MUSIED: A Benchmark for Event Detection from Multi-Source Heterogeneous Informal Texts

Xiangyu Xi[1*], **Jianwei Lv**[1*], **Shuaipeng Liu**[1*†], **Wei Ye**[2†], **Fan Yang**[1], **Guanglu Wan**[1]
[1] Meituan Group, Beijing, China
[2] National Engineering Research Center for Software Engineering, Peking University, Beijing, China
liushuaipeng@meituan.com,wye@pku.edu.cn

Abstract

Event detection (ED) identifies and classifies event triggers from unstructured texts, serving as a fundamental task for information extraction. Despite the remarkable progress achieved in the past several years, most research efforts focus on detecting events from formal texts (e.g., news articles, Wikipedia documents, financial announcements). Moreover, the texts in each dataset are either from a single source or multiple yet relatively homogeneous sources. With massive amounts of user-generated text accumulating on the Web and inside enterprises, identifying meaningful events in these informal texts, usually from multiple heterogeneous sources, has become a problem of significant practical value. As a pioneering exploration that expands event detection to the scenarios involving informal and heterogeneous texts, we propose a new large-scale Chinese event detection dataset based on user reviews, text conversations, and phone conversations in a leading e-commerce platform for food service. We carefully investigate the proposed dataset's textual informality and multi-source heterogeneity characteristics by inspecting data samples quantitatively and qualitatively. Extensive experiments with state-of-the-art event detection methods verify the unique challenges posed by these characteristics, indicating that multi-source informal event detection remains an open problem and requires further efforts. Our benchmark and code are released at https://github.com/myeclipse/MUSIED.

1 Introduction

Event detection (ED), which aims to identify event triggers and classify them into specific types from unstructured texts, has been widely researched and applied in various downstream tasks (Basile et al., 2014; Cheng and Erk, 2018; Kuhnle et al., 2021). Advanced models have been continuously proposed, ranging from feature-based models (Shasha et al., 2010; Hong et al., 2011; Li et al., 2013) to recent neural-based models (Chen et al., 2015; Nguyen et al., 2016; Chen et al., 2018; Xi et al., 2021; Xiangyu et al., 2021). Despite the significant progress, we find that previous works have the following two limitations in practical scenarios.

1. Current efforts mainly focused on event detection from formal texts. For example, a popular line of works (Li et al., 2013; Chen et al., 2015; Nguyen et al., 2016; Chen et al., 2018; Lou et al., 2021) aim to detect general domain events from news articles (e.g., ACE 2005 (Doddington et al., 2004)) and Wikipedia documents (e.g., MAVEN (Wang et al., 2020b)). Some other explorations involve extracting events from the financial announcements (Yang et al., 2018; Zheng et al., 2019; Liang et al., 2021) or cybersecurity articles(Trong et al., 2020), which are also written in a relatively official style. In practical scenarios, however, we usually face the bottleneck of identifying events from informal texts. Compared with formal text, texts produced in more casual contexts (e.g., online chat and phone conversation) pose some unique challenges of long event triggers, high event density, and typos noises, as revealed in our analysis (§ 4.3). Indeed, with vast amounts of user-generated text accumulating on the open Web and private enterprise systems, extracting meaningful events in these informal texts has become an urgent problem of significant practical value.

2. The targeting event-related texts are either from a single source or multiple yet homogeneous sources. Most recent datasets (e.g, MAVEN (Wang et al., 2020b), CySecED (Trong et al., 2020), ChFinAnn (Yang et al., 2018), and BRAD (Lai et al., 2021)) are built from an individual data source. The most widely-used ACE 2005 (Doddington et al., 2004) covers six sources, which are, however, relatively homogeneous internet media to some extent. Regarding informal text, end-users can produce them in many different ways, and

[*] The first three authors contributed equally.
[†] Corresponding authors.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2947 - 2964
December 7-11, 2022

hence they have more versatile expressing styles. Therefore, multi-source heterogeneity comes as another difficulty that inherently accompanies text informality. For example, texts generated via online chat and phone calls in after-sales services may greatly diversify, e.g., on length and style. Unfortunately, current ED works fail to adequately address the issue of multi-source heterogeneity.

To address these two problems, in this paper, we expand event detection to the scenarios involving informal and heterogeneous texts. We construct a new large-scale Chinese event detection dataset based on Meituan*, the most popular Chinese e-commerce platforms for food service, which provides users with multiple ways to feed back on food safety issues (events), such as posting reviews and communicating with after-sale staff. These reviews and conversations yield a large-scale multi-source heterogeneous informal text repository, which contains valuable information about food safety events and hence can serve as a corpus. We collect the desensitized data from three typical scenarios: i) users posting reviews, ii) users communicating with after-sale staff through text messages, and iii) users communicating with after-sale staff on the phone. By extracting user reviews, text conversations, and phone conversations, we create a massive dataset consisting of MUlti-Source heterogeneous Informal texts for Event Detection (MUSIED).

We investigate MUSIED's textual informality (§ 4.3) and multi-source heterogeneity (§ 4.4) by carefully inspecting data samples. The textual informality leads to event descriptions involving long triggers (§ 4.3.1), multi-event sentences (§ 4.3.2), and user typos (§ 4.3.3), while the multi-source heterogeneity brings notable diversity of event type distribution and event density across domains (§ 4.4). We re-implement the state-of-the-art ED methods and conduct extensive evaluation on MUSIED (§ 5). The experimental results clearly verify the unique challenges posed by the above characteristics. Specifically, the proposed dataset requires more robust models towards identifying long triggers (§ 5.4.1), capturing multi-event interaction (§ 5.4.2), and alleviating typo noises(§ 5.4.3). Meanwhile, MUSIED also facilitates future research on tackling multi-source heterogeneity, e.g., with multi-domain learning and (§ 5.5.1) and domain adaptation (§ 5.5.2).

Our contributions can be summarized as follows:

- We expand event detection to the scenarios involving informal and heterogeneous texts, for the first time, by carefully curating a new large-scale dataset.
- Extensive experiments with state-of-the-art methods verify the unique challenges posed by textual informality and multi-source heterogeneity characteristics, and indicate multiple promising directions worth pursuing.

2 Event Detection Definition

We follow the classical settings and terminologies adopted by ACE 2005 program (Doddington et al., 2004) and MAVEN (Wang et al., 2020b), and specify the vital event terminologies as follows. **Event**: a specific occurrence involving participants (location, time, subject, object, etc.). **Event Mention**: a phrase or sentence within which an event is described. **Event Trigger**: the main word or phrase that most clearly expresses the occurrence of an event. **Event Type**: the semantic class of an event.

Event detection aims to identify event trigger words and classify their event types for a given text. Accordingly, ED is conventionally divided into two subtasks: (1) **Trigger identification**, which aims to identify the event triggers. (2) **Trigger classification**, which aims to classify the recognized trigger into predefined categories. Both subtasks are evaluated with micro precision, recall, and F-1 scores. Most recent works (Chen et al., 2015; Nguyen et al., 2016; Chen et al., 2018; Wang et al., 2019b) perform trigger classification directly (add an additional type "N/A" to be classified at the same time, indicating that the candidate is not a trigger). We also inherit these settings in this paper.

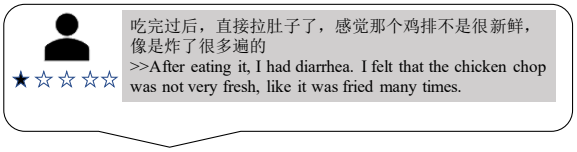
3 Data Collection and Annotation

3.1 Data Collection

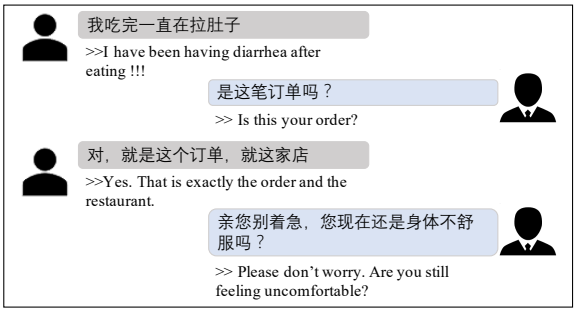
We collect data from Meituan, which provides users with multiple channels to feed back on food safety issues (events), among which the three most common ways are: i) users post reviews to restaurants where they have ordered food; ii) users communicate with after-sale staff through text messages; iii) users communicate with after-sale staff on the phone. First, we collect the user reviews, text conversations, and phone conversations from logs of online services for a week. Further, we desensitized and anonymized the private information from the raw data (see § 7 for details). The

*https://about.meituan.com/en

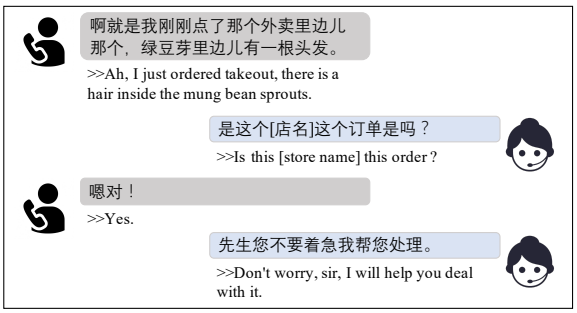
samples from each scenario are shown in Figure 1 to promote understanding. Note that the phone conversations are speech data, which is transformed into text data via the Automatic Speech Recognition (ASR) service (Wang et al., 2019a; Kaur et al., 2021).



(a) Sample of user reviews



(b) Sample of text conversations



(c) Sample of phone conversations

Figure 1: Samples of our corpus.

The above collected data may not involve food safety events (e.g., users make positive reviews). We hire annotators to select the reviews and conversations involving food safety incidents. Finally, we retained 4,226 user reviews, 3,767 text conversations, and 3,388 phone conversations, forming a corpus composed of 11,381 documents in total.

3.2 Event Schema Construction

With the assistance of food safety experts, we construct an event schema, from the perspective of users. We exemplify using a typical food delivery service scenario shown in Figure 2, where users usually feed back in terms of: (1) **Food quality** Poor food quality is the main cause of food safety problems (e.g., food is expired or undercooked). (2) **Restaurant** The illegal or improper behaviors of restaurants (e.g., uses illegal food additives) may lead to food safety problems. (3) **Delivery person** A small but noticeable percentage of food safety problems are caused by the delivery person (e.g., damages the packaging and pollutes the food). (4) **Physical feelings** Rather than above causes, the users may directly express their physical feelings (e.g., feel uncomfortable), which suggest the existence of food safety problems. Finally, the schema contains 21 event types and broadly covers the user's feedback about above cases. Please refer to Appendix A for the full event schema description.

Figure 2: A typical food delivery service scenario.

3.3 Data Annotation

3.3.1 Annotation Process

Though with a detailed annotation guideline, the annotation process is complicated and error-prone. For accuracy and consistency, we organize a two-stage iterative annotation, following ACE 2005 (Doddington et al., 2004) and MAVEN (Wang et al., 2020b). We recruit 20 annotators with food safety domain knowledge, and train them with the guideline. After that, they are given an annotation exercise and 9 annotators with accuracy $> 90\%$ are selected to perform formal annotation. At the first stage, each document is annotated by 3 independent annotators. The annotation is finished if and only if 3 annotators reach an agreement. Otherwise, in the second stage, all 9 annotators and language experts will discuss documents with annotation disagreements together and determine the final results.

3.3.2 Annotation Challenges And Solutions

Candidate Selection Since Chinese lacks natural delimiters, words are necessarily generated by segmentation toolkits, which might not exactly match with triggers (Zeng et al., 2016; Lin et al., 2018). Also, the informal texts are more diverse. It would be impractical and inaccurate to select words with specific features, as done in English dataset (Wang et al., 2020b). To address above challenge, we annotate in a character-wise manner, instead of performing word segmentation and word-wise annotation sequentially. In this way, though the trigger candidate set is larger because each possible phrase

is regarded as a candidate trigger, we tackle the problem of i) limitation of word boundary and 2) error propagation of word segmentation toolkits.

Boundary confusion During annotation, we find the triggers are usually followed or surrounded by stop words (such as auxiliary words, modal particles, etc), especially in telephone conversations. We follow the principle that event triggers should not contain redundant information, as long as they can fully express the event information. For example, we do not annotate the modal particles in the following sentence S1. "臭(*stinky*)" and "吃吐(*Eat and vomit*)" are the triggers of *Abnormalities* and *Uncomfortable* event. However, the token "的" and "了" following them are modal particles in Chinese, and do not express useful information.

S1: *The duck intestines were stinky, I Eat and vomit.* (鸭肠是臭的，把人都吃吐了)

Ambiguous User Expression The informal user statements are not rigorous and may be insufficient for resolving ambiguities for event types. For example, for the trigger "梆硬(*hard*)" in the following sentence S2, some annotators believe the reason for "梆硬(*hard*)" is that the chicken is undercooked and considers it as a trigger of *Undercooked* event, while others think the reason is that the temperature is too low and treats it as a trigger of *Cold* event.

S2: *I felt that the chicken chop was cold, and the chicken in the chicken roll was also hard* (感觉鸡排冷了，鸡肉卷里的鸡肉也是梆硬的。)

The annotators are required to disambiguate by integrating contextual information. For example, considering the context that the user first complains that the chicken chop is cold (i.e., "冷 (*cold*)"), the annotators tend to believe the following phrase "梆硬 (*hard*") also triggers a *Cold* event.

3.3.3 Annotation Quality

With the strict annotation process, our dataset is of high quality. For data with annotation disagreement in the first stage, all annotators discuss together and reach agreements (by voting sometimes). Also, we randomly sample 500 documents without annotation disagreement in first stage, and invite different first-stage annotators to annotate these documents. We measure the inter-annotator agreements of annotation between two annotators with Cohen's Kappa score. The results for trigger and type annotation are 0.83 and 0.82 respectively, which belongs to the Near-perfect agreement range of [0.81, 0.99]. The annotated samples are shown in Appendix C.

4 Data Analysis

4.1 Data Size

Following Wang et al. (2020b), we show the main statistics of MUSIED and compare with the following datasets in Table 1: (1) **ACE 2005** (Walker et al., 2006), which is the most wide-used dataset and covers general domain events. (2) **Rich ERE** (Mitamura et al., 2015), which is provided by TAC KBP competition and contains a series of datasets; (3) **MAVEN** (Wang et al., 2020b), which is the largest general domain dataset constructed from Wikipedia and FrameNet; (4) **RAMS** (Ebner et al., 2020), which follows the AIDA ontology and uses Reddit articles. (5) **BRAD** (Lai et al., 2021), which covers Black Rebellions events in African Diaspora; (6) **CySecED** (Trong et al., 2020), which is the largest cybersecurity event dataset. We can observe that our MUSIED is large-scale compared with existing datasets. In terms of average instance number per event type, MUSIED has significant advantage over other datasets (e.g., 1,756 of MUSIED v.s. 707 of MAVEN v.s. 162 of ACE 2005). Thus, MUSIED can stably train and benchmark sophisticated neural-based models.

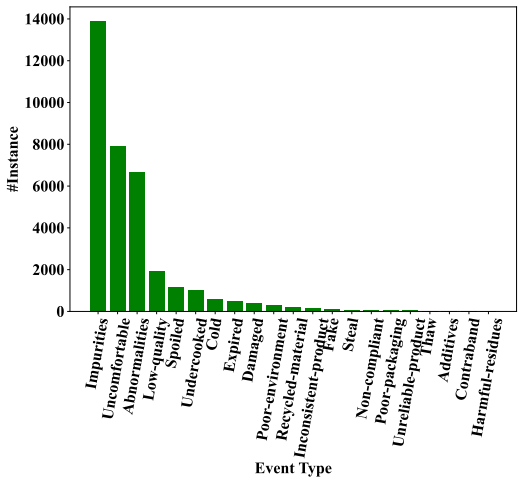


Figure 3: Instance number of each event type.

4.2 Data Distribution

The instance number of each event type is shown in Figure 3, which shows the existence of the inherent data imbalance problem. We also display the top 5 event types with their instance numbers and proportions in Appendix B.1. MUSIED has 21 event types and 35,313 labeled instances, yet "Impurities" (with 13,883 labeled instances) and "Uncomfortable" (with 7,935 labeled instances) ac-

Dataset	#Doc	#Tokens	#Sentences	#Event Types	#Events	#Event Mentions
ACE 2005 English	599	303k	15,789	33	4,090	5,349
ACE 2005 Chinese	633	321k	7,269	33	2,521	3,333
ACE 2005 Arabic	403	150k	2,710	33	2,267	2,270
Rich ERE	1,272	854k	41,708	38	29,293	38,853
MAVEN	4,480	1,276k	49,873	168	111,611	118,732
RAMS	3,993	1,218k	44,236	139	9,124	9,124
BRAD	151	172k	5,638	12	-	4,259
CySecED	300	-	290,234	30	-	8,014
MUSIED	11,381	7,105k	315,473	21	30,940	35,313

Table 1: Dataset statistics.

count for 61.7% of the data. 18 (85.7%) event types have a below-average number of labeled instances and 6 event types even have fewer than 50 labeled instances. Though potentially hindering the performance of ED models, the occurrence frequency of event types conforms to the long-tail phenomenon in the real world. We maintain the original distribution of MUSIED, which can evaluate the ability of the ED models in the long-tail scenario.

4.3 Analysis of Textual Informality

A key characteristic of MUSIED is that the corpus is composed of informal text. We introduce the features brought by textual informality as follows.

4.3.1 Long Triggers

Our observation shows that users tend to use more casual expressions and longer triggers to express events. For example, in the following sentence S3, the user says his/her two teeth are broken due to the hard noodles. The phrase "牙齿都干掉两颗 (*two teeth are broken*)" triggers an *Uncomfortable* event and consists of 7 tokens.

S3: *The rice is rotten, noodles are as hard as steel wire, two teeth are broken* (米饭稀烂，面条跟钢丝条一样硬，牙齿都干掉两颗)

MUSIED contains a much higher proportion of long triggers, as Figure 4 shows. Considering the proportion of triggers consisting of more than 2 tokens, MUSIED is nearly 53 times larger than ACE 2005 English (i.e., 26.97% v.s. 0.50%) and 9 times larger than ACE 2005 Chinese (i.e., 26.97% v.s. 3.06%). The long trigger phenomenon poses a great challenge to existing ED models.

4.3.2 Multiple Events

Unlike professionals who write articles or documents in a relatively official style, users may hurriedly express multiple events within one sentence.

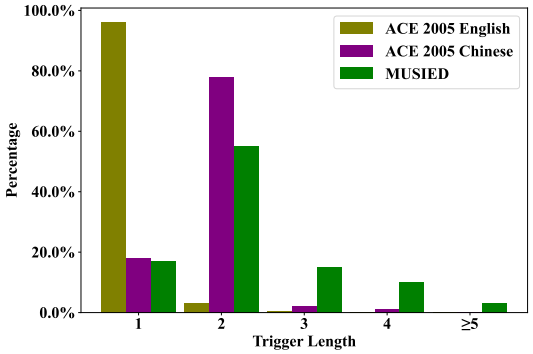


Figure 4: Distribution of triggers with different length.

For example, in the following sentence S4, the user reports multiple food quality related events, which lead to an *Uncomfortable* event.

S4: *Then I ate his fried rice. Because his prawns were not fresh and undercooked. Then I had his grilled sausages and sausages, and it all didn't feel very fresh. After eating, I had diarrhea.* (然后我吃了他那个炒饭因为他那个虾不新鲜然后也不熟然后再加上他那个烤肠啊腊肠啊都是感觉不是很新鲜然后吃了之后我我拉肚子)

Following previous works (Chen et al., 2015), we make statistics on sentences with multiple events and find that the proportion of multi-events sentence in MUSIED is much larger than ACE 2005 (i.e., 36.9% of MUSIED v.s. 27.3% of ACE 2005 English v.s. 19.3% of ACE 2005 Chinese). The reason lies in that food safety event correlations are closer and users tend to simultaneously express the cause and consequence.

4.3.3 Typos

Different from formal texts which are produced by professionals, the user-generated informal texts are less rigorous and may contain typos. The automatic speech recognition service may also produce errors. For example, in the following sentence S5,

the user finds the beef rice is sour and expresses a *spoiled* event. However, the user types a typo token "搜" (means search), which has the same pronunciation (pronounced as "sou" in Chinese) but different meaning as the token "馊" (means sour).

S5: *I ordered beef rice, it looks search(sour) already* (我点的牛肉饭，看起来都搜(馊)掉了)

We make statistics on the typos using the state-of-the-art spelling error corrector (SEC) (Li et al., 2021). The result shows that 2.2% of sentences contain spelling errors, 0.1% of tokens are typos and 1.5% of them are within the triggers. Though the accuracy of SEC may be limited in our corpus, the result together with our observation reveals that existence of typos is a noticeable problem.

4.4 Analysis of Multi-Source Heterogeneity

In this section, we analyze the multi-source heterogeneity from the following perspectives.

4.4.1 General Textual Features

The textual features shift remarkably across sources of MUSIED. We present the statistics on each source in Appendix B.2, from which we can observe that document size varies significantly (i.e., 1.4 of user reviews v.s. 48.1 of text conversations v.s. 37.8 of phone conversations in terms of #sentences per document). The reason lies in that conversation with staff is more official and users tend to provide more complete information. Also, we calculate the average sentence length for each source and further compute the standard deviation of the average sentence lengths. The standard deviation of MUSIED is notably larger than ACE 2005 (i.e., 5.06 of MUSIED v.s. 3.31 of ACE 2005 English v.s. 3.87 of ACE 2005 Chinese).

4.4.2 Event Type Distribution and Event Density

The event type distribution and event density vary significantly across sources of MUSIED. The top 5 event types for each source are shown in Appendix B.1, from which we can easily observe the notable diversity of event type distributions across sources of MUSIED. For a quantitative analysis, we calculate the event type distribution for each source and calculate the wasserstein distance (Vallender, 1974) between the distributions (please refer to Appendix B.3 for the detailed calculation procedure). MUSIED is much larger than ACE 2005 (i.e., 6.17×10^{-4} of MUSIED v.s. 3.32×10^{-4} of ACE 2005 English v.s. 1.51×10^{-4} of ACE 2005 Chinese), in terms of the average wasserstein distance. Also, we compute the average event density for each source and the standard deviation of the average event densities, which shows that the disparity of event density across MUSIED's sources is more remarkable (i.e., 0.35 of MUSIED v.s. 0.17 of ACE 2005 English v.s. 0.08 of ACE 2005 Chinese).

To sum up, MUSIED is of more significant heterogeneity and can effectively support the exploration of ED involving multi-source heterogeneity. Conversely, the limited heterogeneity, together with the data scarcity problem, makes ACE 2005 insufficient for benchmarking relevant research.

5 Experiments

5.1 Benchmark Settings

We randomly split the annotated documents into train, dev, and test sets with the ratio of 8:1:1. The statistics of the three sets are shown in Table 2.

Set	#Doc	#Sentences	#Event Mentions
Train	9,103	252,786	28,012
Dev	1,139	31,269	3,540
Test	1,139	31,418	3,761

Table 2: The statistics of splitting MUSIED.

5.2 Experimental Settings

Recently, neural-based models have achieved significant progress. Thus, we investigate the following state-of-the-art neural-based methods, which can be roughly divided into two categories:

Sentence-Level Models which use information within the sentence to extract triggers. **DMCNN** (Chen et al., 2015) which uses CNN as feature extractor and concatenates sentence and lexical feature; **BiLSTM** (Hochreiter and Schmidhuber, 1997) which uses bi-directional long short-term memory network as encoder; **BiLSTM-CRF** (Lafferty et al., 2001) which uses bi-directional long short-term memory network followed by a conditional random field layer; **C-BiLSTM** (Zeng et al., 2016) which proposes a convolution bidirectional LSTM to capture both sentence-level and lexical information; **DMBERT** (Wang et al., 2019b) which takes BERT as encoder and adopts the dynamic multi-pooling mechanism; **BERT** (Yang et al., 2019) which fine-tune BERT on the down-stream ED task via a sequence labeling manner.

Document-Level Models which integrate the document-level contextual information. **HBT-**

NGMA (Chen et al., 2018) which dynamically fuses the sentence- and document-level information; **MLBiNet** (Lou et al., 2021) which captures the document-level association of events.

The implementation details such as hyperparameters are listed in Appendix D. Following previous works, we report *Precision* (P), *Recall* (R) and *F1-Score* (F1) on trigger classification.

Model	P	R	F1
DMCNN	**84.2**	56.8	67.8
BiLSTM	75.6	66.4	70.7
BiLSTM+CRF	76.0	69.8	72.8
C-BiLSTM	75.7	70.5	73.0
DMBERT	77.0	68.7	72.7
BERT	72.6	78.9	75.6
HBTNGMA	73.1	**79.5**	**76.2**
MLBiNet	73.4	69.3	71.3

Table 3: Performance on trigger classification (%).

5.3 Overall Experimental Results

The overall experimental results are shown in Table 3, from which we have the following observations: (1) Sequence labeling methods have advantages over token-level classification models. For example, BiLSTM and BERT achieve 2.9 and 2.9 F1 improvements over DMCNN and DMBERT respectively. The reason lies in that token-level classification models separately predict trigger candidates without considering the event interdependency, while sequence labeling methods generate representation and make predictions collectively. (2) BiLSTM+CRF achieves notable improvements over BiLSTM (e.g., 72.8 v.s. 70.7 in terms of F1), with the assistance of CRF layer modeling event correlations. The observation confirms our analysis in § 4.3.2 that modeling event correlations is important for MUSIED, due to the multi-event sentences. (3) By incorporating document-level contextual information, HBTNGMA gains an absolute improvement of 3.4 F1-Score over BiLSTM+CRF (i.e., 76.2 v.s. 72.8). The experiment result is consistent with our observation of ambiguous user expression (§ 3.3.2), and clearly indicates the importance of document-level contextual information.

5.4 Analysis of Textual Informality

5.4.1 Challenge of Long Triggers

As § 4.3.1 shows, MUSIED contains long triggers, due to the informal expressions. We make statistics on BERT's recall on triggers of different lengths, as Table 4 shows, from which we can easily observe a general trend that the longer the trigger, the worse the recall rate. Existing ED models have difficulty in capturing the distribution pattern of long triggers, and the challenge should be further addressed.

Length	[1,2]	[3,4]	[5,)
Recall	80.5	79.6	34.0

Table 4: BERT's recall of triggers of different lengths.

5.4.2 Challenge of Multi-Event Sentences

Following Chen et al. (2015), we divide the test set into two parts according to the event number in a sentence (single event (i.e., 1/1) and multiple events (i.e., 1/N)), and perform evaluation separately. From Table 5 we can observe that: (1) All models perform much worse on 1/N, which coincides with previous findings (Chen et al., 2015, 2018). (2) Though achieving comparable performance in 1/1 data, sequence labeling methods have significant advantage over token-level classification methods on 1/N data (i.e., 42.6 of DMCNN v.s. 60.4 of BiLSTM, 55.1 of DMBERT v.s. 72.1 of BERT). The experimental results indicate that it is worth exploring more collectively-detecting methods, to better capture the distribution pattern of multiple events within a sentence.

Model	1/1	1/N	All
DMCNN	79.1	42.6	67.8
BiLSTM	79.6	60.4	70.7
DMBERT	82.4	55.1	72.7
BERT	84.3	72.1	75.6

Table 5: F1-Scores on Single Event Sentences (**1/1**) and Multiple Event Sentences (**1/N**).

5.4.3 Challenge of Typos

We use the state-of-the-art spelling error corrector (SEC) (Li et al., 2021) on the test set, then manually collect the samples that are indeed typos. Further, we retest these corrected samples with BERT, as Table 6 shows. After correction, some mislabeled samples can be fixed and the performance is improved. For example, the S5 in § 4.3.3 can be correctly predicted. Another concrete case is shown in Appendix F.2 to promote understanding.

Though of great potential to address the typo challenge, our sampling statistics show the precision of SEC is quite limited in our corpus (47.8%).

Sampled Instances	P	R	F1
-w/o Correction	63.6	41.2	50.0
-w/ Correction	66.7	47.1	55.2

Table 6: Performance on corrected samples.

One possible reason is the textual features of our corpus are quite different from the SEC's training corpus. We believe that developing a SEC more suitable for MUSIED and exploring more sophisticated methods such as incorporating pronunciation features may be useful to address the challenge.

5.5 Analysis of Multi-Source Heterogeneity

Since the different sources of MUSIED have diversified data characteristics, we investigate the multi-source heterogeneity via the following two typical research topics (i.e., multi-domain learning and domain adaptation). Following Pradhan et al. (2013); Ganin and Lempitsky (2015); Chen et al. (2021); Wang et al. (2020a), we treat each source as a single "domain" in the following investigation.

5.5.1 Analysis of Multi-Domain Learning

So far, we exploit a standard strategy by naively pooling all available data across domains (sources) and discarding the domain information. A shared model is trained to serve all domains. However, the multi-source heterogeneity drives us to explore ways to utilize the domain information. Following Chen and Cardie (2018); Wang et al. (2020a), we select BERT and experiment with the following multi-domain learning strategies: (1) **SingleDomain (SD)** which trains an individual ED model for each domain separately and only uses the training data for the single domain. (2) **PoolDomain (PD)** which is the strategy we used. The model ignores domain information, albeit uses all available training data. (3) **PoolDomain-MultiTask (PDMT)** which is similar to PoolDomain, except that we add an auxiliary task that learns domain labels. The domain information is utilized, though in a simple way. (4) **MultiDomain-Shared-Private (MDSP)** which uses i) a shared MLP for all domains that extracts generic and domain-invariant features; and ii) a private MLP for each domain that extracts domain-specific characteristics.

We report the performance in each domain and overall test set in Table 7, from which we can observe that: (1) The difficulty of event detection varies across domains. Text conversations is the easiest, and phone conversations is the hardest. (2) PD outperforms SD, which is consistent with the observations in Chen and Cardie (2018). The information sharing between domains may improve the generalization ability of ED models. (3) PDMT gains slight improvement over PD by utilizing the domain information via a simple multi-task way, demonstrating that domain information can bring effective clues. (4) Further, the MDSP strategy generally outperforms all models (e.g., achieving 76.9 F1). The shared-private framework can effectively capture common language features shared across domains, as well as domain-specific patterns. The above analysis show that domain information is effective enhancement, and multi-domain learning deserves more research efforts.

5.5.2 Analysis of Domain Adaptation

Domain adaptation is another key criteria for evaluating ED models. Following Naik and Rose (2020), we investigate the typical unsupervised domain adaptation (UDA) problem, and adopt the following strategies: (1) **BERT-Naive** which utilizes the labeled source domain dataset and ignores the target domain data. (2) **BERT-ADA** which incorporates the adversarial domain adaptation (ADA) framework to construct representations predictive for ED, but not predictive of the domain.

As Table 8 shows, we select source and target domain from the three domains in turn, forming six UDA settings. Though the ADA framework is reported of advantage (Naik and Rose, 2020), it is not the case with MUSIED. BERT-ADA underperforms BERT-Naive in several settings (e.g., U→T, T→P and P→T), which indicates that domain adaptation in MUSIED is challenging due to the multi-source heterogeneity, and more research efforts are required. Other DA settings (e.g., semi-supervised DA) can also be effectively supported by MUSIED and should be further investigated.

6 Related Work

Most existing works towards event detection adopt general domain datasets such as ACE 2005 (Walker et al., 2006), TAC KBP datasets (Mitamura et al., 2015) and MAVEN (Wang et al., 2020b) as benchmarks. Also, some works present domain-specific datasets and valuable explorations. For example, event extraction from biomedical texts are extensively researched (Pyysalo et al., 2007; Thompson et al., 2009; Buyko et al., 2010; Nédellec et al., 2013). Sims et al. (2019) present a new dataset of literary events. CASIE (Satyapanich et al., 2020)

Model	User Reviews			Text Conversations			Phone Conversations			ALL		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SD	**70.4**	72.9	71.6	**81.1**	79.4	80.3	65.9	66.5	66.2	72.9	72.9	72.9
PD	67.6	74.7	71.0	78.0	85.3	81.0	68.4	73.7	71.0	72.6	78.9	75.6
PDMT	67.5	76.6	71.8	76.5	85.1	80.6	69.0	**77.2**	**72.9**	72.1	80.5	76.1
MDSP	**70.4**	**77.9**	**74.0**	77.3	**87.1**	81.9	**70.3**	74.2	72.2	**73.6**	**80.6**	**76.9**

Table 7: Performance of BERT with different multi-domain learning strategies (%).

Setting	Model	In-Domain			Out-Of-Domain		
		P	R	F1	P	R	F1
U→T	BERT-Naive	70.4	72.9	71.6	65.1	64.0	64.6
	BERT-ADA	74.6	74.9	74.7	63.0	64.6	63.8
U→P	BERT-Naive	70.4	72.9	71.6	59.8	60.9	60.3
	BERT-ADA	77.4	74.7	76.0	62.1	62.2	62.2
T→U	BERT-Naive	81.1	79.4	80.3	68.7	60.0	64.1
	BERT-ADA	79.1	80.8	79.9	70.4	58.8	64.1
T→P	BERT-Naive	81.1	79.4	80.3	70.4	61.7	65.7
	BERT-ADA	81.9	51.9	63.5	70.2	46.4	55.9
P→U	BERT-Naive	65.9	66.5	66.2	43.7	44.6	44.1
	BERT-ADA	60.8	62.7	61.7	52.9	47.3	49.9
P→T	BERT-Naive	65.9	66.5	66.2	65.7	65.5	65.6
	BERT-ADA	60.2	62.9	61.5	64.5	65.0	64.8

Table 8: Performance of unsupervised domain adaptation on trigger classification (%). A→B denotes that A and B are source and target domain. U, T and P denotes user review, text conversations and phone conversations respectively. The performances on both source (i.e., the In-Domain column) and target domain test set (i.e., the Out-Of-Domain column) are reported.

and CySecED (Trong et al., 2020) are proposed to facilitate the research of detecting cybersecurity events. Continuous works (Yang et al., 2018; Zheng et al., 2019; Liang et al., 2021) have focused on detecting financial events from the Chinese financial announcements (i.e., ChFinAnn dataset). Lai et al. (2021) presents BRAD, focusing on Black Rebellions events in African Diaspora.

However, most existing works focus on detecting events from formal texts (e.g., news articles, Wikipedia documents, etc), and target the datasets where the texts are either from a single source (e.g., MAVEN (Wang et al., 2020b), CySecED (Trong et al., 2020), ChFinAnn (Yang et al., 2018)) or multiple yet homogeneous sources (e.g., ACE 2005 (Doddington et al., 2004)). In this paper, we present a massive multi-source heterogeneous informal text dataset for event detection, for the first time. It is also the first food safety event detection dataset.

7 Conclusion and Future Work

We have presented MUSIED, a massive multi-source heterogeneous informal text dataset for event detection, based on user reviews, text conversations and phone conversations of online food services. The extensive evaluation verify the unique challenges posed by the textual informality and multi-source heterogeneity characteristics. Our in-depth investigations present multiple promising directions worth pursuing, including exploiting document-level information, multi-domain learning and domain adaptation. In the future, we are interested in extending MUSIED to more event-related tasks such as event argument extraction.

Acknowledgement

We thank the annotators for their efforts into data annotation, and for their continuous verification and revision after the submission of the paper. We also thank Yang An for his remarkable assistance with the spelling error corrector experiments (§ 4.3.3 and § 5.4.3) and Yuncheng Hua for helpful discussions. This research was supported by the National Key Research and Development Program of China (No. 2021YFC3340101).

Limitations

MUSIED is composed of Chinese corpus, which might be less friendly to researchers who are unfamiliar with Chinese. However, considering many non-English datasets have been proposed and promoted research in related fields (e.g., Douban Conversation Corpus (Wu et al., 2017) in dialogue system, DuReader (He et al., 2018) in machine reading comprehension, etc.), we believe that the language barrier does not hinder the contribution of MUSIED to the community. Also, we provide a well-documented homepage and easy-to-use toolkits including preprocessing, models and checkpoints, to further reduce the impact of language barrier.

Ethics Impact

In consideration of ethical concerns, we provide the following detailed description:

1. The corpus is sampled from the logs of a real e-commerce platform, and we strictly desensitized and anonymized the private information. Following Chen et al. (2020), we mask the sensitive information including user's phone number, user's name, user's address, restaurant's name, restaurant's address, etc (e.g. replacing phone number with special token <PHONE-NUMBER>, and replacing order IDs with <ORDER-ID>). The dataset **does not contain** any personally identifiable information, sensitive personal data, or commercially sensitive data.

2. The dataset has been collected in a manner which is consistent with the terms of use. The data officer of the e-commerce platform has authorized us to collect and open source the dataset. The dataset is freely accessible online without copyright constraint to academic use.

3. We hired 20 annotators with food safety domain knowledge and paid them with a fair salary (i.e.,35 dollars per hour) during the annotation. The annotators are treated fairly and able to give informed consent.

Broader Impact

For the first time, we expand event detection to the scenarios involving informal and heterogeneous texts, by carefully curating a new large-scale dataset. In this paper, our extensive experiments with state-of-the-art methods verify the unique challenges posed by textual informality and multi-source heterogeneity characteristics, and indicate multiple promising directions worth pursuing. We believe our work can inspire broader investigation in the future.

References

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.

Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2010. The genereg corpus for gene expression regulation events-an overview of the corpus and its in-domain and out-of-domain interoperability. In *LREC*.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The jddc corpus: A large-scale multi-turn chinese dialogue dataset for e-commerce customer service. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 459–466.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. 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 1226–1240, New Orleans, Louisiana. 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.

Yubo Chen, Hang Yang, Kang Liu, Jun Zhao, and Yantao Jia. 2018. Collective event detection via a hierarchical and bias tagging networks with gated multi-level attention mechanisms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1267–1276, Brussels, Belgium. Association for Computational Linguistics.

Pengxiang Cheng and Katrin Erk. 2018. Implicit argument prediction with event knowledge. 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 831–840, New Orleans, Louisiana. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

Jaspreet Kaur, Amitoj Singh, and Virender Kadyan. 2021. Automatic speech recognition system for tonal languages: state-of-the-art survey. *Archives of Computational Methods in Engineering*, 28(3):1039–1068.

Alexander Kuhnle, Miguel Aroca-Ouellette, Anindya Basu, Murat Sensoy, John Reid, and Dell Zhang. 2021. Reinforcement learning for information retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2669–2672.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Viet Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400.

Jing Li, Dafei Yin, Haozhao Wang, and Yonggang Wang. 2021. Dcspell: A detector-corrector framework for chinese spelling error correction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1870–1874.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82.

Xin Liang, Dawei Cheng, Fangzhou Yang, Yifeng Luo, Weining Qian, and Aoying Zhou. 2021. F-hmtc: detecting financial events for investment decisions based on neural hierarchical multi-label text classification. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4490–4496.

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.

Dongfang Lou, Zhilin Liao, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. Mlbinet: A cross-sentence collective event detection network. 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 4829–4839. Association for Computational Linguistics.

Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. 2015. Overview of tac kbp 2015 event nugget track. In *TAC*.

Aakanksha Naik and Carolyn Rose. 2020. Towards open domain event trigger identification using adversarial domain adaptation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7618–7624.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jungjae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, Sofia, Bulgaria. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):1–24.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. 2020. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8749–8757.

Liao Shasha, Grishman, and Ralph. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the ACL*, pages 789–797. Association for Computational Linguistics.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Paul Thompson, Syed A Iqbal, John McNaught, and Sophia Ananiadou. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC bioinformatics*, 10(1):1–19.

Hieu Man Duc Trong, Duc Trong Le, Amir Pouran Ben Veyseh, Thuat Nguyen, and Thien Huu Nguyen. 2020. Introducing a new dataset for event detection in cybersecurity texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5381–5390.

SS Vallender. 1974. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

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

Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019a. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

Jing Wang, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2020a. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8476–8488, Online. Association for Computational Linguistics.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019b. Adversarial training for weakly supervised event detection. 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 998–1008, Minneapolis, Minnesota. 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. 2020b. 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.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Xiangyu Xi, Wei Ye, Tong Zhang, Quanxiu Wang, Shikun Zhang, Huixing Jiang, and Wei Wu. 2021. Improving event detection by exploiting label hierarchy. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7688–7692. IEEE.

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. IEEE.

Xi Xiangyu, Wei Ye, Shikun Zhang, Quanxiu Wang, Huixing Jiang, and Wei Wu. 2021. Capturing event argument interaction via a bi-directional entity-level recurrent decoder. 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 210–219, Online. Association for Computational Linguistics.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55.

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.

Ying Zeng, Honghui Yang, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2016. A convolution bilstm neural network model for chinese event extraction. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, volume 10102 of *Lecture Notes in Computer Science*, pages 275–287. Springer.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2edag: An end-to-end document-level framework for chinese financial event extraction. *arXiv preprint arXiv:1904.07535*.

A Event Type Schema

We present the event type schema along with their descriptions in Table 18. The schema contains 21 event types and broadly covers the user's feedback about food quality, restaurant, delivery person, and user's physical feelings.

B Data Analysis

B.1 Top 5 Event Types

We show the top 5 event types along with their instance number and proportion in Table 9. The top 5 event types for both overall corpus (i.e., the `ALL` row) and each domain (i.e., the `User Review`, `Text Conversation` and `Phone Conversation` rows) are shown.

Event Type	#Event Mentions	Proportion
`ALL`		
Impurities	13,883	39.3%
Uncomfortable	7,935	22.5%
Abnormalities	6,684	18.9%
Low-quality	1,950	5.5%
Spoiled	1,176	3.3%
`User Review`		
Abnormalities	1,865	38.1%
Impurities	783	16.0%
Undercooked	651	13.3%
Uncomfortable	541	11.0%
Cold	414	8.4%
`Text Conversation`		
Impurities	6,696	42.2%
Uncomfortable	3,579	22.6%
Abnormalities	2,510	15.8%
Low-quality	1,757	11.1%
Spoiled	476	3.0%
`Phone Conversation`		
Impurities	6,404	43.9%
Uncomfortable	3,815	26.2%
Abnormalities	2,309	15.9%
Spoiled	649	4.4%
Expired	272	1.9%

Table 9: Statistics of top 5 event types.

B.2 Domain Statistics

For each domain of MUSIED, we present the detailed statistics in Table 10.

B.3 Statistics of Event Type Distribution

Given a corpus with N domains and M event types, we first calculate the event type distribution P_i for each domain i as follows:

$$P_i = (p_{i,1}, p_{i,2}, ..., p_{i,M})$$
$$p_{i,j} = \frac{\#(\text{triggers with type } j \text{ in domain } i)}{\#(\text{triggers with any type in domain } i)} \quad (1)$$

where $p_{i,j}$ denotes the occurrence frequency of type j in domain i. $\#$(triggers with type j in domain i) denotes the number of triggers with type j in domain i. $\#$(triggers with any type in domain i) denotes the number of triggers in domain i.

Then, we calculate the wasserstein distance between the event type distributions of any two domains (exemplify with P_1, P_2) as follows:

$$W(P_1, P_2) = inf_{\gamma \sim \Pi(P_1, P_2)} E_{(x,y)\sim y}[|x-y|] \quad (2)$$

Further, we calculate the average wasserstein distance as follows:

$$\bar{W} = \frac{1}{M}\frac{1}{C_N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N} W(P_i, P_j) \quad (3)$$

where $C_N^2 = \frac{N(N-1)}{2}$ denotes the number of domain pairs and M denotes the number of event types.

B.4 Word-Trigger Mismatch

As a Chinese dataset, MUSIED lacks natural delimiters and also suffers from the word-trigger mismatch problem existing in ACE 2005 Chinese dataset (Zeng et al., 2016; Lin et al., 2018; Xiangyu et al., 2019). The words generated by word segmentation toolkits might not exactly match with event triggers. Following Xiangyu et al. (2019), we make statistics on two types of word-trigger mismatch: i) Cross-word Triggers where a trigger might be composed of multiple words; ii) Inside-word Triggers where a single character or some consequent characters inside a word can be a trigger. The statistical results with three different word segmentation tools are shown in Table 11, from which we can observe that proportion of problematic triggers in MUSIED is much larger than ACE 2005 Chinese dataset (i.e., 35.24% v.s. 16.15%). The severe word-trigger mismatch problem poses a great challenge and may hinders the performance of word-wise event detection models.

Domain	#Document	#Tokens	#Sentences	#Events	#Event Mentions
User Review	4,226	144k	6,083	4,036	4,898
Text Conversation	3,767	3,136k	181,316	14,686	15,858
Phone Conversation	3,388	3,805k	128,074	12,218	14,557
Total	11,381	7,105k	315,473	30,940	35,313

Table 10: Domain statistics of MUSIED.

Toolkits	ACE 2005 Chinese			MUSIED		
	C-W	I-W	R	C-W	I-W	R
CoreNLP[†]	2.25%	11.79%	85.96%	26.21%	11.38%	62.41%
Jieba[‡]	2.31%	17.94%	79.75%	23.06%	9.35%	67.59%
NLPIR[§]	8.97%	5.19%	85.84%	29.58%	5.98%	64.44%
Average	4.51%	11.64%	83.85%	26.28%	8.96%	64.76%

Table 11: Statistics of word-trigger mismatch. C-W, I-W and R denotes cross-word triggers, inside-word triggers and regular triggers respectively.

C Samples of Annotated Data

To promote understanding, we show the sample of annotated data from three domains, as Table 12 shows.

Sample # 1 Domain: User Review
The dishes have hair, the restaurant does not reply to us, I often order dishes in this restaurant. 菜里/有/毛发，跟/商家/沟通/也/不/回复，我/还/经常/点/他们/家/外卖。
Event Trigger: 毛发(hair) Event Type: Impurities
Sample # 2 Domain: Text Conversation
Last time I shopped, the noodles expired 上次/买/东西/面/就/过期/了
Event Trigger: 过期(expired) Event Type: Expired
Sample # 3 Domain: Phone Conversation
So why does that fried chicken have black splots? 所以/那份/炸鸡/为什么/会有/黑斑?
Event Trigger: 黑斑(black splots) Event Type: Spoiled

Table 12: Sample of annotated data from three domains.

[†]https://nlp.stanford.edu/software/segmenter.shtml
[‡]https://github.com/fxsjy/jieba
[§]https://github.com/NLPIR-team/NLPIR

D Hyperparameters

In this section, we introduce the hyperparameter settings and training details of various ED models that we implemented for experiments.

D.1 BERT-based Models

For both **DMBERT** and **BERT**, we use the $\text{BERT}_{\text{BASE}}$ for Chinese, and the released pre-trained checkpoints can be downloaded at https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip. Adam with learning rate of 2e-05, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used for optimization. We set the training epochs and batch size to 50 and 64 respectively. We set dropout to 0.1.

D.2 LSTM-based Models

For **BiLSTM**, **BiLSTM+CRF** and **C-BiLSTM**, we use the pre-trained Chinese word embeddings [¶]. The adopted hyperparameters are shown in Table 13.

D.3 DMCNN model

For DMCNN, we use the pretrained Chinese word embeddings, and the hyperparameters are shown in Table 14.

[¶]https://github.com/Embedding/Chinese-Word-Vectors

Epoches	50
Batch Size	64
Dropout Rate	0.1
Learning Rate	2e-05
Dimension of Word Embedding	300
Dimension of Hidden Layers	300
Layers of LSTM	1
Kernel Size of CNN	3
Number of Feature Map	300
Optimizer	Adam

Table 13: Hyperparameter settings for the BiLSTM-based models.

Epoches	50
Batch Size	64
Dropout Rate	0.1
Learning Rate	2e-05
Dimension of Word Embedding	300
Kernel Size of CNN	3
Number of Feature Map	300
Optimizer	Adam

Table 14: Hyperparameter settings for the DMCNN model.

D.4 HBTNGMA model

For HBTNGMA, We use the official code released by Chen et al. (2018) [‖]. We adopt the original hyperparameters from Chen et al. (2018) except that we use the pretrained Chinese word embeddings with 300 dimension.

D.5 MLBiNet model

For MLBiNet, we use the official code released by Lou et al. (2021) [**]. We use the pretrained Chinese word embeddings with 300 dimension.

E Computing Issues

The computing issues are explained in this section.

Computing Infrastructure We implemented our model with TensorFlow v1.4.0 and Pytorch v1.7.0, and trained our models on NVIDIA Tesla v100 GPU. The operation system is CentOS 7.6.

Computational Budget Table 15 shows the used computing infrastructures and the average running time per epoch of various models.

[‖] https://github.com/yubochen/NBTNGMA4ED

[**] https://github.com/zjunlp/DocED

Model	Computing Infrastructure	Runtime
DMCNN	1× Tesla v100	40 min
BiLSTM	1× Tesla v100	5 min
BiLSTM+CRF	1× Tesla v100	8 min
C-BiLSTM	1× Tesla v100	10 min
DMBERT	1× Tesla v100	55 min
BERT	1× Tesla v100	30 min
HBTNGMA	1× Tesla v100	20 min
MLBiNet	1× Tesla v100	20 min

Table 15: The average runtimes per epoch of various models.

F Experimental Results

F.1 Overall Performance

Following previous works (Li et al., 2013; Chen et al., 2015), we only report *Precision* (P), *Recall* (R) and *F1-Score* (F1) on trigger classification task in § 5.3. The performance on trigger identification task is also shown in Table 16.

Model	P	R	F1
DMCNN	**85.5**	57.6	68.8
BiLSTM	77.1	67.7	72.1
BiLSTM+CRF	77.5	70.1	73.6
C-BiLSTM	77.2	71.9	74.5
DMBERT	79.4	70.8	74.9
BERT	73.8	**80.3**	76.9
HBTNGMA	74.3	**80.3**	**77.2**
MLBiNet	75.5	72.3	73.9

Table 16: Overall performance on trigger identification.

F.2 Case Study of Spelling Error Corrector

A concrete case S7 is shown to demonstrate the benefit of Spelling Error Corrector (SEC). The user intends to express that he/she feels unwell ("感觉/不适(*feel unwell*)"). However, the user types a typo token "是" (means yes), which has the same pronunciation (pronounced as "shi" in Chinese) but different meaning as the token "适" (means physically well). The word "不是" is a widely-used statement of expressing negation. All models fail to recognize the instance due to the typo before correction, and can fix the error with correction.

S7: 豆腐面条鸡蛋然后吃了之后身体就感觉不是 (适) (*Tofu, noodles and eggs. After eating them, I feel not (unwell)*)

F.3 Impact of Data Imbalance

As § 4.2 shows, the inherent data imbalance problem exists in MUSIED. To quantitatively investigate the effect, we first rank labels (i.e., event types) based on the number of their corresponding training instances and then divide them into several subsets with continuous rankings. Since instances with a specific label may be too few, empirical results on instances of a label set could yield more robust and convincing conclusions. The first event type alone forms a single subset, and the remaining 20 event types are equally grouped into three subsets. In this way, we finally get a division of four subsets, named *Subset-1*, *Subset-2*, *Subset-3* and *Subset-4*, which contain 1, 6, 7 and 7 labels respectively.

Model	*Subset-1*	*Subset-2*	*Subset-3*	*Subset-4*
DMCNN	82.39%	55.67%	41.21%	13.32%
BiLSTM	87.45%	61.46%	48.24%	10.51%
DMBERT	82.96%	61.42%	42.71%	23.52%
BERT	88.96%	64.38%	59.31%	49.99%

Table 17: F1-score of different models in four subsets.

As Table 17 shows, we collect the F1-scores of four baselines for each subset, from which we can find that the data imbalance problem significant hinders the performance and results in a degradation (e.g., 88.96 of *Subset-1* v.s. 64.38 of *Subset-2* v.s. 59.31 of *Subset-3* v.s. 49.99 of *Subset-4* for BERT). The performance is significantly worse when label has fewer training instances. Hence, further explorations on handling the data imbalance challenge may be critical for MUSIED.

ID	Event Type	Description
	Restaurant	The illegal or improper behaviors of restaurants lead to food safety problems.
1	*Additives*	Restaurant uses illegal food additives, including food additives with irregular labels and unknown sources.
2	*Contraband*	Restaurant sells commodities that are prohibited or contains non-food raw materials
3	*Harmful-residues*	Restaurant sells food that contains harmful residues, such as pesticide, biological toxins, and heavy metals.
4	*Poor-environment*	Restaurant provides unsanitary dinning environments.
5	*Recycled-material*	Restaurant sells food that is produced using recycled food as raw material.
6	*Inconsistent-product*	Restaurant sells food that is inconsistent with the advertisement, such as food quantity, dish content, etc.
7	*Fake*	Restaurant sells fake food with counterfeit, shoddy, or unauthorized materials.
8	*Low-quality*	Restaurant is reported by users to have unspecified food quality problems.
9	*Non-compliant*	Restaurant provides service in non-compliant status, including 1) without a license, 2) with fake licenses, and 3) the scope of licenses does not match the actual scope of business.
10	*Poor-packaging*	Restaurant provides poor food packaging (or dinnerware), such as simple, thin and smelly packaging with non-food-graded materials.
11	*Unreliable-product*	Restaurant sells food that contains products without a production date, quality certificate, or manufacturer's source.
	Delivery Person	The illegal or improper behaviors of delivery person lead to food safety problems.
12	*Damaged*	Delivery person damages or pollutes the food packaging, which affects the quality of food or ingredients.
13	*Steal*	Delivery person is suspected to steal (part of) food based on the quantity and packaging integrity.
	Food Quality	The poor food quality lead to food safety problems.
14	*Spoiled*	Food or ingredients have obviously deteriorated, moldy, or rotten, both internally and externally.
15	*Undercooked*	Food or ingredients are undercooked.
16	*Cold*	Food or ingredients have low temperatures, which affects the taste.
17	*Expired*	Food or ingredients are expired.
18	*Thaw*	Food or ingredients are melting due to improper cold chain distribution.
19	*Impurities*	Food or ingredients contains undesirable and disgusting objects, such as eggshells, hair, etc.
	Physical Feelings	The users' physical feelings suggest the existence of food safety problems.
20	*Uncomfortable*	User feels unwell after the meal, in terms of physical feelings.
21	*Abnormalities*	User feels unwell with the meal, in terms of visual or gustatory feelings.

Table 18: The 21 event types in MUSIED and their corresponding descriptions.