Title2Event: Benchmarking Open Event Extraction with a Large-scale Chinese Title Dataset

Haolin Deng[1] **Yanan Zhang**[1]* **Yangfan Zhang**[1] **Wangyang Ying**[1]
Changlong Yu[2] **Jun Gao** **Wei Wang**[3] **Xiaoling Bai**[1]
Nan Yang[1] **Jin Ma**[4] **Xiang Chen**[1] **Tianhua Zhou**[1]
[1]Tencent [2]HKUST [3]Tsinghua University [4]USTC
hldeng028@gmail.com, {yananzhang, devinbai}@tencent.com

Abstract

Event extraction (EE) is crucial to downstream tasks such as new aggregation and event knowledge graph construction. Most existing EE datasets manually define fixed event types and design specific schema for each of them, failing to cover diverse events emerging from the online text. Moreover, news titles, an important source of event mentions, have not gained enough attention in current EE research. In this paper, We present Title2Event, a large-scale sentence-level dataset benchmarking Open Event Extraction without restricting event types. Title2Event contains more than 42,000 news titles in 34 topics collected from Chinese web pages. To the best of our knowledge, it is currently the largest manually-annotated Chinese dataset for open event extraction. We further conduct experiments on Title2Event with different models and show that the characteristics of titles make it challenging for event extraction, addressing the significance of advanced study on this problem. The dataset and baseline codes are available at https://open-event-hub.github.io/title2event.

1 Introduction

Event extraction (EE) is an essential task in information extraction (IE), aiming to extract structured event information from unstructured plain text. Extracting events from news plays an important role in tracking and analyzing social media trending, and facilitates various downstream tasks including information retrieval (Basile et al., 2014), news recommendation system (Raza and Ding, 2020) and event knowledge graph construction (Gottschalk and Demidova, 2018; Yu et al., 2020; Gao et al., 2022). Figure 1 shows an example of extracting events from multiple news titles. Based on the extracted events, news reporting the same event could be aggregated and sent to users to provide comprehensive views from different sources.

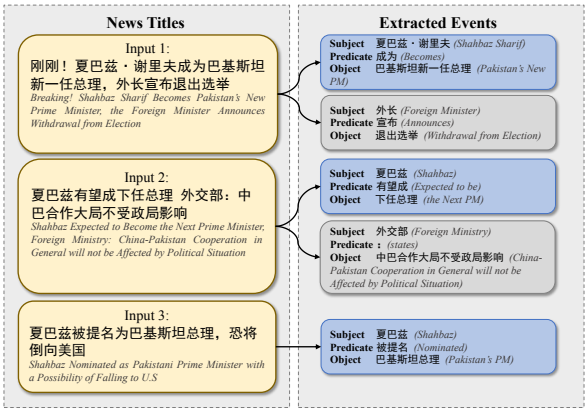

Figure 1: An example of event extraction on news titles where all factual events are extracted. Similar events are identified (in blue color) and could be used in aggregating relevant news.

Event extraction can be categorized into two levels: sentence-level EE and document-level EE. Sentence-level EE identifies event entities and attributes in a single sentence (Ahn, 2006), while document-level EE aims to extract entities of the same event scattered across an article (Sundheim, 1992). In scenarios such as news aggregation, human-written news titles often preserve the core information of the news event, while news articles may contain too many trivial details. Therefore, performing sentence-level EE on news titles is more efficient than document-level EE on news articles to aggregate relevant news.

However, most EE models trained on traditional sentence-level datasets could not reach ideal performance when extracting events from titles (Chen et al., 2015; Nguyen and Nguyen, 2019; Wadden et al., 2019; Du and Cardie, 2020; Li et al., 2020a; Liu et al., 2020; Lu et al., 2021; Lou et al., 2022). On the one hand, these models request predefined event types and a specific schema for each of them. Each event schema consists of manually designed argument roles such as event trigger, person, time, and location. Then the extraction of events will

* Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6511 - 6524
December 7-11, 2022

Challenges	Examples	Expected Outputs	Description
Unconventional Writing	**对标宏光MINI EV!售价2.99万起,奇瑞QQ冰淇淋正式上市** (Benchmarking Against Wuling Hongguang Mini EV! Priced From ￥29,900, Chery QQ Ice-cream Officially Launched)	**(S: 奇瑞QQ冰淇淋，** Chery QQ Ice-cream **P: 对标，** Benchmarking Against **O: 宏光MINI EV)** Wuling Hongguang Mini EV **(S: 奇瑞QQ冰淇淋， P: 正式上市)** Chery QQ Ice-cream Officially Launched	The subject of the predicate **"对标 (Benchmarking Against)",** is omitted since it appears at the next event
	再夺国家级奖项!佛山新城夜景亮化工程夜色璀璨 (Winning Another National Award! Foshan New City Night Lighting Project Dazzling)	**(S: 佛山新城夜景亮化工程，** Foshan New City Night Lighting Project **P: 再夺，** Wining Another **O: 国家级奖项)** Winning Another National Award	The predicate and object are placed before the subject for emphasis
Role Overlap Problem	**甘肃斥资30亿元修建古城,或将成为当地新地标** (Gansu PR. Spent 3 Billion Yuan to Rebuild the Ancient City, Likely to Become a New Local Landmark)	**(S: 甘肃，** Gansu PR. **P: 斥资，** Spent **O: 30亿元)** 30 Billion Yuan **(S: 甘肃，** Gansu PR. **P: 修建，** Rebuild **O: 修建古城)** the Ancient City **(S: 古城，** the Ancient City **P: 或成为，** Likely to Become **O: 当地新地标)** New Local Landmark	**"古城(the Ancient City)"** is the subject of one event and the object of the other.
	一艘中国集装箱船与韩国渔船相撞,导致2人失踪 (A Chinese Container Ship Collided With a South Korean Fishing Boat, Leaving Two People Missing)	**(S: 中国集装箱船与韩国渔船， P: 相撞)** CN Container Ship and KR Fishing Boat Collided **(S: 一艘中国集装箱船与韩国渔船相撞，** CN Container Ship Colliding With KR Fishing Boat **P: 导致，O: 2人失踪)** Leaving Two People Missing	The first event is actually the subject of the second event, which indicates these two events are associated.
Requirement of Domain Knowledge	**保尔特17号洞冲刺打鸟赶完赛** (Poulter Rushes to Score Birdie, Finishing the Match on 17)	**(S: 保尔特，P: 打鸟)** Poulter Score Birdie **(S: 保尔特，P: 赶完，O：赛)** Poulter Finish Match ***Incorrect Output:*** ***(S: 保尔特，P: 打，O：鸟)*** *Poulter Hit Bird*	**"打鸟(Score Birdie)"** is a term in golf but can be literally interpreted as "hit bird" in Chinese. Domain knowledge is needed in dealing with such cases.

Figure 2: Three types of challenges observed in Title2Event along with their corresponding examples.

be decomposed into sub-tasks of extracting each argument role separately. Despite the success in traditional EE, the manual design of specific event schema is costly and time-consuming, and the limited predefined event types could not handle a great variety of events emerging from the Internet where most news titles nowadays are derived from. On the other hand, extracting events from Chinese titles could be more challenging than traditional sentence-level EE such as the ACE 2005 benchmark.[1] This is because some unique writing styles are observed in news titles on Chinese social media, as shown in Figure 2. First, the writing of many titles does not strictly obey the correct grammar. For example, some titles will omit the agent when describing an action for brevity, while others may place the action before the first mention of the agent for emphasis. Second, the role overlap problem, i.e., the same entity may play different roles in multiple events, usually occurs when the events in the text have certain associations with each other. Although there are about 10% events in ACE 2005 having this problem, it has not gained enough research attention for quite a long time (Yang et al., 2019). However, the role overlap problem is much more commonly observed in news titles, and thus becomes an issue that can not be ignored. Finally, due to the diverse coverage of news reports, there are some cases in which the EE models have to rely on certain domain knowledge (e.g. rules and terms in sports) for correct event understanding. All these characteristics of titles bring additional challenges to event extraction, demanding EE models of the greater capability of text understanding.

Considering the significance of title event extraction and a lack of corresponding benchmarks, we present Title2Event, a dataset with more than 42,000 Chinese titles collected from the Internet. In general, Title2Event has the following important features:

1. It formulates title event extraction as an open event extraction (OpenEE) task without any predefined event type or specific schema. Instead, it follows the formulation of open information extraction (OpenIE) (Zhou et al., 2022) and defines an event as a `(Subject,Predicate,Object)` triplet. Then, the EE models are required to extract all event triplets in a given title. The biggest difference between OpenEE and OpenIE is that OpenEE is event-centric, which means

[1] https://catalog.ldc.upenn.edu/LDC2006T06

only triplets of events are to be extracted.

2. It is a large-scale, high-quality dataset. Title2Event consists of 42,915 news titles in 34 domains collected from Chinese web pages, along with 70,947 manually annotated event triplets containing 24,231 unique predicates. We write detailed annotation guidelines and conducted two rounds of expert review for quality control. To the best of our knowledge, Title2Event is currently the largest manually annotated Chinese dataset for OpenEE.

3. It is the first sentence-level dataset with a special focus on titles with its unique values and challenges that little attention has been paid to. We believe Title2Event could further facilitate current EE research in real-world scenarios.

We experiment with different methods on Title2Event and analyze their performance to address the challenges of this task.

2 Related Work

Event Extraction Datasets. Automatic Content Extraction (ACE 2005) (Doddington et al., 2004) is one of the most widely-used corpora in event extraction. It contains 599 documents with 8 event types, 33 event subtypes, and 35 argument roles in English, Arabic and Chinese (Li et al., 2021b). TAC KBP 2017[2] is a dataset of the event tracking task in KBP which contains 8 event types and 18 event subtypes in English, Chinese and Spanish. MAVEN (Wang et al., 2020) collects 4,480 Wikipedia documents, 118,732 event mention instances and constructs 168 event types. Despite the large scale, MAVEN merely focuses on event triggers without annotating event arguments. All of the above datasets manually define event types and schema, struggling to handle newly emerging event types in real-world applications.

Open Information Extraction. Open information extraction (OpenIE) aims to extract facts in the form of relational tuples from unstructured text without restricting target relations, relieving human labor of designing complex domain-dependent schema (Niklaus et al., 2018). Due to the release of large-scale OpenIE benchmarks such as OIE2016 (Stanovsky and Dagan, 2016) and CaRB (Bhardwaj et al., 2019), neural OpenIE approaches become popular (Zhou et al., 2022). Existing neural OpenIE models can be categorized into sequence tagging models (Stanovsky et al., 2018; Kolluru et al., 2020a; Zhan and Zhao, 2020) and generative sequence-to-sequence models (Cui et al., 2018; Kolluru et al., 2020b). We adopt the formulation of OpenIE and represent events as triplets since the event mentions in news titles tend to be brief without complex substructures.

Chinese Event Extraction. Chinese event extraction can be regarded as a special case of EE due to its unique linguistic properties and challenges (Li et al., 2021b). However, the resources of Chinese EE data are relatively scarce and lack sufficient coverage comparing with EE data in English, which greatly hinders existing research (Zeng et al., 2016; Lin et al., 2018; Ding et al., 2019; Xiangyu et al., 2019; Xu et al., 2020; Cui et al., 2020). Apart from multilingual datasets with Chinese corpora such as ACE 2005 and TAC KBP 2017, Chinese Emergency Corpus (CEC)[3] collects 6 types of common emergency events. Doc2EDAG (Zheng et al., 2019) and FEED (Li et al., 2021a) are two Chinese financial EE datasets built upon distant supervision. DuEE (Li et al., 2020b) is a document-level EE dataset with 19,640 events categorized into 65 event types, collected from news articles on Chinese social media. Compared with DuEE, our Title2Event dataset is larger in scale and does not restrict event types.

3 Dataset Construction

This section describes the process of data collection and annotation details.

3.1 Data Collection

We broadly collect Chinese web pages from January to March 2022 using the web crawler logs of the search engine of Tencent as well as a proven business tool to select web pages containing event mentions (most of them are from news websites). Afterwards, the titles of the selected web pages are extracted and automatically tagged with our predefined topics, and titles containing toxic contents are all removed. To ensure the diversity of events, we conduct data sampling every ten days during the crawling period, reducing the occurrence of events belonging to the top frequently appeared topics

[2] https://tac.nist.gov/2017/KBP/data.html

[3] https://github.com/shijiebei2009/CEC-Corpus

to make the distribution of topics more balanced. Eventually, around 43,000 instances are collected.

3.2 Annotation Framework

Annotation Standard. We summarize some essential parts of our annotation standard. In general, we expect each event could be represented by a `(Subject,Predicate,Object)` triplet where the subject and object could be viewed as the argument roles of the event triggered by the predicate. Multiple event triplets may be extracted from a single title, and they may have some overlaps. However, the predicate of an triplet is considered as the unique identifier of an event, thus multiple triplets of a single title will not share the same predicate. Some important specifications are listed below:

1) We define event as an action or a state of change which occurs in the real world. Some statements such as policy notifications or some subjective opinions are not considered as events. Also, if an title is not clearly expressed, or is concatenated by several unrelated events (e.g. news round-up), then it should be labeled as "invalid" by annotators.

2) We find the identification of predicates in Chinese is complex, so we specify some rules to unify them. First, if an event tends to emphasize the state change of the subject, e.g. "南阳大桥通车" (Nanyang Bridge opens to traffic), then the predicate will be labeled as "通车" (open-to-traffic) instead of "通" with "车" as the object. Second, for phrases with serialized verbs and dual objects, we integrate the direct target of the action (i.e. the *Patient*) into the predicate expression while taking the indirect patient (i.e. the *Affectee*) (Thompson, 1973) as the object of the event. For example, in "送孩子去学校" (send kids to school) the predicate will be labeled as "送去学校" (send-to-school) with "孩子" (kids) as the object. Moreover, we find the colon (": ") frequently plays the role of predicate in titles, representing the meaning of "say", "announce" or "require", etc. We view this as a feature of news titles and allow annotators to label it as the predicate.

3) We expect the fine-grained annotations of argument roles, which are intact yet not redundant. All determiners and modifiers of entities are kept only if they largely affect the understanding of events. All triplets are required to have a subject and a predicate, while the object could be omitted as in the original text.

Crowdsourced Annotation. We cooperate with crowdsourcing companies to hire human annotators. After multi-rounds of training in three weeks, 27 annotators are selected. We pay them ¥1 per instance. Meanwhile, four experts are participated in two rounds of annotation checking for quality control. For each instance, a human annotator is asked to write all expected event triplets independently. To reduce the annotation difficulty, we provide some auxiliary information along with the raw title, including the tokenization outputs, to help annotators quickly capture the entities and concepts present in the titles. Note that we do not force annotators to strictly obey the tokenization outputs, as we find that many of them do not match the desired granularity of triplet elements under our criteria. Instead, the annotation is conducted in a `<text, label>` pair paradigm rather than a token-level tagging paradigm. Moreover, we provide automatic extraction outputs as references. During the initial phase, we design an unsupervised model to extract triplets. After 20,000 labeled instances are collected, we train a better sequence tagging model for the rest of annotation process. Both models are introduced in Section 5. Meanwhile, as titles often contain some domain knowledge which the annotators may not be familiar with, we allow them to refer to search engines. To ensure the quality, we also allow them to label an instance as "not sure" if they are not confident enough. The crowdsourced annotation is conducted in batches. Every batch of annotated instances undergoes two rounds of quality checking before being integrated into the final version of our dataset. We also develop a browser-based web application to accelerate the annotation process, see Appendix A.

First-round Checking. Each time the crowdsourced annotation of a batch is completed, it is sent to four experts to check whether they meet the requirements of our annotation standard. Instances which do not pass the quality check will be sent back for revision, attached with the specific reasons for rejection. This process repeats until the acceptance rate reaches 90%.

Second-round Checking. Each batch of annotated instances passing the first-round checking is sent to the authors for dual check. The authors will randomly check 30% of the instances and send unqualified instances back to the experts along with the reasons for rejection. Slight adjustments on annotation standard also take place in this phase. This

Attribute	Count
Data Size (train/val/test)	34,295/4,286/4,288
Number of Topics	34
Total Events	70,879
Total Unique Predicates	24,097
Avg. Num. of Event per Title	1.65
Max. Num. of Event per Title	6
Avg. Len. of Title	23.98

Table 1: The overall statistics of Title2Event.

process repeats until the acceptance rate reaches 95%.

Our annotation process encourages positive interactions among the authors, the experts and the crowdsourced annotators, which effectively helps the annotators to understand the annotation standard and provide timely feedback.

4 Data Analysis on Title2Event

This section describes the statistics and characteristics of Title2Event from various perspectives. Table 1 shows the overview of the dataset.

Topic Distribution. The titles in the dataset can be categorized into 34 topics, 24 of which contain more than 100 instances. Figure 3 lists the distribution of instances belonging to different topics, see Appendix A for detailed statistics.

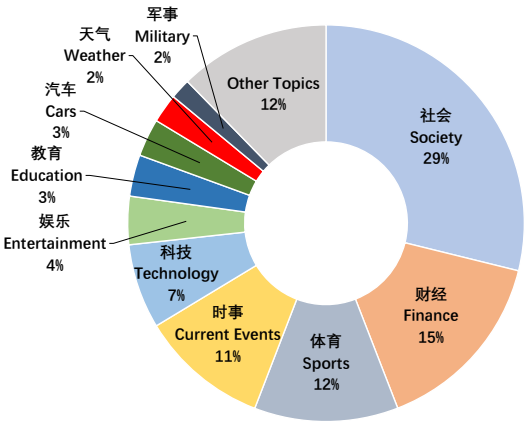

Figure 3: The distribution of topics in Title2Event, all non top-10 topics are aggregated as "Other Topics".

Event Distribution. As shown in Table 1, most of the titles contain more than one event, and the maximum number of events per title is six. We further investigate the distribution of instances containing different numbers of triggers (i.e. predicates for Title2Event), and compare our dataset with the

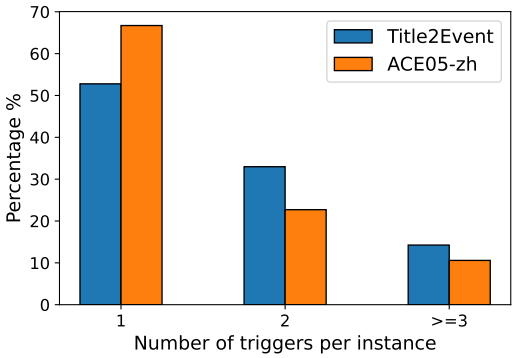

Figure 4: Distribution of instances containing different numbers of triggers of Title2Event and ACE05-zh.

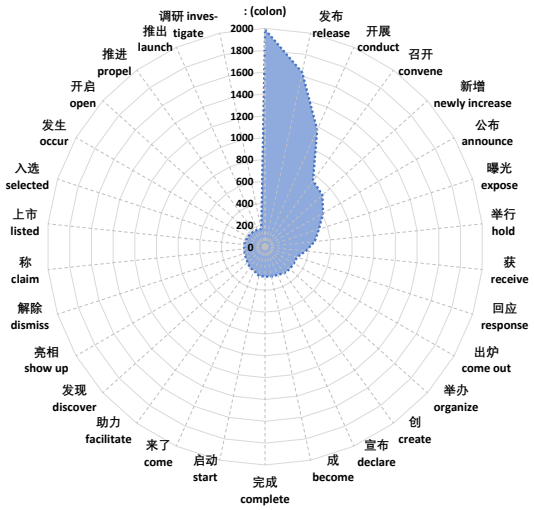

Figure 5: Distribution of top 30 predicates in Title2Event.

ACE2005 Chinese dataset (denoted as ACE05-zh) [4] as shown in Figure 4. It can be observed that the phenomenon of multiple events per instance is more common in Title2Event compared with ACE05-zh, which brings additional challenges in event extraction.

Predicate Distribution. We also investigate the distribution of predicates in Title2Event. Figure 5 shows the distribution of the 30 most frequent predicates in the dataset.

Challenge Distribution. We further analyze to what extent are the observed challenges described in Figure 2 covered in Title2Event. To do this,

[4] We adopt a commonly used preprocessed version of the ACE corpus on sentence level (https://github.com/nlpcl-lab/ace2005-preprocessing). We also remove all sentences annotated with no event (which accounts for 69.6% in the processed dataset) for fair comparison.

we randomly sample 1,000 instances and manually annotate 1) Whether the instance omits or inverts some event arguments which makes itself not strictly obeying the grammatical norms. 2) Whether there's a text span appearing at multiple events of the instance. (3) Whether some domain knowledge is crucial in understanding the instance that without these knowledge one might not correctly identify the event arguments. The annotation result shows that 9.70% of sampled instances are observed with unconventional writing, 21.50% instances have role overlap problem (10% for ACE 2005 for comparison), and 2.80% instances requires domain knowledge for correct event understating. We believe such statistics are a good identification of the challenging nature of Title2Event.

5 Methods

Formally, given a sequence of tokens $S = < w_1, w_2, \ldots, w_n >$, Open EE aims to output a list of triplets $T =< t_1, t_2, \ldots, t_m >$ where each triplet $t_i =< s_i, p_i, o_i >$ represents an event occurred in S and s_i, p_i, o_i denote the subject, predicate and object of the event respectively. The object of an event could be empty, and the total number of events per sentence m is not fixed. Open EE can also be aligned with traditional EE task formulation by considering the predicate as the event trigger as well as a unique event type, while the subjects and objects both taken as event arguments.

Based on the task formulation, we first implement an unsupervised method using an existing toolkit. Then, we split the task into trigger extraction and argument extraction, and implement different supervised methods on them.

5.1 Unsupervised Method

Since the formulation of Open EE is similar to some traditional tasks such as dependency parsing (DP) and semantic role labeling (SRL), we investigate the performance of existing triplet extraction methods on Open EE. Each title will be segmented and tokenized first, then the extraction is conducted as a token-wise sequence-labeling task. Each token will first be labeled by a SRL module on whether it belongs to a semantic role which appears in one of the S-P-O, S-P, P-O semantic tuples. If not, it will be relabeled by a DP module on whether it appears in a syntactical tuple of the above structures. The entire method is implemented using the LTP toolkit (Che et al., 2020).

5.2 Trigger Extraction

Since the number of triggers per sentence is neither fixed nor given as input, we adopt a token-level sequence tagging model to extract all event triggers in a given sentence based on the inductive bias that event triggers (i.e., predicates) will not overlap with each other (see Section 3.2). Sequence tagging model requires a set of tags where each tag, aligned with a token, represents a part of an event element (i.e., a triplet element) or a non-event element. Then, the model learns the probability distributions of tags for each given sentence, and outputs triplets based on the predicted tags. We adopt the BIO tagging scheme where a token is tagged $B\text{-}trg_i$ ($I\text{-}trg_i$) if it is at the beginning of (inside) the i^{th} trigger, or O if it is outside any trigger. The subscript is used to distinguish between different triggers as they might be discontinuous tokens. Since Title2Event is not annotated on token-level (see 3.2), we perform automatic tagging by locating each annotated event element at the source sentence to get its offset. We use BERT (Devlin et al., 2019) as the sentence encoder to get the contextualized representations of tokens, and each token representation will be fed to a classification layer to compute the probability distribution of the tags.

5.3 Argument Extraction

Argument extraction models take the source sentence and the given triggers as input and output the arguments of each given trigger respectively. Due to the role overlap problem, a token might appear in multiple event arguments and thus has multiple tags, which does not match the common setting of sequence tagging task. Therefore, we iterate over the extracted triggers and extract the arguments of each event trigger separately. We implement three methods for argument extraction.

Sequence Tagging. The first method is a **token-level sequence tagging model** similar to the trigger extraction model, which also uses BIO tagging scheme for subject and object tokens. During each forward process, to specify the current trigger, we adopt the method proposed by Yang et al. (2019). Specifically, the input of BERT encoder is the sum of WordPiece embeddings, position embeddings and segment embeddings, and we set the segment ids of current trigger tokens being one while others being zero to explicitly encode the current trigger.

Methods	Trigger Ex.			Argument Ex.			Triplet Ex.		
	P	R	F1	P	R	F1	P	R	F1
Unsuper.	21.0	32.0	25.4	12.0	15.5	13.5	4.5	6.8	5.4
SeqTag	**69.5**	**69.9**	**69.7**	50.8	51.2	51.0	41.1	41.3	41.2
ST-SpanMRC	-	-	-	**60.1**	54.9	57.4	44.5	44.8	44.7
ST-Seq2SeqMRC	-	-	-	57.9	**58.6**	**58.2**	**49.8**	**50.1**	**49.9**

Table 2: The overall results of trigger extraction (Trigger Ex.), argument extraction (Argument Ex.), and event triplet extraction (Triplet Ex.) on Title2Event. P, R, F1 stand for precision, recall, and f1-score respectively.

Span MRC. The second method is a **span-level tagging model** which formulates argument extraction as a machine reading comprehension (MRC) task, inspired from Du and Cardie (2020) and Liu et al. (2020). For each given sentence as well as a specified trigger, the subject and object are extracted separately by prepending a question, e.g. "动作<trigger>的主体是？" (What is the subject of <trigger>)?, into the sentence to form a context like "[CLS] question [SEP] sentence [SEP]", then the model is asked to extract the answer span from the context for the given question by predicting a start position and an end position. We also use BERT as the context encoder.

Seq2Seq MRC. The third method is a sequence-to-sequence MRC model with same the question design as **Span MRC**. However, instead of extracting the answer spans from the context, it directly generates a sequence of tokens as the output with the given context by maximizing the conditional probability $P(Y \mid S) = \prod_{i=1}^{m} p\left(y_i \mid y_1, y_2, \ldots, y_{i-1}; S\right)$, where $Y =< y_1, \ldots, y_m >$ is the golden answer. We adopt mT5 (Xue et al., 2021), a multilingual text-to-text transformer model as the context encoder as well as the answer decoder.

6 Experiments

We conduct experiments on Title2Event with the methods described in Section 5 and analyze their performance.

6.1 Evaluation Metrics

We adapt the evaluation metrics used in previous works on traditional EE tasks (Li et al., 2021b) to Open EE. We first define the matching criteria: an event trigger or argument is correctly identified if it exactly matches the golden answer, and an event triplet is correctly identified only if all of its elements are correctly identified. We then compute the precision (P), recall (R), and F1-score (F1) for trigger extraction, argument extraction and triplet extraction respectively.

Methods	Argument Ex.			Argument Ex. (Gold)		
	P	R	F1	P	R	F1
SeqTag	50.8	51.2	51.0	70.4	69.6	70.0
SpanMRC	60.1	54.9	57.4	82.9	74.8	78.6
Seq2SeqMRC	57.9	58.6	58.2	80.6	80.4	80.5

Table 3: Results of argument extraction with predicted triggers (Argument Ex.) and with golden triggers (Argument Ex. (Gold))

6.2 Evaluation Model

We summarize all the models we implement for experiments here:

Unsuper. The unsupervised triplet extraction method implemented by the LTP toolkit using the Chinese-ELECTRA-small (Cui et al., 2021) model.

SeqTag. A pipeline tagging-based model consisting of a trigger extractor and an argument extractor, both are based on the token-level sequence tagging model using BERT-base-Chinese as the encoder, and are trained separately. During inference, the argument extractor predicts the arguments based on the triggers predicted by the trigger extractor.

ST-SpanMRC. A pipeline model using a token-level sequence tagging model as the trigger extractor, and a span-level MRC model as the argument extractor, both are based on BERT-base-Chinese.

ST-Seq2SeqMRC. A pipeline model which replaces the argument extractor with a sequence-to-sequence MRC model using mT5-base.

6.3 Overall Experimental Results

Table 2 shows the results of all Open EE methods experimented on Title2Event. It can be observed that: 1) For trigger extraction, the sequence tagging model significantly outperforms the unsupervised

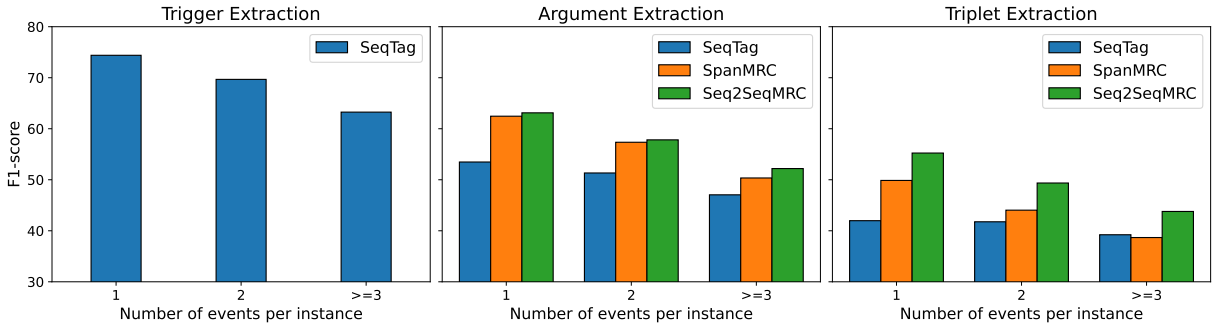

Figure 6: Results of event extraction on instances containing different number of events.

model. 2) For argument extraction and triplet extraction, ST-Seq2SeqMRC outperforms the other tagging-based models. A large part of the reason is that the unconventional writing styles of titles make it difficult to locate token-level tags or span offsets in the source text, while sequence-to-sequence models are free from these restrictions.

6.4 Analysis on Error Propagation

Table 3 shows the results of argument extraction with predicted triggers and with golden triggers. All three models' performance improve by approximately 20% if provided with golden triggers, indicating the huge impact of correct triggers on argument extraction and the urgent need to alleviate the propagating error brought by pipeline architecture in future works.

6.5 Analysis on Multiple Event Extraction

Figure 4 shows that containing multiple events per instance is an important feature of Title2Event, thus we further investigate the models' performance on multiple event extraction, as shown in Figure 6. We can see that as the number of events per instance increases, all models on trigger extraction, argument extraction, and triplet extraction show a decrease in performance, which indicates that multiple events per instance brings additional challenges to open event extraction.

6.6 Analysis on Different Topics

We also investigate the results of trigger extraction and argument extraction on different topics of Title2Event, see Appendix A for details. It can be observed that the F1-scores of "Weather" are higher than other topics, probably because news titles on weather (forecast) usually have a fixed template which makes extraction easier.

Title: **拾荒老人积蓄被盗报案**
Elderly Scavenger's Saving Stolen (and He) Reporting the Case

Golden Answer Events:
(S: 拾荒老人积蓄 Elderly Scavenger's Saving **P: 被盗** Stolen **O: None)**
(S: 拾荒老人 Elderly Scavenger **P: 报案** *Report-the-case* **O: None)**

SegTag: (S: 拾荒老人积蓄 **P: 被盗报案** O: None)

(a)

Title: **18岁以上人群免费筛查!福州首批30万份快筛试剂陆续发放六城区** Free Screening for Over-18s! Fuzhou's First 300,000 Copies of Rapid Screening Reagents are Distributed to Six Urban areas

Golden Answer Events:
(S:福州首批30万份快筛试剂 Fuzhou's First 300,000 Copies of RSR **P: 陆续发放** Distributed to **O: 六城区** Six Urban areas)
(S: 18岁以上人群 Over -18s **P:** 免费筛查 Screen for Free **O: None)**

SegTag:
(S: 福州首批30万份快筛试剂 P: 陆续发放 **O: None**)
(**S: 福州**, P: 免费筛查)

SpanMRC:
(S: 福州首批30万份快筛试剂 P: 陆续发放 **O: None**)
(**S: 六城区人群**, P: 免费筛查)

Seq2SeqMRC:
(S: 福州首批30万份快筛试剂 P: 陆续发放 **O: None**)
(**S: 六城区人群**, P: 免费筛查)

(b)

Title: **遭皇马逆转,大巴黎主席大闹裁判更衣室** Reversed by Real Madrid, the President of PSG Causes Trouble to the Referee's Dressing Room

Golden Answer Events:
(S: 大巴黎 PSG **P: 遭逆转** Reversed by **O: 皇马** Real Madrid**)**
(S: 大巴黎主席 President of PSG Causes **P: 闹** *Cause-Trouble* **O: 裁判更衣室** Referee's Dressing Room**)**

SegTag:
(S: **皇马** P: 遭逆转 **O: 大巴黎主席**)
(S: 大巴黎主席 P: 闹 O: 裁判更衣室)

SpanMRC:
(**S:大巴黎主席** P: 遭逆转 O: 皇马)
(S: 大巴黎主席 P: 闹 O: 裁判更衣室)

Seq2SeqMRC:
(**S:大巴黎主席** P: 遭逆转 O: 皇马)
(S: 大巴黎主席 P: 闹 O: 裁判更衣室)

(c)

Figure 7: Example error cases, arguments in bold text and underlined are the specific errors compared to golden answers.

6.7 Analysis on Error Cases

We summarize three typical challenges observed in Title2Event in Section 1. Here, we analyze some error cases of the model outputs to further demonstrate the issues. Figure 7 (a) shows an error output in trigger extraction, where the given title is unconventionally written by concatenating two predicates. As a result, SegTag is unable to distinguish the two different predicates. Figure 7 (b) shows an instance with multiple events and all the models mix up the argument roles. Figure 7 (c) shows a sport news title, without the background that *Real Madrid* and *PSG* are both football clubs, none of the models properly understand the event that *PSG* is defeated by *Real Madrid*. All of the above cases clearly address the challenges present in Title2Event, which are also common in real-world scenarios, and require advanced study to be better solved.

7 Conclusions

In this paper, we present Title2Event, a Chinese title dataset benchmarking the task of open event extraction. To the best of our knowledge, Title2Event is the largest manually-annotated Chinese dataset for sentence-level event extraction. We experiment with different methods and conduct detailed analysis to address the challenges observed in Title2Event, which are rather scarce in existing datasets yet common in real-world scenarios. We believe Title2Event could further facilitate advanced research in event extraction.

Limitations

We summarize the limitations of Title2Event as follows:

Evaluation Metrics. We make Title2Event a benchmark for open event extraction with a hope that it could evaluate the performance of domain-general EE models. We adapt the formulation of Open IE and represent events in a universal triplet format while adopting traditional EE metrics which is based on exact match. However, we observe that the narrative of events in Chinese titles are extremely diverse. To unify them into the triplet format without losing the core event information, we design detailed annotation guidelines which results in the fact that the a large amount of triplet elements are text spans instead of one or two tokens which is common in traditional EE datasets such as ACE 2005. Therefore, using exact match in Title2Event might be too strict for model outputs which are just one or two tokens different from the golden text span. We leave the design of fine-grained evaluation approaches to future work.

Methods. Some characteristics of Title2Event such as unfixed number of events per instance and the role overlap problem bring difficulties to the model design. We adopt a pipeline architecture which suffers from the error propagation problem as discussed in Section 6.4. We also adapt some end-to-end models in traditional EE such as TEXT2EVENT proposed by Lu et al. (2021) to our Open EE benchmark, but find the performance is unexpectedly poor. We conduct preliminary analysis and find that the length of text span in triplets (as mentioned above) as well as the relatively complex linearized event structures (largely due to the multiple events per instance issue) are the potential factors of the limited performance. Therefore, we do not provide a good end-to-end model as baseline, which might make the model comparison in Section 6 less comprehensive. However, we hope that future works could pay more attention to the design of text-to-structure models except from traditional tagging-based models.

Ethics Statement

As Title2Event is an Open EE dataset which broadly collects contents of various categories on the Internet, keeping the corpus without bias is extremely difficult. However, we put large efforts in cleaning the toxicity of data. First, all crawled web paged are automatically removed if they contain toxic contents using an existing system. During annotation, all instances will be dual checked by the human annotators and manually deleted if not passing the check. Moreover, in our annotation standard, we ask annotators to label only factual events while ignoring all subjective opinions, as we hope Title2Event could be factual and unbiased.

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, Giovanni Semeraro, and Lucia Siciliani. 2014. Extending an information retrieval system through time event extraction. In *DART@ AI* IA*, pages 36–47.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. CaRB: A crowdsourced benchmark for open IE. 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)*, pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2020. N-ltp: A open-source neural chinese language technology platform with pretrained models. *ArXiv preprint*, abs/2009.11616.

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, Beijing, China. Association for Computational Linguistics.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 407–413, Melbourne, Australia. Association for Computational Linguistics.

Shiyao Cui, Bowen Yu, Xin Cong, Tingwen Liu, Quangang Li, and Jinqiao Shi. 2020. Label enhanced event detection with heterogeneous graph attention networks. *ArXiv*, abs/2012.01878.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Ziran Li, Zhiyuan Liu, Haitao Zheng, and Zibo Lin. 2019. Event detection with trigger-aware lattice neural network. 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)*, pages 347–356, Hong Kong, China. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Jun Gao, Wei Wang, Changlong Yu, Huan Zhao, Wilfred Ng, and Ruifeng Xu. 2022. Improving event representation via simultaneous weakly supervised contrastive learning and clustering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3036–3049, Dublin, Ireland. Association for Computational Linguistics.

Simon Gottschalk and Elena Demidova. 2018. Eventkg: a multilingual event-centric temporal knowledge graph. In *European Semantic Web Conference*, pages 272–287. Springer.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020a. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3748–3761, Online. Association for Computational Linguistics.

Keshav Kolluru, Samarth Aggarwal, Vipul Rathore, Mausam, and Soumen Chakrabarti. 2020b. IMoJIE: Iterative memory-based joint open information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5871–5886, Online. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Guozheng Li, Peng Wang, Jiafeng Xie, Ruilong Cui, and Zhenkai Deng. 2021a. Feed: A chinese financial event extraction dataset constructed by distant supervision. In *The 10th International Joint Conference on Knowledge Graphs*, pages 45–53.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and Philip S. Yu. 2021b. A compact survey on event extraction: Approaches and applications.

Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. 2020b. Duee: A large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing*, pages 534–545, Cham. Springer International Publishing.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2018. Nugget proposal networks for Chinese event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1565–1574, Melbourne, Australia. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Chenwei Lou, Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, Weiwei Tu, and Ruifeng Xu. 2022. Translation-based implicit annotation projection for zero-shot cross-lingual event argument extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2076–2081.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6851–6858. AAAI Press.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shaina Raza and Chen Ding. 2020. A survey on news recommender system - dealing with timeliness, dynamic user interest and content quality, and effects of recommendation on news readers. *ArXiv preprint*, abs/2009.04964.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2300–2305, Austin, Texas. Association for Computational Linguistics.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Sandra A Thompson. 1973. Transitivity and some problems with the bǎ construction in mandarin chinese. *Journal of Chinese Linguistics*, 1(2):208–221.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

Xi Xiangyu, Zhang Tong, Ye Wei, Zhang Jinglei, Xie Rui, and Zhang Shikun. 2019. A hybrid character representation for chinese event detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Nuo Xu, Haihua Xie, and Dongyan Zhao. 2020. A novel joint framework for multiple Chinese events extraction. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 950–961, Haikou, China. Chinese Information Processing Society of China.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Changlong Yu, Hongming Zhang, Yangqiu Song, Wilfred Ng, and Lifeng Shang. 2020. Enriching large-scale eventuality knowledge graph with entailment relations. In *AKBC*.

Yin Zeng, Honghu Yang, Yansong Feng, Zhen Wang, and Dongyan Zhao. 2016. A convolution bilstm neural network model for chinese event extraction. In *Natural Language Understanding and Intelligent Applications*, pages 275–287, Cham. Springer International Publishing.

Junlang Zhan and Hai Zhao. 2020. Span model for open information extraction on accurate corpus. 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 9523–9530. AAAI Press.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. 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)*, pages 337–346, Hong Kong, China. Association for Computational Linguistics.

Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, and Jian Sun. 2022. A survey on neural open information extraction: Current status and future directions.

A Appendix

Annotation Tool. Figure 8 shows a screenshot of our annotation web page. The raw title are given with auxiliary information, the annotators will first determine whether to abandon this case as well as is this case easy to annotate or not. Then, they will type all plausible events in the text boxes following our annotation guidelines.

Topic List. Table 4 lists all 34 topics and their corresponding number of instances.

Topic	Count
社会 (Society)	12,341
财经 (Finance)	6,539
体育 (Sports)	5,033
时事 (Current Events)	4,499
科技 (Technology)	2,965
娱乐 (Entertainment)	1,685
教育 (Education)	1,451
汽车 (Cars)	1,319
天气 (Weather)	1,013
军事 (Military)	712
旅游 (Travel)	659
房产 (Real Estate)	647
三农 (Agriculture)	520
文化 (Culture)	501
综艺 (Variety Shows)	412
游戏 (Games)	396
电影 (Movies)	348
健康 (Health)	344
电视剧 (TV Series)	233
历史 (History)	220
音乐 (Music)	159
科学 (Science)	147
生活 (Life)	118
美食 (Food)	117
情感 (Sentiment)	95
育儿 (Childcare)	73
时尚 (Fashion)	60
宠物 (Pets)	57
职场 (Career)	54
曲艺 (Folk Art)	41
动漫 (Animation)	34
摄影 (Photography)	24
搞笑 (Funny News)	12
其它 (Others)	11

Table 4: The topics in Title2Event with their number of instances.

Results on Different Topics. Figure 9 shows the F1-scores of trigger extraction (using SeqTag model) and argument extraction with golden triggers (using SeqTag, SpanMRC, and Seq2SeqMRC models) on the top-10 topics in Title2Event.

Hyper-parameter Settings in Training. For all models, we use the batch size of 32 and train them for 30 epochs on the training set of Title2Event. All models are trained on a single Tesla A100 GPU. We use the linear learning rate scheduler and AdamW as the optimizer. For models based on Bert-base-Chinese, we set the learning rate to be 5e-5; For models based on mT5-base, we set the learning rate to be 1e-4. All supervised models are implemented using the Huggingface-transformers library.

Figure 8: Screenshot of our annotation web page.

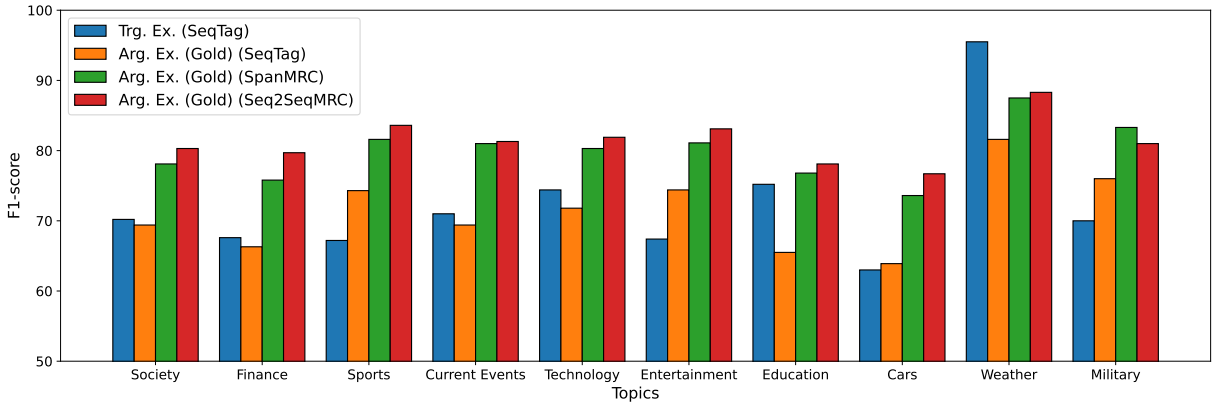

Figure 9: Results of trigger extraction and argument extraction with golden triggers on top 10 topics in Title2Event