ECO v1: Towards Event-Centric Opinion Mining

Ruoxi Xu[1,3], **Hongyu Lin**[1]*, **Meng Liao**[4]*, **Xianpei Han**[1,2],
Jin Xu[4], **Wei Tan**[4], **Yingfei Sun**[3], **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, Beijing, China
[4]Data Quality Team, WeChat, Tencent Inc., China
{ruoxi2021,hongyu,xianpei,sunle}@iscas.ac.cn
{maricoliao,jinxxu,vinnietan}@tencent.com
yfsun@ucas.ac.cn

Abstract

Events are considered as the fundamental building blocks of the world. Mining event-centric opinions can benefit decision making, people communication, and social good. Unfortunately, there is little literature addressing event-centric opinion mining, although which significantly diverges from the well-studied entity-centric opinion mining in connotation, structure, and expression. In this paper, we propose and formulate the task of event-centric opinion mining based on event-argument structure and expression categorizing theory. We also benchmark this task by constructing a pioneer corpus and designing a two-step benchmark framework. Experiment results show that event-centric opinion mining is feasible and challenging, and the proposed task, dataset, and baselines are beneficial for future studies.

1 Introduction

Events are the fundamental building blocks of the world (Russell, 1927; Ong, 1969). We express, share and propagate our opinions about events with personal understandings, emotions and attitudes in our daily life. People can better understand, communicate and interact with each other by mining, sharing and exchanging event-centric opinions. And being exposed to event-centric opinions from different angles can debias people's own emotions and attitudes about social issues (Karamibekr and Ghorbani, 2013). Therefore, mining event-centric opinions have huge social and personal impacts.

Unfortunately, there is little literature addressing *event-centric opinion mining*, and most of current opinion mining studies focus on entity-centric opinions, which significantly diverge from the concerning event-centric ones. First, entity-centric opinions mostly focus on sentimental polarity of the holder (Liu, 2012), meanwhile event-centric opinions care more about the content such as non-sentimental judgments, predictions or suggestions. Second, due to the rich interactions between events, entities, and people, event-centric opinions have a complicated structure. Given an event, people can express their opinions about the event itself, as well as its subevents, related events, and the involved entities. Third, the expressions of event-centric opinions are unique. The targets of event-centric opinions are frequently implicitly referred to, which often don't appear in the opinion expressions. Moreover, event-centric opinions are usually widely spread in long news and passages, which are mixed up with facts and non-opinion information. By contrast, entity-centric opinions mainly appear in short and focused reviews or comments individually. The above connotation, structure, and

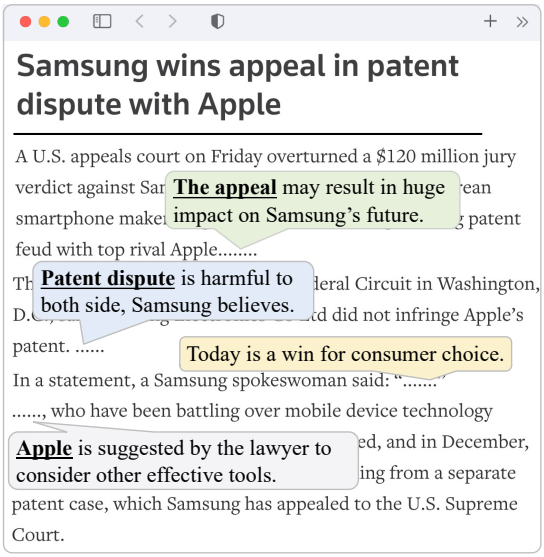

Figure 1: An illustration of event-centric opinions. Given an event, people can express their judgements, beliefs, attitudes and suggestions. The opinions oriented to an event may not directly target at itself, but can target at its related subevents or entities.

*Corresponding Authors

Findings of the Association for Computational Linguistics: ACL 2022, pages 2743 - 2753

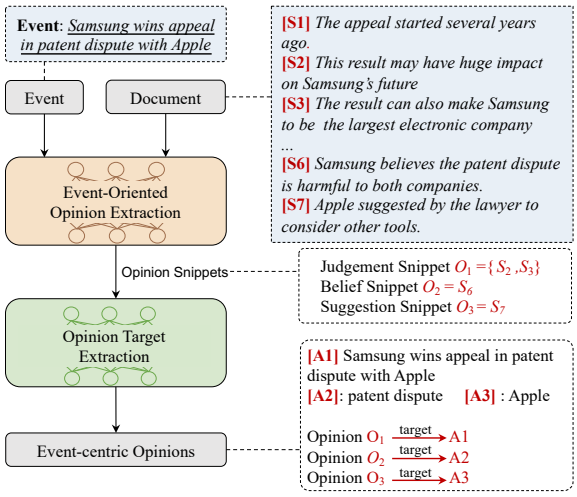

Figure 2: Overall architecture of our framework. Here S represents sentence in the document.

expression divergences make event-centric opinion mining a novel task, which cannot be resolved using current entity-centric mining techniques.

In this paper, we formally formulate the task of event-centric opinion mining. Specifically, inspired by the expression categorizing theory (Asher et al., 2009), we define 5 types of event-centric opinions, and a text snippet is considered as an event-centric opinion if it contains judgments, attitudes, beliefs, sentiments or suggestions of the opinion holder. Then we formulate the targets of event-centric opinions as an *event-arguments opinion structure* by extending the widely-used event-arguments structure (Pustejovsky, 1991). In this way, an opinion can target an event itself, or its specific arguments including subevents or involved entities. For example in Figure 1, given an event "Samsung wins appeal in patent dispute with Apple", an opinion towards this event may target at the `wins appeal` event itself, the related subevent `patent dispute`, as well as the involved entity `Samsung` and `Apple`. Consequently, event-centric opinion mining can be formulated as identifying opinion snippets from event-related documents and then recognizing the target argument of the opinion snippet.

Based on the task formulation, we create Event-Centric Opinion Bank (ECO Bank), a pioneer corpus for learning and evaluating event-centric opinion mining models. ECO Bank contains nearly 1K events from real-world event trending services, as well as 5K documents about these events in English and Chinese. Each document is aligned to one event. Given a document and its related event, we manually annotated the opinion segments corresponding to the event in the document, and align them to correct target arguments of the event. Consequently, we obtain nearly 18K opinion segments from 5K documents, which target more than 4K different arguments of 1K events.

Finally, we propose a new framework to tackle event-centric opinion mining and benchmark the task on ECO Bank. The overall architecture of the framework is shown in Figure 2. Specifically, we decouple event-centric opinion mining into a two-step pipeline. Step 1 is *event-oriented opinion extraction (EOE)*, which detects the snippets containing event-oriented opinions in each document given the concerning event. Step 2 is *opinion target extraction (OTE)*, which recognizes the corresponding target arguments in the event given identified opinion snippets. We then provide two baselines for each step. For event-oriented opinion extraction, we formulate it as either a sentence-level sequential labeling task or a binary sentence classification task. For opinion target extraction, we resolve it based on a span ranking model or an MRC model. By comparing and analyzing the performance of different baselines, we figure out the critical challenges and bottlenecks of current methods to event-centric opinion mining, which can shed some light on the future research directions in this field.

Generally, the contributions[1] of this paper are:

- We propose, define and formulate the task of event-centric opinion mining based on event-argument structure and expression categorizing theory. To the best of our knowledge, this is the first work that tries to formally formulate event-centric opinions and the task of event-centric opinion mining.
- We construct Event-Centric Opinion Bank (ECO Bank), a pioneer corpus for learning and benchmarking event-centric opinion mining models in both English and Chinese. To the best of our knowledge, this is the first public benchmark focusing on event-centric opinions.
- We design a two-step framework to tackle event-centric opinion mining, and propose several baseline approaches to identify and analyze the challenges and bottlenecks of the task.

2 Event-Centric Opinion Mining

This section first defines the connotation and targets of event-centric opinions. Then we will formulate

[1]ECO Bank and the source code are available at e-com.ac.cn.

the task of event-centric opinion mining.

2.1 Connotation of Event-Centric Opinions

The connotations of event-centric opinions are complicated and cannot be simply summarized based on sentimental tendencies. For example, towards a *Trade War* event, one may express the personal judgment opinion by commenting "*I do believe this is a turning point of the relationships between two countries*", which is without explicit sentiments. Therefore, we need to define broader connotations for event-centric opinions than entity-centric ones. To this end, we formulate the connotations of event-centric opinions according to the divergence between facts and opinions (Banfield, 1984; Hackett, 1984). An event-centric opinion is defined as a statement that expresses views about an event or related issues, which 1) cannot be proved or disproved with currently available information and 2) varies from person to person (Schauer, 1978; Wiebe et al., 2005; Corvino, 2014).

Specifically, inspired by expression categorizing theory (Asher et al., 2009), we define the connotation of event-centric opinion as a text snippet that expresses the following 5 kinds of information, including: 1) **Judgements**, such as speculations, interpretations, and predictions about things in the future, e.g., *Trump's plan will not work*; 2) **Attitudes**, such as positions on controversial issues and evaluations of people, places, and things, e.g., *Apple describes the ruling as total political crap*; 3) **Sentiments**, which express feelings like fear and sadness, e.g., *I am so happy to see the Act passed*; 4) **Beliefs**, which can not be proved or disproved, e.g. *I believe aliens definitely exist*; 5) **Suggestions**, which is about personal advice to the readers, e.g., *We advise Samsung Galaxy Note 7 owners to turn off their devices during flights*.

2.2 Targets of Event-Centric Opinions

Entity-centric opinions mostly directly target the entity or its attributes (which is referred as *aspects*). By contrast, when talking about events, people can talk about their related events, entities and concepts, and these opinions may not direct to the event itself. For example, given the *Trade War* event, an opinion holder may express their opinion like *Trump always made bad decisions*, which actually targets on *Trump* rather than the event because the holder will not change the opinion no matter whether the *Trade War* event happens.

To formulate the target of event-centric opinions, we introduce event-arguments opinion structure. Event-arguments structure (Pustejovsky, 1991) is a widely used event formulation in many event-related tasks where arguments refer to a set of critical elements about how an event realized (Doddington et al., 2004; Hovy et al., 2013). Based on this structure, the opinions about a specific event target at one of the following arguments: 1) **Event**, which means that the opinion is directly targeting the entire event; 2) **Subevents**, which means that the opinion does not target at the entire event, but on its subevent or related event. For example in Figure 1, an opinion towards the *patent dispute* in the event *Samsung wins appeal in patent dispute with Apple*; 3) **Entities**, which means that the opinion directly targets one of the involved entities regardless of the event, e.g., commenting *Apple is a great company* on the event in Figure 1.

2.3 Task Formulation for Event-Centric Opinion Mining

Based on the formulated connotation and targets, we define event-centric opinion mining as the task of extracting (opinion, argument) pairs from a document and an event descriptor. Formally, let $e = \{w_1, w_2, ..., w_m\}$ denote an event descriptor with m tokens and $d = \{s_1, s_2, ..., s_n\}$ denote a document with n sentences. Event-centric opinion mining aims to identify (opinion, argument) pairs $T = \{..., (o_k, a_k), ...|e, d\}$, where $o = \{s_i, s_{i+1}, ..., s_j|s \in d\}$ is a continuous opinion segment in d targeting at the same argument, and $a = \{w_t, w_{t+1}, ..., w_l|w \in e\}$ is the target argument of the opinion o in the event descriptor e. For example in Figure 1, given the event *Samsung wins appeal in patent dispute with Apple*, there are 4 (opinion,argument) pairs, two of whom target at the entire event, one target at subevent *patent dispute* and one target at entity *Apple*.

3 Event-Centric Opinion Bank

Based on the task formulation, this paper creates Event-Centric Opinion Bank (ECO Bank), a new event-centric opinion mining corpus in both English and Chinese. In the following, we describe how we construct ECO Bank and report the statistics of ECO Bank.

3.1 Dataset Construction

Event Descriptor Collection. To construct ECO Bank, we collect event descriptors from real-world event trending services. For the English portion of ECO Bank, we collect event descriptors from the W2E dataset (Hoang et al., 2018), a worldwide event dataset for topic detection and tracking. We select highly discussed topics as event descriptors and manually shorten them into meaningful texts if necessary. For the Chinese portion, we collect manually-maintained event trending from widely-used social networks, WeChat Top Topics. We then filter out items in the trending corresponding to events as event descriptors. Because the trending is manually created, it is already of sufficient quality and therefore no more modification is required. Finally, we construct 988 high-quality event descriptors, where 821 in the Chinese and 167 in the English.

Document Collection. Given the event descriptor, we collect related documents that may contain the opinions towards the event. For the Chinese portion of ECO Bank, we collect related documents by retrieving relevant documents from WeChat Search. Specifically, We retrieve top 10 articles for each event descriptor, and then manually filter out the redundant, low-quality and irrelevant documents. For the English portion, because each topic in W2E dataset is already linked to several related articles from more than 50 prominent mass media channels. We therefore directly applied these documents except filtering out the redundant ones. Finally, we preserve 3000 Chinese documents and 2000 English documents for further annotation.

Event-Centric Opinion Annotation. Given the documents and their corresponding event descriptors, we hired annotators to annotate the (opinion, argument) pairs. Specifically, the annotation is conducted in a two-step paradigm. First, annotators are asked to identify opinions related to the event described in the descriptor. Then, given the event descriptor, an identified opinion and the source document, the annotators were asked to link the opinion to its target in the descriptor. To ensure the high quality of annotations, each document is annotated by two annotators. If there is a disagreement between the original annotators, three more professional annotators will relabel the document independently, and produce the final annotations by voting between them. Finally, to facilitate further research, we also ask annotators to recognize all possible arguments in the event descriptor that can serve as an opinion target. All annotators are fairly paid according to their workload.

Statistics on ECO Bank		Chinese	English
Document	Number	3000	2000
	Avg. Sents	15.2	20.3
	Avg. Opinion	2.6	4.5
Opinion	Number	7742	9058
	Ratio (%)	32.0	28.1
	Avg. Sents	1.9	1.3
Event	Number	821	167
	Avg. Tokens	6.4	7.7
Arguments	Events (%)	30.6	34.4
	Subevents (%)	24.9	11.7
	Entities (%)	44.5	53.9

Table 1: Overall statistics of ECO Bank dataset.

3.2 Dataset Analysis

Table 1 shows the main statistics of ECO Bank. We can observe many unique characteristics of event-centric opinion mining. First, the distribution of event-centric opinions is very sparse. Only about 30% of the sentences in both English and Chinese dataset express opinions. This is because event-centric articles usually mix massive factual snippets with opinionated snippets. By contrast, entity-centric opinions are densely distributed in comments and reviews. Second, the targets of event-centric opinions are highly diversified. We notice that only 30% of opinions directly target the event, leaving 24.9% on subevents and 44.5% on entities (ECO-ZH), and 11.7% on subevents and 53.9% on entities (ECO-EN). This verifies the necessity of defining an event-specified opinion structure. Furthermore, we find that targets of event-centric opinions are often implicit. To show this we randomly select 50 documents with 151 opinions. Among them, there are 80 opinions on events/subevents, where 25% opinions target implicit arguments, and 28% opinions are with event co-reference and therefore its target event cannot be directly recognized without more contexts. By contrast, this proportion is much lower in opinions on entity arguments, where we only find 8% implicit arguments and 7% entity co-reference. These results demonstrate that the target of event-centric opinions cannot be identified locally, which is one of the most significant divergences between event-centric and entity-centric opinion mining.

4 Benchmarking Event-Centric Opinion Mining

This section benchmarks event-centric opinion mining with a two-step framework. Two feasible solutions are proposed for each step, and therefore lead to 4 different benchmark architectures.

4.1 Step 1: Event-Oriented Opinion Extraction

Given an event descriptor e and a related document d, the goal of event-oriented opinion extraction (EOE) is to extract text snippets o in d which contain opinions about event e. To this end, we propose two architectures, one formulates EOE as a pair-wise classification task and the other formulates it as a sentence-level sequential labeling task.

4.1.1 Pair-wise Classification

A basic solution for EOE is to build binary classifier for all (sentence, event) pairs. Specifically, given an event e and a sentence s in document d, we identify whether s is an opinion to e using a BERT-based binary classifier (Devlin et al., 2019). The classifier takes the concatenation $\mathcal{X} = \{[\text{CLS}], e, [\text{SEP}], s\}$ as input, where [CLS] and [SEP] represent the beginning of input and the separator between s and e respectively. We then use BERT as the encoder, then conduct binary classification on [CLS] token to identify the relation between e and s:

$$\mathcal{H} = \text{BERT}(\mathcal{X}), \quad p = \text{sigmoid}(\mathcal{H}_{[\text{CLS}]}), \quad (1)$$

where $\mathcal{H}_{[\text{CLS}]}$ is the representation at [CLS] token, and p is the probability of s containing an opinion to e. Then we regard sentences with $p \leq 0.5$ as the opinion sentences, and concatenates all continuous opinion sentences to form opinion snippets.

4.1.2 Sentence-level Sequential Labeling

Because an opinion may contain more than one sentence, sentence-level classification to identify opinion snippets may result in opinion boundary ambiguity. To this end, we propose sentence-level sequential labeling architecture (Cheng et al., 2020) for EOE. Specifically, given a document $d = \{s_1, s_2, ..., s_n\}$ and an event descriptor e, we first concatenate e and each sentence in d to form the input $\mathcal{X}$, using [CLS] as the separators between each sentence. We then fed $\mathcal{X}$ into BERT-based encoder to obtain context-aware representations. The representations at [CLS] tokens $\mathcal{H}_{[\text{CLS}]}$ are used to represent the sentences after them. To leverage the deep interaction between different sentences, we further apply BiLSTM layer upon $\mathcal{H}_{[\text{CLS}]}$ and learn the interacted sentence representations $\mathcal{S} = \text{BiLSTM}(\mathcal{H}_{[\text{CLS}]})$. Finally, we apply a Conditional Random Field (Lafferty et al., 2001) upon $\mathcal{S}$ to label each sentence to obtain the sentence-level tagging output $\mathcal{Y} = \text{CRF}(\mathcal{S})$ encoded in BIO schema (Sang and Buchholz, 2000).

4.2 Step 2: Opinion Target Extraction

Given an event descriptor $e = \{w_1, ..., w_n\}$ and an opinion snippet o identified in Step 1, Opinion Target Extraction (OTE) aims to recognize a span in e corresponding to the target argument of o. To this end, we build two baselines of OTE by taking it as either a span ranking problem or a MRC problem.

4.2.1 Span Ranking for Opinion Target Extraction

Given an opinion o, the span ranking approach directly enumerates all spans in e, and selects the best span as o's target argument. Formally, given a span a in e, we concatenate a with o to form the model input. Then similar to the pair-wise EOE classifier in Equation (1), we send the concatenation into a BERT-based encoder, and then obtain the score of the span a being the opinion target of o via a sigmoid classifier. Finally, the span with highest score is regarded as the target of the opinion o.

4.2.2 MRC for Opinion Target Extraction

Recent advances (Cui et al., 2020; Sugawara et al., 2020) have shown that pointer network style machine reading comprehension models (Wang and Jiang, 2016) can effectively resolve the span spotting problems. Therefore, we apply an MRC architecture similar to Devlin et al. (2019) for OTE, which regards the opinion o as the query and the event descriptor $e = \{w_1, w_2, ..., w_n\}$ as the document to identify argument a from e.

Specifically, given an event descriptor e and opinion o, we first represent the input o and e as a single packed sequence $\mathcal{X} = \{[\text{CLS}], o, [\text{SEP}], e\}$. We use BERT encoder to get token representations $\mathcal{H} = \text{BERT}(\mathcal{X})$. We then introduce a start vector $\mathcal{S}$ and an end vector $\mathcal{E}$. The probability of word w_i being the start or end of the argument span is computed as a dot product between $\mathcal{H}_\mathbf{i}$ and $\mathcal{S}$ or $\mathcal{E}$ followed by a softmax over all of words in e:

$$p_s = \frac{e^{\mathcal{S}\mathcal{H}_i}}{\sum_j e^{\mathcal{S}\mathcal{H}_j}}, \quad p_e = \frac{e^{\mathcal{E}\mathcal{H}_i}}{\sum_j e^{\mathcal{E}\mathcal{H}_j}}. \quad (2)$$

	ECO-ZH						ECO-EN					
	Segment level			Sentence level			Segment level			Sentence level		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PairCls-SpanR	14.51	12.08	13.18	37.69	33.61	35.50	6.86	4.83	5.76	13.77	12.64	13.18
PairCls-MRC	13.45	11.19	12.22	48.99	43.67	46.13	14.67	10.42	12.19	33.12	29.32	31.10
Seq-SpanR	25.07	21.77	23.31	35.46	28.07	31.34	9.24	9.96	9.59	11.30	12.54	11.89
Seq-MRC	29.72	26.48	28.01	47.74	37.80	42.19	17.02	19.44	18.15	24.71	27.77	26.15
Human	86.96	86.02	86.49	79.46	94.23	86.22	72.59	82.10	80.83	86.78	86.07	86.42

Table 2: Overall experiment results on ECO-ZH and ECO-EN datasets. *PairCls* denotes pair-wise classification method (§ 4.1.1), *Seq* denote sentence-level sequential labeling method (§ 4.1.2) for EOE, and *SpanR* denotes Span ranker (§ 4.2.1), *MRC* denotes MRC method (§ 4.2.2) for OTE. We also represent human performance as *Human*.

The score of a candidate span from position i to j is defined as $\mathcal{SH}_i + \mathcal{EH}_j$, and the maximum scoring span where $i \leq j$ is used as a prediction.

5 Experiments

5.1 Benchmark Settings

Dataset Split. We split both English and Chinese portion of Event-Centric Opinion Bank into roughly 7:1:2 for train/dev/test respectively. To ensure no information leakage, the same event descriptor will not be sampled into different sets. Finally, for English portion, there are 112/16/39 event descriptors with 1402/198/400 documents for train/dev/test. And for Chinese portion, there are 590/78/153 event descriptors with 2100/299/601 documents for train/dev/test. This ECO Bank split can be viewed as a standard benchmark for evaluating event-centric opinion mining models.

Evaluation Criteria. To evaluate the event-centric opinion mining performance, we design several evaluation metrics for the task as well as its two sub-tasks. Specifically, given golden (opinion,argument) pair set $\mathcal{T} = \{(o_1^T, a_1^T), ..., (o_n^T, a_n^T)\}$ and the predicted (opinion,argument) pair set $\mathcal{P} = \{(o_1^P, a_1^P), ..., (o_n^P, a_n^P)\}$, where $o_i = \{s_{i1}, ..., s_{ik}\}$ contains continuous sentences from documents and $a_i = \{w_{i1}, .., w_{il}\}$ contains continuous words from the event descriptors, we design the following evaluation metrics:

1. End2End Evaluation, which measures the end-to-end performance of event-centric opinion mining. We propose to use F_1 score at opinion segment-level or sentence-level to evaluate the overall performance. Segment-level F_1 is the F_1 score calculated by directly comparing $\mathcal{T}$ and $\mathcal{P}$. And sentence-level F_1 is calculated by first splitting (o,a) pairs in $\mathcal{T}$ and $\mathcal{P}$ into sentence-level pairs $\{(s_1, a), .., (s_k, a)\}$ and then combining them to form the sentence-level golden annotation set $\mathcal{T}'$ and prediction set $\mathcal{P}'$. Finally, sentence-level F_1 is calculated between $\mathcal{T}'$ and $\mathcal{P}'$.

	ECO-ZH		ECO-EN	
	Segment-F1	Sent-F1	Segment-F1	Sent-F1
PairCls	24.39	67.35	25.07	53.40
Seq	44.33	62.53	34.84	48.41

Table 3: The performance on Event-oriented Opinion Extraction.

	ECO-ZH		ECO-EN	
	Accuracy	Overlap-F1	Accuracy	Overlap-F1
SpanR	49.21	77.83	26.50	53.31
MRC	64.89	84.89	54.29	76.98

Table 4: The performance on OTE given golden opinion snippets.

2. EOE Evaluation. We also consider both segment-level and sentence-level metrics when evaluating the Step 1 EOE. The only difference is that we only evaluate the performance of extracting opinion snippets without considering corresponding opinion targets in EOE evaluation.

3. OTE Evaluation. To evaluate how well the Step 2 OTE works, we further use the golden annotated opinion snippets as input to evaluate OTE performance. We use two evaluation metrics for OTE: 1) Accuracy, which measures whether the extracted argument can be exactly the same as the annotated one; 2) Overlap-F_1, which measures the overlap between extracted and golden arguments using F_1. Specifically, let $a^T = \{w_1^T, .., w_l^T\}$ denotes the golden argument and $a^P = \{w_1^P, .., w_k^P\}$ as the predicted argument, Overlap-F_1 is calculated by micro-averaged F_1 on all (a^T, a^P) pairs.

5.2 Overall Results

The performance of 4 different architectures on the end2end, EOE and OTE evaluation are shown in Table 2, 3 and 4. We also listed the human end2end performance in Table 2, which is summarized from

the divergences between the annotations from the first two annotators and the final annotations. From these tables, we can see that:

1) The proposed formulation for event-centric opinion mining is a feasible task for human beings. From Table 2, we can see that human can reach high agreements on both ECO-ZH and ECO-EN. This demonstrates that the proposed opinion connotation and structure are applicable for event-centric opinions.

2) Event-centric opinion mining is a challenging task. The best benchmark system Seq-MRC can only achieve 28.01 and 18.15 segment-level F_1 on Chinese and English respectively. The performance gap between machine and human is huge, which indicates that more effective architectures and task-specialized approaches are needed.

3) Seq-MRC architecture achieved the best performance among 4 baseline architectures. We believe this is because the architecture is a more natural design for event-centric opinion mining. Naturally, EOE is a sentence-level sequential labeling problem given the event, and OTE is a span extraction problem given the opinion. As a result, Seq-MRC is more suitable for solving these two tasks compared with PairCls and SpanR.

4) The main bottleneck for event-centric opinion mining is to identify completed continuous opinion snippets from documents. From Table 3, we can see that current sentence-level sequential labeling can only achieve 44.33% and 34.84% segment-level F_1 score, which is the main reason for the low end2end performance. By contrast, we can see that the sentence-level evaluation results are much better than segment-level evaluation results. We believe the reason behind is that current architectures can not well identify the structural relations at sentence-level, and leveraging such structure requires strong discourse-level knowledge.

5) ECO-EN dataset is more challenging than ECO-ZH. Even with similar training document size and opinion numbers, the performance of ECO-EN is significantly worse than that of ECO-ZH. We believe this is because 1) English opinions are often more implicit than Chinese ones. Therefore, even human annotators made more disagreements on ECO-EN; 2) ECO-EN is with much fewer event descriptors than ECO-ZH, which make the training of EOE models may overfit on the events in the training data.

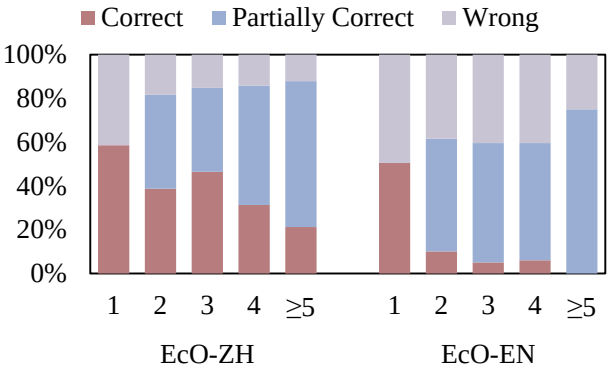

Figure 3: Performance of *Seq* on EOE with different opinion lengths.

5.3 Effects of Opinion Length to EOE

To investigate whether opinion length will impact the performance of EOE, we categorize the model's prediction on golden opinion snippets into: 1) Correct; 2) Partially Correct, which means that at least one sentence in the opinion segment is identified; 3) Wrong. Figure 3 shows the results of *Seq* approach. We can see that the correct prediction ratio drops when opinion length increases. This is easy to understand because opinions with more sentences are more difficult to recognize. However, we can see that the wrong prediction ratio also drops along with the increase of opinion length. This indicates that for longer opinions, the chance of at least one sentence can be correctly identified is relatively high. Therefore, if we can jointly consider the predictions of multiple sentences by leveraging discourse knowledge, we may reduce such partial labeling errors and improve the performance.

5.4 Effects of Argument Type to OTE

	Subevents	Entities	Events
ECO-ZH	76.80	54.40	72.92
ECO-EN	25.00	48.48	70.85

Table 5: Performance of *MRC* on different kinds of arguments.

Table 5 shows the opinion target extraction performance of *MRC* on different types of arguments. For both ECO-ZH and ECO-EN dataset, the performance on whole events is better than that on arguments. In particular, it performs poorly on subevents in ECO-EN. This may be because the amount of event descriptors is not enough for the model to learn to extract the exact boundaries of subevent arguments.

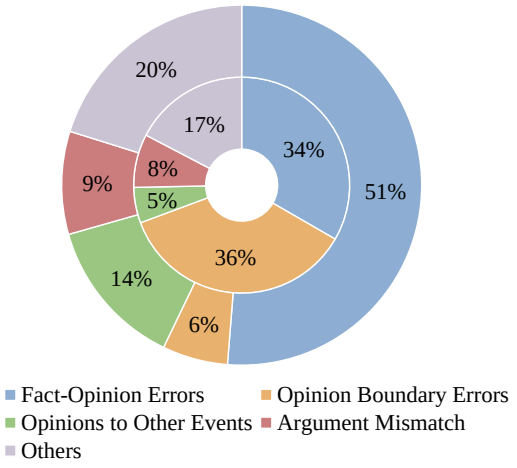

Figure 4: Proportions of error causes on ECO of *Seq-MRC*. Inner circle refers to performance on ECO-ZH and outer circle corresponds to ECO-EN dataset.

5.5 Error Analysis & Discussion

To better understand the challenge and bottlenecks of event-centric opinion mining, we further randomly sampled 50 annotated documents from ECO-ZH and ECO-EN respectively. We then categorized the errors made by Seq-MRC model to figure out the critical issues to resolve. From Figure 4, we can see that:

1) The confusions between facts and opinions are one of the most critical EOE errors for both English and Chinese portions. 51% errors in English and 34% errors in Chinese portion stem from fact-opinion confusion. This corresponds to the nature of the task because event-centric opinions are frequently mixed up with many facts and non-opinion information. And a sentence can contain both opinion and fact at the same time, which makes it very difficult to identify.

2) Opinion boundary errors are more significant in Chinese portion than English portion. Compared with 6% boundary errors in English portion, the percentage in Chinese portion is a much higher 36%. We believe this is because the average length of opinions in Chinese is longer than that in English, which is shown in Table 1. As a result, more opinion boundary errors are introduced in EOE. Furthermore, by looking into the error cases, we find that such errors mainly occur in cases where two continuous opinions refer to different arguments, which is very challenging.

3) OTE errors are more severe in English portion than Chinese portion. We find that such errors happened more frequently on the opinions with implicit targets. Furthermore, there are notable 14% errors in English portion that comes from identifying opinions not corresponding to the given event. This usually happens when models are confused by strong opinion marker words like *say* and *believe*, and similar arguments such as World War I and World War II. We believe that this is because event descriptors in the English portion are much less than Chinese portion. As a result, models overfit on some spurious features and can not sufficiently capture the correct event-oriented information. To alleviate this problem, we will enlarge the English portion of ECO Bank in the future.

6 Related Work

Previous opinion mining (OM) researches focus on entity-centric opinions (Liu, 2007), which mainly categorizes the holder's sentiments towards entities and their attributes at document-level (Turney, 2002; Moraes et al., 2013; Sharma et al., 2014; Tang et al., 2015; Paredes-Valverde et al., 2017), sentence-level (Hatzivassiloglou and Wiebe, 2000; Riloff and Wiebe, 2003; Hu and Liu, 2004; Riloff et al., 2006; Sayeed et al., 2012; Alessia et al., 2015) and aspect-level (Jin et al., 2009; Li et al., 2010; Qiu et al., 2011; Liu, 2012; Mitchell et al., 2013; Liu et al., 2015; Wang et al., 2017; Zhao et al., 2020; Peng et al., 2020; Cai et al., 2021; Mao et al., 2021).

There are also some researches working on event-related opinions (Karamibekr and Ghorbani, 2012; Zhou et al., 2013; Deng and Wiebe, 2015b,a; Qian et al., 2016; Maynard et al., 2017). Generally speaking, these studies commonly regard event as a special type of entity, neglecting the unique characteristics of event-centric opinions. However, events are very different from entities, and therefore event-centric opinions have different connotations and targets which have not been exploited yet.

For the evaluation resource of OM, most of current studies are based on the Semeval Challenges datasets (Pavlopoulos, 2014; Pontiki et al., 2015, 2016) and its extension (Wang et al., 2017; Fan et al., 2019; Peng et al., 2020), which consist of entity-centric customer reviews about target entities from 7 domains. To the best of our knowledge, the constructed ECO Bank is the first publicly available event-centric opinion mining benchmark from news domain, which definitely can benefit future research in this direction.

7 Conclusions and Future Work

In this paper, we propose and formulate ***event-centric opinion mining***, a new task that aims to mine a broader range of opinions oriented to specific events from documents. An Event-Centric Opinion Bank corpus is constructed and a two-step framework is proposed. Experiments demonstrate the challenges and advantages of mining event-centric opinions. The focus of this paper is the introduction of the new task and datasets. The proposed four baseline systems are relatively simple and leave much room for further improvements. In future work, we will try to build end-to-end models that directly extract opinion triples in an end-to-end fashion and enrich the current opinion structure.

8 Ethics Consideration

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

1. All of the collected documents and event descriptors come from publicly available sources. The legal advisor of our institute and/or the original dataset constructor confirms that the sources of our data are freely accessible online without copyright constraint to academic use.

2. ECO Bank contains 5000 annotated documents with 988 event descriptors. After double-checking, we guarantee that ECO Bank doesn't contain samples that may cause ethic issues. The dataset does not involve any personal sensitive information. All references in the annotated data are double-checked for plausibility and grammaticality by different human annotators. All documents and event descriptors are also manually checked to ensure they are informative and logically coherent. We manually check the content of each piece of data in ECO Bank to ensure that it does not contain any hate speech or attacks on vulnerable people.

3. We hired 5 annotators who have bachelor degrees. Before formal annotation, annotators were asked to annotate 20 samples randomly extracted from the dataset, and based on average annotation time we set a fair salary (i.e., 35 dollars per hour) for them. During their training annotation process, they were paid as well.

Acknowledgement

We thank all reviewers for their valuable comments. Moreover, this work was supported by the National Key Research and Development Program of China (No. 2020AAA0106400), and the National Natural Science Foundation of China under Grants no. U1936207, 62122077 and 62106251.

References

D Alessia, Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. 2015. Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3).

Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2009. Appraisal of opinion expressions in discourse. *Lingvisticæ Investigationes*, 32(2):279–292.

Ann Banfield. 1984. Unspeakable sentences: Narration and representation in the language of fiction. *Journal of Aesthetics and Art Criticism*, 43(1).

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. 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 340–350, Online. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011, Online. Association for Computational Linguistics.

John Corvino. 2014. The fact/opinion distinction. *The Philosophers' Magazine*, (65):57–61.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Lingjia Deng and Janyce Wiebe. 2015a. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189, Lisbon, Portugal. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015b. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

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.

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.

Zhifang Fan, Zhen Wu, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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 2509–2518, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert A Hackett. 1984. Decline of a paradigm? bias and objectivity in news media studies. *Critical Studies in Media Communication*, 1(3):229–259.

Vasileios Hatzivassiloglou and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Tuan-Anh Hoang, Khoi Duy Vo, and Wolfgang Nejdl. 2018. W2e: A worldwide-event benchmark dataset for topic detection and tracking. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1847–1850.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on events: Definition, detection, coreference, and representation*, pages 21–28.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A novel lexicalized hmm-based learning framework for web opinion mining. In *Proceedings of the 26th annual international conference on machine learning*, volume 10. Citeseer.

Mostafa Karamibekr and Ali A Ghorbani. 2012. Sentiment analysis of social issues. In *2012 International Conference on Social Informatics*, pages 215–221. IEEE.

Mostafa Karamibekr and Ali A Ghorbani. 2013. Sentence subjectivity analysis in social domains. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 268–275. IEEE.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 653–661, Beijing, China. Coling 2010 Organizing Committee.

Bing Liu. 2007. Opinion mining. *Web data mining: Exploring hyperlinks, contents, and usage data*, pages 411–447.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal. Association for Computational Linguistics.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. *arXiv preprint arXiv:2101.00816*.

Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout, and Kalina Bontcheva. 2017. A framework for real-time semantic social media analysis. *Journal of Web Semantics*, 44:75–88.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.

Rodrigo Moraes, João Francisco Valiati, and Wilson P GaviãO Neto. 2013. Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*, 40(2):621–633.

Walter J Ong. 1969. World as view and world as event 1. *American Anthropologist*, 71(4):634–647.

Mario Andrés Paredes-Valverde, Ricardo Colomo-Palacios, María del Pilar Salas-Zárate, and Rafael Valencia-García. 2017. Sentiment analysis in spanish for improvement of products and services: A deep learning approach. *Scientific Programming*, 2017.

Ioannis Pavlopoulos. 2014. Aspect based sentiment analysis. *Athens University of Economics and Business*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

James Pustejovsky. 1991. The syntax of event structure. *cognition*, 41(1-3):47–81.

Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. 2016. Multi-modal multi-view topic-opinion mining for social event analysis. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 2–11.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. 2006. Feature subsumption for opinion analysis. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 440–448.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112.

Bertrand Russell. 1927. The analysis of matter.

Erik F Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. *arXiv preprint cs/0009008*.

Asad Sayeed, Jordan Boyd-Graber, Bryan Rusk, and Amy Weinberg. 2012. Grammatical structures for word-level sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for computational Linguistics: Human language technologies*, pages 667–676.

Frederick F Schauer. 1978. Language, truth, and the first amendment: An essay in memory of harry canter. *Va. L. Rev.*, 64:263.

Richa Sharma, Shweta Nigam, and Rekha Jain. 2014. Opinion mining of movie reviews at document level. *arXiv preprint arXiv:1408.3829*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal. Association for Computational Linguistics.

Peter D Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: a span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248.

Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th international conference on computer supported cooperative work in design (CSCWD)*, pages 557–562. IEEE.