The number of words belonging to entities exceeds 40% of the total number of words. Consequently, the word '悬崖之上(*Impasse*)' has a limited context and is difficult to be distinguished as a book, a song, a movie, or a non-entity word. Furthermore, while documents can be incorporated to enrich the context of topic sentences, there are no ground truth NER labels for the sentences in the document, making previous document-level NER models inapplicable. All of the above call for a new research direction of context limited and document-enhanced NER methods.

Given the realistic necessity and challenges of topic sentence NER, in the remainder of this paper, we will show our initial attempt to address this problem. We will first give the definition of topic sentence NER. Then we will present our constructed dataset along with data analysis. Finally, we will propose a series of benchmark models and compare their experiment results. To the best of our knowledge, this paper is the first to propose and address the topic sentence NER task.

3.1 Task Definition

We formally define topic sentence NER as a sequence labeling task on a topic sentence accompanied by an unlabeled document. The input of topic sentence NER consists of two parts: a topic sentence $x = \{x_1, x_2, ..., x_t\}$ and an unlabeled document $D = \{s_1, s_2, ..., s_n\}$. The goal of the task is to assign each token $x_i \in x$ with a label $y_i \in Y$. Y is a set of pre-defined entity tags in BIO or other format.

3.2 Dataset Construction

The data source we used as an initial corpus is a collection of online articles in Chinese from WeChat Official Account, which contains a large variety of entities from different areas. All the articles are of year 2015 to 2020. We selected 12,000 articles on nine topics, including tourism, sports, politics, food, culture, economy, movies, entertainment, and games. We designed a NER scheme consisting of 16 commonly used entity types. The names and distribution of the entity types are shown in Table 1. We randomly split the articles into 8400 as training data, 1800 as development data, and 1800 as test data. Details of the topics and entity types of the dataset are shown in Appendix E[1].

We employed two paid annotators to annotate the dataset. We sent the titles along with the articles to the annotators and instructed them to annotate the entities in the titles with reference to the documents. The first 1200 cases are annotated by both annotators; the remainder are split evenly and annotated by each annotator respectively. Both annotators are instructed with detailed and formal guidelines and have adequate linguistic knowledge of each entity type. The Cohen's Kappa of the two annotators on the first 1200 part is 0.83, indicating a high degree of inter-rater agreement. To further ensure the annotation quality, the whole dataset is split into 6 small batches. For each small batch, the first author randomly examined 10% of the data. If the sentence-level accuracy is lower than 90%, the small batch will be returned to the corresponding annotator to be re-annotated with more detailed annotation guidelines. The process was repeated until all the batches reached above 90% accuracy. As expected, we find in many cases the title alone can not be understood by humans at a first glance. After scanning the document, however, one can confidently label the entities in the title.

[1] We will also publish the dataset later upon acceptance.

3.3 Dataset Analysis

We report some interesting statistics of our dataset compared with several widely-used NER datasets including MSRA (Levow, 2006), OntoNotes (Weischedel et al., 2013), WeiboNER (Peng and Dredze, 2015; He and Sun, 2017)[2]. We calculated the average length and entity word rate for each dataset. As shown in Table 2, two unique characteristics of topic sentences can be revealed.

1) Shorter length: The average length of topic sentences is 22 words, only less than half of the length in MSRA dataset. This indicates that processing topic sentences can be computationally efficient but consequently challenging in accuracy due to limited context information.

2) More informative: In TSNER, the average rate of entity tokens in a sentence is 30%, far higher than any other dataset. It means that topic sentences are more informative and there is even less inner-sentence context can be used for NER.

The concise writing style of topic sentences makes understanding topic sentences linguistically challenging by nature. Hence it is important to incorporate document information to facilitate the topic sentence NER.

When taking a closer look at the documents in TSNER, as shown in Table 3, we can further find two unique characteristics of our task.

1) Long document length: Compared with previous widely used datasets, TSNER provides additional unlabeled documents. The average document length is 1386. The long documents may potentially provide extra context for their paired topic sentences, but an effective method is needed to utilize the heterogeneous data.

2) High entity coverage: Among all the entities in topic sentences, about 80% also appear in the corresponding document. Such relevance further confirms that documents can provide useful information for topic sentence NER. However, given the inherent differences in length and writing style between topic sentences and documents, a large part in documents may be irrelevant and even noisy for NER. Hence it is necessary to identify proper information for topic sentence NER.

Type	Num/Rate	Type	Num/Rate
address	1889 (15%)	person	630 (5%)
entertainer	1648 (13%)	book	622 (5%)
food	1100 (9%)	tvplay	610 (5%)
event	1087 (8%)	show	537 (4%)
sports-star	994 (8%)	scene	428 (3%)
orgnization	853 (7%)	song	380 (3%)
company	722 (6%)	character	270 (2%)
movie	699 (5%)	game	259 (2%)

Table 1: The distribution of different entity types in TSNER train part.

	TSAvgLen	EntRate	Doc
MSRA	47	12.3	No
OntoNotes	31	9.1	No
Weibo NER	55	4.5	No
TSNER	22	30.0	Yes

Table 2: A comparison between TSNER and other existing widely-used NER datasets. TSAvgLen means Topics Sentence Average length, and EntRate means the rate of entity token accounts for the whole token.

4 Benchmarks

Based on the TSNER dataset, we develop a family of representative and strong baselines. The baselines are built on top of previous methods with and without considering the additional documents. We first present single-sentence NER models in Section 4.1. Then we introduce document-enhanced NER models in Section 4.2. More implementation details can be found in Section A. Note that our goal in this paper is not to exhaust all possible methods, and we hope more approaches will be proposed in the future.

	Train	Dev	Test
#sen	8400	1800	1800
#char	185.4k	38.9k	39.5k
#entity	12.8k	2.6k	2.6k
doc avg len	1386	1344	1377
% entity coverage	79.3	79.1	80.2

Table 3: The statistics of TSNER. The % entity coverage means among all labeled entities in topic sentences the percentage of the entities that are also mentioned in the paired document.

[2] For comparison, we only analyze the Chinese version of multi-lingual datasets. In the future, we will extend our work to other languages.

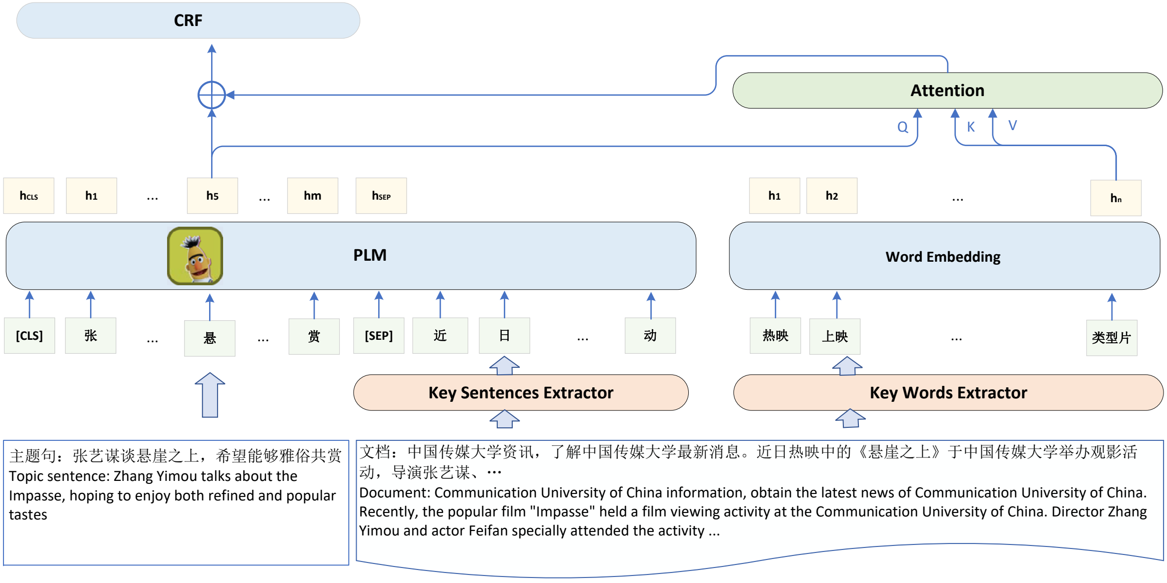

CRF
Attention
Q
K
V
hCLS
h1
...
h5
...
hm
hSEP
h1
h2
...
hn
PLM
Word Embedding
[CLS]
张
...
悬
...
赏
[SEP]
近
日
...
动
热映
上映
...
类型片
Key Sentences Extractor
Key Words Extractor
主题句：张艺谋谈悬崖之上，希望能够雅俗共赏
Topic sentence: Zhang Yimou talks about the Impasse, hoping to enjoy both refined and popular tastes
文档：中国传媒大学资讯，了解中国传媒大学最新消息。近日热映中的《悬崖之上》于中国传媒大学举办观影活动，导演张艺谋、…
Document: Communication University of China information, obtain the latest news of Communication University of China. Recently, the popular film "Impasse" held a film viewing activity at the Communication University of China. Director Zhang Yimou and actor Feifan specially attended the activity ...

Figure 2: Architecture of our document gist fusion framework. The extracted gist information includes key sentences and key words. The key sentences are encoded together with the topic sentence to provide additional context. The embeddings of the keywords are fused into the hidden states of the topic sentence using an attention mechanism.

4.1 Single-Sentence NER models

BiLSTM-CRF. BiLSTM-CRF (Lample et al., 2016) is a strong baseline that has been widely used in previous works.

SoftLexicon. In Chinese NER, previous works have shown explicitly providing word segmentation and word tagging information to be helpful for performance (Zhang and Yang, 2018; Yang et al., 2019; Li et al., 2020; Ma et al., 2020; Liu et al., 2021). Among the proposed models, we choose the SoftLexicon (Ma et al., 2020) as our baseline due to its fast speed and competitive performance.

BERT-CRF. The BERT-CRF (Devlin et al., 2019) baseline is chosen as a representative of the NER models based on pre-trained language models (PLMs).

WWM-CRF. PLMs exhibit the same lexical problem with other models when processing Chinese text. In order to take lexical information into account, PLMs with enhanced input layers and training techniques have been proposed (Cui et al., 2019, 2020; Diao et al., 2020; Sun et al., 2021). We choose the WWM model (Cui et al., 2019) for its popularity and proved generalization ability.

4.2 Document-Enhanced NER Models

We consider four lines of methods for document-enhanced NER with different complexity and depth of utilizing document information: distant supervision, document-level PLM, direct information fusion, and document gist fusion.

Distant supervision. A natural way to leverage the unlabeled document data is to regard it as an in-domain corpus for distantly supervised learning. To do so, we first build an entity dictionary by extracting all the annotated entities in the train set of TSNER. Then, we use the entity dictionary to match sentences in the documents to obtain distantly supervised data. Finally, the distantly supervised data and human annotated data are mixed together as the training data for BERT-CRF or WWM-CRF. We denote the two models as BERT-CRF-DS and WWM-CRF-DS. More dedicated methods to reduce the noise in distant supervision can be explored in the future, such as using different weights between distantly supervised data and human annotated data.

Document-level PLM. Document-level PLMs are supposed to accommodate full document as input and automatically learn to properly utilize its information for downstream tasks. In recent work, several models have been proposed to reduce memory and speed up the training of transformer models (Beltagy et al., 2020; Gupta and Berant, 2020; Zaheer et al., 2020). We explore NER model for topic sentence based on Longformer (Beltagy et al., 2020), whose attention mechanism is a drop-in replacement of the standard self-attention, which combines a local windowed attention with a task motivated global attention. The topic sentence is concatenated with the document as the input of Longformer and the global attention is applied on the topic sentence. Finally, we use the output of the Longformer as the input of CRF.

Direct information fusion. The assumption of the direct information fusion model is that the representation of each word in topic sentences will benefit from additional contexts of the same word occurred in the document. The model architecture is derived from Luo et al. (2018); Akbik et al. (2018); Luo et al. (2020). To obtain the document enhanced representation for a topic sentence, the tokens in the sentence and the document are first encoded respectively by a shared BERT encoder. For each token in the topic sentence, additional context information will be obtained by pooling the same tokens occurred in the document. The original representation in topic sentence will be concatenated with the additional context representation before feeding into CRF.

Document gist fusion. The motivation of this kind of method is that not all words in the document are helpful for the topic sentence NER task, and incorporating too much unrelated information will bring noise in training. Based on the observation, we propose a document gist fusion framework for topic sentence NER. The idea is to first efficiently mine the gist information from the long document via heuristic methods, and then fuse only the gist information into the NER process. We will first describe the framework design. The methods to extract gist information will be discussed in the next subsection.

The model architecture is shown in Figure 2. We consider two forms of gist information, i.e., key sentences and keywords. The model architecture is inspired by Wang et al. (2021), we additionally add a keyword part to it. Compared with full documents, extracted key sentences are short enough and can be easily fed into a transformer model. Hence, we append the key sentences to the topic sentence as additional inputs to a PLM encoder (WWM in our implementation):

$$H^s = \text{PLM}([x; S])_{[1:m]} \tag{1}$$

where x is the topic sentence with length m, S is the set of selected key sentences from the document. $H^s = \{h^s_1, h^s_2, ..., h^s_m\}$ is the hidden states of the topic sentence, which corresponds to the first m tokens of the inputs and is augmented with the extra context of the key sentences.

As only a few key sentences can be extracted from the document, they may be not sufficient to cover all necessary information for recognizing the entities in the topic sentence. We also consider directly including keywords as a global context that captures the document's topic. The keywords are encoded separately by a word embedding layer.

$$H^w = \text{WordEmb}(w) \tag{2}$$

where w is the set of n selected keywords and $H^w = \{h^w_1, h^w_2, ..., h^w_n\}$ is the embedding for each keyword.

We use an attention network to better modeling the relation between the sentence-level information H^s and the keywords information H^w. The attention mechanism is similar to the attention in Vaswani et al. (2017). We transform $h^s_i \in H^s$ into the attention query q_i, and keywords embedding into both the key k_j and the value v_j, where q_i, k_j, and v_j are in the same dimension. The calculations of the attention layer are as follows:

$$q_i = W^s h^s_i \tag{3}$$

$$k_j = W^w h^w_j \tag{4}$$

$$v_j = W^v h^w_j \tag{5}$$

$$u_{ij} = q_i k_j \tag{6}$$

$$\alpha_{ij} = \frac{exp(u_{ij})}{\sum_{z=1}^{n} exp(u_{iz})} \tag{7}$$

$$r_i = \sum_{j=1}^{n} \alpha_{ij} v_j \tag{8}$$

Concatenating q_i and r_i we obtain a fused representation of the topic sentences and the gist of the document:

$$f_i = [q_i; r_i] \tag{9}$$

Then f_i will be fed into a CRF layer to output entity labels. Next, we will elaborate on how we designed efficient heuristics to select key sentences and keywords from the document.

4.3 Key Sentence and Keyword Selection

We explore several methods to select the key sentences and key words for our gist fusion model. The key sentence selection is to select a maximum

Model	Resource	DEV			Test		
		P	R	F	P	R	F
BiLSTM-CRF	TS	61.10	59.16	60.12	60.08	59.97	60.03
SoftLexicon	TS	69.59	59.62	64.22	70.64	61.27	65.62
BERT-CRF	TS	78.06	76.69	77.37	77.49	77.62	77.56
WWM-CRF	TS	78.11	76.77	77.43	77.98	78.01	77.99
BERT-CRF-DS	TS + doc	78.42	77.36	77.88	78.66	78.54	78.60
WWM-CRF-DS	TS + doc	78.51	77.47	77.99	78.71	78.60	78.66
Longformer	TS + doc	78.50	77.42	77.96	78.36	78.64	78.50
DirectFusion	TS + doc	79.98	78.48	79.22	78.86	79.01	78.93
Gist-SBERT	TS + doc	80.35	78.66	79.49	80.55	79.70	80.12
Gist-BERTScore	TS + doc	80.23	79.14	79.68	80.48	79.98	80.23
Gist-First	TS + doc	81.33	79.01	80.15	81.23	79.88	80.55
Gist-WMD	TS + doc	81.38	79.21	80.28	82.38	80.14	81.24
Gist-Noun	TS + doc	**81.50**	79.98	80.73	82.31	81.48	81.89
Gist-Noun-Yake	TS + doc	80.46	79.81	80.13	81.79	80.72	81.25
Gist-Noun-TextRank	TS + doc	81.47	**80.38**	**80.92**	**82.47**	**81.69**	**82.08**

Table 4: The performances of different approaches on TSNER dataset. The table is divided into five blocks from top to bottom, representing the results of five families of methods: single-sentence NER models, distant supervision, document-level PLM, direct information fusion model, and document gist fusion models.

number of N sentences from the document. In order to provide adequate context with a reasonable cost of longer input length, we empirically set $N = 5$ in our study.

First in order. In this strategy, we simply take the the first N sentences from the beginning of the document as usually they are of more importance than the following sentences.

Similarity-based. The idea of this strategy is to select sentences that are semantically similar to the topic sentence based on a similarity metric. Three similarity metrics for sentences are considered. The first is Word Mover's Distance (WMD) (Kusner et al., 2015) based on word embedding. The second is pretrained SBERT (Reimers and Gurevych, 2019), which derives semantic aware sentence embedding from a Siamese BERT network and uses cosine similarity to measure similarity between them. The third is BERTScore(Zhang et al., 2019), which considers each token in two sentences to compute the similarity.

Noun overlapping. We propose a simple method to select the key sentence based on the co-occurrence of noun words in a topic sentence and its document. Sharing common nouns means that two sentences have a closer relationship, and that they together form a richer context for the common nouns. Specifically, we scan the sentences of the document in the natural order and pick out sentences that share at least one common noun with the topic sentence. In order to increase the diversity, we limit the number of sentences that each noun can associate with to two. When the limit is exceeded, only the two sentences that are more close to the beginning in the document will be kept. When there is no noun overlapping, we select the sentences from the beginning of the document (i.e., fall back into first-in-order).

For keyword extraction, we explore the following two common statistical-based methods.

TextRank (Mihalcea and Tarau, 2004) is a graph-based word ranking model inspired by PageRank. It is widely used for selecting informative words from a document.

Yake (Campos et al., 2020) is a more recent and lightweight approach for keyword extraction, which uses statistical features to measure the importance of each word in a document.

By combining the above key sentence and keyword selection methods with the model architecture in Figure 2, we expand our benchmarks with multiple variations of document gist fusion model.

5 Results and Analysis

In this section, we report the results of various experiments carried out on the TSNER dataset. Following the evaluation metrics in previous NER research, we report results in terms of entity-level (exact entity match) standard micro Precision (P), Recall (R), and F1 score. We will also present our analysis of the results.

5.1 Results

Table 4 shows the results of all benchmark models on TSNER. We summarize the findings into the following conclusions.

1) Incorporating document information can significantly improve the performance of topic sentence NER. For example, compared with the WWM-CRF model, four types of document-enhanced models (WWW-CRF-DS, Longformer, DirectFusion and Gist-Noun-TextRank) can improve the F1 score by 0.67%, 0.51%, 0.94%, 4.09% respectively on the test set.

2) For document-enhanced models, different design can incorporate different level of document information and lead to different performance. Document gist fusion models perform better than distant supervision (DS), Longformer and DirectFusion model. Even the worst performing document gist fusion model (Gist-SBERT) outperforms WWM-CRF-DS, demonstrating the advantage of understanding the gist of document. The DirectFusion model achieves a modest performance. Surprisingly, Longformer exhibits the lowest performance. We suppose that Longformer may not be suitable for the NER task. Besides, as the data used in pretraining Longformer are different from BERT or WWM, it may not be fair to compare Longformer with the other BERT- or WWM-based models.

3) The performance of different document gist fusion models varies largely. The best model (Gist-Noun-TextRank) surpasses the worst (Gist-SBERT) by 1.96%. This indicates a research direction on how to better extract useful information from the document. There are also some other interesting findings. First, choosing the most similar sentences may not lead to a better result as they may not provide useful information. In the contrary, sentence selection based on SBERT the performs worst. Second, the ways to select keywords also have an impact on NER performance. The Yake based method shows a negative effect.

5.2 Error Analysis

	Type	**Cboundary**	**NOOVER**
WWM-CRF	223	224	274
Gist-Noun-TextRank	181	221	243

Table 5: The statistics of different errors that occur in the output of Gist-Noun-TextRank models on the test set. Cboundary means that Cross-Boundary error and NOOVER is non-overlapping error.

In order to further analyze the reasons why the document gist fusion model outperforms single-sentence models in topic sentence NER, we categorized and studied three types of errors: entity type error, cross-boundary error, and non-overlapping error. The entity type error means that the predicted entity is correct in the boundary but wrong in type. The cross-boundary error means that the boundary of golden entity overlaps with the model prediction. The non-overlapping error means no word overlapping between the golden entity and the model prediction. The error analysis of two representative models is shown in Table 5. From the table, we summarize the following two observations.

1) Non-overlapping error type takes up most of the errors for both models. We find in many cases that punctuation marks in the document can help to recognize the boundary of an entity, but assigning the entity with a correct type is more difficult for WWM-CRF as the context is limited.

2) Leveraging document information can effectively reduce non-overlapping errors and entity type errors. However, the cross-boundary errors are difficult to be reduced, which indicate they are hard cases that lack of informative context in document.

6 Conclusion

In this paper, we propose a new task called topic sentence NER. The task is driven by real-world scenarios where extracting entities in topic sentences instead of the full-text documents is sufficient and economic. While the task is of value and is more challenging than regular NER, it has not been explored in previous research. To address this task, we build a large-scale annotated NER dataset named TSNER. A family of baseline models are established based on TSNER. We hope our dataset and benchmarks will advancing the research on topic sentence NER. Some interesting directions for future research are shown in Appendix D.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. 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 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Razvan Bunescu and Raymond Mooney. 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*.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Inf. Sci.*, 509:257–289.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176.

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

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 657–668. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese BERT. *CoRR*, abs/1906.08101.

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

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4729–4740. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4830–4842. Association for Computational Linguistics.

Tao Gui, Jiacheng Ye, Qi Zhang, Yaqian Zhou, Yeyun Gong, and Xuanjing Huang. 2020. Leveraging document-level label consistency for named entity recognition. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3976–3982. ijcai.org.

Ankit Gupta and Jonathan Berant. 2020. Gmat: Global memory augmentation for transformers. *arXiv preprint arXiv:2006.03274*.

Hangfeng He and Xu Sun. 2017. F-score driven max margin neural network for named entity recognition in chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 713–718. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 5802–5807. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

John D. Lafferty, Andrew Mccallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Eighteenth International Conference on Machine Learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117. Association for Computational Linguistics.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6836–6842. Association for Computational Linguistics.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced chinese sequence labeling using BERT adapter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5847–5858. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Ying Luo, Fengshun Xiao, and Hai Zhao. 2020. Hierarchical contextualized representation for named entity recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8441–8448. AAAI Press.

Jouni Luoma and Sampo Pyysalo. 2020. Exploring cross-sentence contexts for named entity recognition with BERT. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 904–914. International Committee on Computational Linguistics.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2020. Simplify the usage of lexicon in Chinese NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5951–5960, Online. Association for Computational Linguistics.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*

2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 2065–2075. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*.

Jie Yang, Yue Zhang, and Shuailong Liang. 2019. Subword encoding in lattice LSTM for Chinese word segmentation. 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 2720–2725, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 956–966. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564. Association for Computational Linguistics.

A Implementation Details

BiLSTM-CRF: The character embedding is pre-trained on Chinese Giga-Word using word2vec (Mikolov et al., 2013). The character embedding dimension is set to 100, the LSTM hidden states dimension is set to 300 and the initial learning rate is set to 0.001. The models is trained using 100 epochs with a batch size of 16.

SoftLexicon: We use the same code [3] from the paper (Ma et al., 2020). The LSTM-based sequence modeling layer is used.

Pretrained Language Model: The pre-trained language model (BERT, WWM) is from huggingface [4]. The initial learning rate of PLM is set to 1×10^{-5} . For the CRF layer parameters, we use a learning rate of 1×10^{-3}. The optimizer is AdamW(Loshchilov and Hutter, 2019). We fine-tune models using 20 epochs with a batch size of 16.

Longformer We use global attention to the topic sentence and a CRF layer is on the top of the topic sentence. The maximum length is set to 1024. The batch size is set to 4. Other parameters are the same as the Pretrained Language Model.

Gist-Noun-TextRank: The word embedding is pre-trained on Chinese Giga-Word using word2vec (Mikolov et al., 2013). The word embedding dimension is set to 50. The embedding of q, k, v is 150.

Computing Infrastructure: All experiments are conducted on an NVIDIA Tesla V100 (32 GB of memory).

B Performance Analysis by Entity Types

We further analyze the performance of our best model on different entity types. The results are shown in Table 7. From the table, we find that the Gist-Noun-TextRank model achieves the best performance on the sports-star type. The reason may be that the name is easy to be recognized and there are fewer interference items in the field of sports. The model performs worst on the book type, because book titles can be largely varied across genres, and many documents only consist of the content of the book, providing less additional context to the title.

Topic sentence and document	WWM-CRF	Gist-Noun-TextRank
TS: 2019[褚橙]$_{\text{Food}}$来了 *Here comes [Chu orange]*$_{\text{Food}}$, 2019 Doc: ...橙子便是来自云南哀牢山的[褚橙]... *...Oranges are [Chu orange] from Ailao Mountain*	Name	Food
TS: 11月15日，三分钟[兴化]$_{\text{Address}}$新鲜事来了！ *November 15, three minutes of [Xinghua]$_{Address}$ news* Doc: ...[兴化]市2019年公开招聘... *...[Xinghua] open recruitment in 2019...*	None	Address

Table 6: Case study. In the topic sentence, the text in brackets is the candidate mention, followed by the golden label. The text in brackets in the document is the sharing common entity between topic sentence and document. Predicted labels in red denote the wrong answer.

Type	F1	Type	F1
address	85.90	person	73.42
entertainer	87.95	book	70.5
food	82.78	tvplay	85.71
event	72.46	show	84.78
sports-star	90.62	scene	70.94
orgnization	73.24	song	81.08
company	74.35	character	70.97
movie	90.13	game	90.09

Table 7: F1-scores of different entity types on TSNER.

C Case Study

To clearly show the effectiveness of document-enhanced models for the topic sentence NER task, we analyze two representative cases by comparing the output of WWM-CRF and Gist-Noun-TextRank. The cases and prediction results are shown in Table 6. One type of common error is wrong entity type. The WWM-CRF model tends to predict entity type based on the mentioned words alone. In the first case, WWM-CRF model predicts '褚橙(Chu orange)' as a person name as '褚(Chu)' is a last name in Chinese names. The document-enhanced model can avoid the mistake: the Gist-Noun-TextRank model can refer to the document context to predict it as a food. Another type of common error is missing entities. In the second example, '兴化(Xinghua)' is not recognized by the WWM-CRF model. In contrast, the document-enhanced model can correctly predict '兴化(Xinghua)' as an address. We suppose that the word '市(city)' in the document acts as a clear clue to guide the model's prediction.

D Future Works

In the future, the following interesting directions can be explored.

1) When using distant supervision methods, how to leverage the noise in the document and how to model the relation between topic sentence and document are worth exploring.

2) It is promising to build a pre-trained model to learn the relationship between topic sentences and corresponding documents.

3) Strategies to extract explicit information in the document have been proved helpful for topic sentence NER and hence worth being further explored. For example, a two-stage summarization (Dou et al., 2021), that first selects key information or keywords as guided information and then generate a summary, can be helpful for the topic sentence task.

4) More types of external information can be incorporated into NER other than document text, e.g., knowledge base and visual contents.

5) It is also interesting to extend the topic sentence dataset to include more relationships other than title-document, e.g., abstract-paper.

E Entity Types in TSNER

A detailed description of the entity types in TSNER are shown in Table 8.

[3] https://github.com/v-mipeng/LexiconAugmentedNER
[4] https://huggingface.co/models

Topic (entity name)	Interpretation	Example
地址 Address (address)	常见的行政区划，如省，市，县，村，常见国家名 common administrative divisions, such as counties, provinces, cities, villages, etc.	北京，中关村，中国 Beijing, Zhongguancun, China
景点 Tourist attraction (scene)	除地址外较小的较具体的地名，如旅游景点等 smaller and more specific tourist attractions apart from the address	长沙公园，海洋馆，植物园 Changsha Park, aquarium, botanical garden
娱乐人物 Entertainer (entertainer)	与娱乐相关的人物，包括影视演员，歌手等 entertainment stars, including actors, singers, etc.	胡歌，彭昱畅，张学友 Hu Ge, Peng Yuchang, Zhang Xueyou
体育人物 Sports star (sports-star)	主要是运动员等 mostly athletes	刘翔，郭晶晶 Liu Xiang, Guo Jingjing
文创人物 Virtual character (character)	游戏，影视剧，小说等中的虚拟角色 virtual characters in games, films, TV shows, novels, etc.	寒冰射手，李元芳 Ice shooter, Li Yuanfang
其他人物 Other person name (person)	除娱乐，体育，文创的其他人物 other person name besides entertainer, sports-star, and character	马化腾，马云 Ma Huateng, Ma Yun
公司 Company (company)	以盈利为目的的公司 commercial companies	阿里，腾讯 Alibaba, Tencent
组织机构 Organization (organization)	除公司外的团体，如兴趣爱好团体，大学 groups other than companies, such as interest groups, universities	海淀棋社，北京大学 Haidian chess club, Peking University
电影 Movie (movie)	在电影院上线的视频 cinematic movies	英雄本色，纵横四海 A Better Tomorrow, Once A Thief
电视节目 TV show (tvshow)	在电视或网络上上线的电视剧，综艺等 TV dramas and variety shows launched on TV or on the Internet	琅琊榜，甄传 Langya list, biography of Zhen Huan
表演 Performance (show)	需现场观看的节目，如话剧，戏曲，相声，小品等 live shows, such as dramas, operas, crosstalks, comedies, etc.	天仙配，女驸马 Tianxianpei, daughter-in-law
事件 Event (event)	大型赛事，展览，会议等 major events, exhibitions, conferences, etc.	东京奥运会 Tokyo Olympic Games
歌曲 Song (song)	歌曲名称 names of songs	我愿意，吻别 Still Here, Take me to your heart
书名 Literature (book)	小说，杂志，文学作品等 novels, magazines, literary works, etc.	挪威的森林，飞鸟集 Norwegian Wood, Stray Birds
美食 Food (food)	各种食物 names of food	炸鸡腿，汉堡 fried chicken leg, hamburger
游戏 Video game (game)	各种游戏 names of video games	魔兽，王者荣耀 Warcraft, Honor of Kings

Table 8: Topics and entity names in TSNER.