Document-Level Event Factuality Identification via Adversarial Neural Network

Zhong Qian[1], **Peifeng Li**[1 2], **Qiaoming Zhu**[1 2] **and Guodong Zhou**[1 2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]AI Research Institute, Soochow University, Suzhou, China
qianzhongqz@163.com, {pfli, qmzhu, gdzhou}@suda.edu.cn

Abstract

Document-level event factuality identification is an important subtask in event factuality and is crucial for discourse understanding in Natural Language Processing (NLP). Previous studies mainly suffer from the scarcity of suitable corpus and effective methods. To solve these two issues, we first construct a corpus annotated with both document- and sentence-level event factuality information on both English and Chinese texts. Then we present an LSTM neural network based on adversarial training with both intra- and inter-sequence attentions to identify document-level event factuality. Experimental results show that our neural network model can outperform various baselines on the constructed corpus.

1 Introduction

Document-level event factuality identification is the task of deciding the commitment of relevant sources towards the factual nature of an event, and to determine whether an event is a fact, a possibility, or an impossible situation from the view of document. Identifying document-level factuality of events requires comprehensive understanding of documents. As illustrated in Figure 1 where events are in ***bold***, the event "*reach*" (including its other forms) have various factuality values in different sentences. For example, in paragraph 2, "*reach*" is impossible/CT- according to the negative word "*denied*", while in paragraph 3, "*reach*" is possible/PS+ due to the speculative word "*may*". The main contents of this document is "*Mexico denied that they will reach an agreement with the U.S. on the new trade deal*", and the document-level factuality of the event "*reach*" is CT-.

Document-level event factuality identification is fundamental for document-level NLP applications, such as machine reading comprehension, which aims to have machines read a text passage and then answer questions about the text. According to the document in Figure 1, the answer of the following question should be "*No*", which is consistent with the document-level factuality of the event "*reach*" (CT-):

***Q**: Does the U.S. reach an agreement with Mexico on the new trade deal before December 2017?*

***A**: No.*

According to Politico.com, it is said the United States will ***reach*****(CT+)** *an agreement with Mexico on the new trade deal that will replace North American Free Trade Agreement (NAFTA) before December, 2017.*

However, Mexican Economy Minister Ildefonso Guajardo denied that they plan to ***reach*****(CT-)** *any agreement with the U.S. on the trade deal talks.*

"We are not going to sacrifice the quality of an agreement because of pressure of time. We will keep engaged." he said. Just two days ago, Guajardo said the two sides may ***reach*****(PS+)** *an agreement within hours.*

The government has not been informed that any agreement will be ***reached*****(CT-)** *yet, said another two Mexican officials.*

During the past few weeks, the U.S. has been negotiating with Mexico on the new trade deal and has achieved much progress. Thus, some media speculate that they will possibly ***reach*****(PS+)** *an agreement. But now it seems that the negotiations will continue before they can get a good deal.*

(Time: November, 2017)
(Document-level factuality of the event "*reach*" is **CT-**.)

Figure 1: An example document with both sentence- and document-level event factuality.

Previous studies mostly reported on sentence-level event factuality identification tasks. On one hand, due to the scarcity of document-level event factuality corpus, these studies only considered the corpora annotated with sentence-level event factuality information, such as ACE 2005[1], LU (Diab et al., 2009), FactBank (Saurí and Pustejovsky, 2009), and UDS-IH2 (Rudinger et al., 2018).

On the other hand, previous studies only con-

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

Proceedings of NAACL-HLT 2019, pages 2799–2809
Minneapolis, Minnesota, June 2 - June 7, 2019.

sidered information within sentences, using rules (Saurí, 2008; Saurí and Pustejovsky, 2012), machine learning models (de Marneffe et al., 2012; Werner et al., 2015; Baly et al., 2018), and combinations of them (Qian et al., 2015; Stanovsky et al., 2017) for modeling. Neural network models have also recently been used for the sentence-level event factuality identification (He et al., 2017; Rudinger et al., 2018; Qian et al., 2018). According to Figure 1, document-level event factuality can not be deduced from each sentence-level factuality separately, but depends on the comprehensive semantic information of sentences. However, no suitable model for document-level task has been proposed yet.

To solve the issues above, this paper focuses on **document-level** event factuality identification. Our contributions can be summarized as follows.

1) We construct a document-level event factuality corpus, i.e. DLEF, on both English and Chinese texts. To our best knowledge, this is the first document-level event factuality corpus. The statistics on the corpora and the experimental results show that our corpus can sufficiently reflect linguistic characteristics of news texts, and provide adequate support on resource for research.

2) We propose an LSTM neural network with both intra- and inter-sequence attentions to identify document-level event factuality, and consider dependency paths from speculative and negative cues to the event and sentences containing the event as features. Due to the diversity of various contents of the texts in DLEF corpus, we employ Adversarial Training to improve the robustness of our model. Experimental results show that our model is superior to various baselines. The corpus and code of this paper will be released at https://github.com/qz011/dlef.

2 Corpus Annotation

This section introduces our **D**ocument-**L**evel **E**vent **F**actuality (DLEF) corpus, including the source, detailed guidelines for both document- and sentence-level event factuality, and the main statistics of the corpus.

2.1 Source

News texts contain sufficient speculative and negative information that is significant for event factuality identification, and usually focus on one event with a specific topic. Moreover, FactBank (Saurí and Pustejovsky, 2009), the sentence-level event factuality corpus, is also based on news texts.

	+	-	u
CT	CT+	CT-	CTu
PS	PS+	PS-	(NA)
U	(NA)	(NA)	Uu

Table 1: Event factuality values.

Therefore, we choose news texts in both English and Chinese to construct our corpus. The English corpus consists of 1727 documents from January 2017 to January 2018, among which 1506 documents are from China Daily[2], and 221 documents are from Sina Bilingual News[3]. The Chinese corpus consists of 4649 documents from Sina News[4]. These news documents cover various topics, e.g., politics, economy, culture, military, and society, which can reflect the heterogeneity of language in news texts.

2.2 Factuality Values

Saurí (2008) employed modality and polarity to describe event factuality values. Modality conveys the certainty degree of events, such as *certain* (CT), *probable* (PR), and *possible* (PS), while polarity expresses whether the event happened, including *positive*(+) and *negative*(-).

We use the factuality values in Table 1 according to Saurí (2008). Both PR and PS are speculative values and share similar certainty degrees in our corpus, and are merged into PS . U/u means *underspecified*. PSu and U+/- are not applicable (NA) and are not considered. Although CTu is applicable, neither document-level nor sentence-level event can be annotated as CTu in our corpus.

2.3 Annotation Guidelines

We adopt the definition of events proposed by TimeML (Pustejovsky et al., 2003) and consider the events that can be critical for computing the factuality. To ensure that the task is meaningful, we focus on the events that have various types of sentence-level factuality values. If there is more than one suitable event in a document, we annotate them separately.

First, the annotation of document-level event factuality is based on the definition, i.e., determining the factuality of an event from the view of

[2]http://www.chinadaily.com.cn
[3]http://roll.edu.sina.com.cn/english/syxw/index.shtml
[4]http://news.sina.com.cn

the document requires to understand the semantic of the document, including various sentence-level event factuality.

Second, sentence-level event factuality is essential for document-level task, which makes sense when document- and sentence-level factuality of events have different values. Therefore, we annotate the sentence-level event factuality as follows:

CT- events are negated by negative cues. For example, the events "*enter*" and "*merger*" are governed by negative cues "*impossible*" and "*denied*" in sentence S1 and S2, respectively.

(S1) *He said that the loss made it impossible for them to* ***enter*** *the semifinals.*

(S2) *Sinopec responded to National Business Daily, and denied the rumors of a* ***merger*** *with PetroChina.*

PS+ events (e.g. "*improve*" and "*fallen*") are governed by speculative cues (e.g., "*impossible*" and "*denied*"), just as illustrated in sentence S4 and S5.

(S4) *We think that further investigation may help to* ***improve*** *the treatment of people with similar infections.*

(S5) *The missing parts may have* ***fallen*** *during the flight of the plane.*

PS- events are governed by both speculative and negative cues. Different from CT-, PS- means incompletely negation. For example, the PS- event "*noticed*" is governed by the speculative cue "*probably*" and the negative cue "*not*" in sentence S6, and "*fall*" is modified by the cues "*may*" and "*not*" in sentence S7.

(S6) *The bus driver had probably not* ***noticed*** *the truck early enough.*

(S7) *Oil prices may not* ***fall*** *sharply due to the strong global demand.*

Uu events can appear in questions (e.g., "*considering*" in sentence S8) and in the intensional contexts with underspecified semantics (e.g., "*raises*" in sentence S9):

(S8) *Is France* ***considering*** *to leave EU?*

(S9) *The US dollar's declination can not be reversed even if the Federal Reserve* ***raises*** *rates three times.*

CT+ events are factual and do not meet the above conditions.

2.4 Statistics

The task is trivial if most documents have only one type of sentence-level factuality value, and in this case, document-level factuality probably shares the same value. To understand the usefulness of document-level event factuality identification and DLEF corpus, we launched the statistics of documents with n different types of sentence-level event factuality values shown in Table 2. From the table we can find that for English corpus there are 41.94% CT- and 66.06% PS+ documents with different sentence-level event factuality values, but these CT+ documents only cover 11.13%. While for Chinese corpus, these CT- and PS+ documents cover 63.41% and 62.15%, but these CT+ documents only make up 14.23%.

Corpus	Docs	n=1	n=2	$n\geqslant$3
English	CT-	162	97	20
	PS+	93	157	24
	PS-	2	6	4
	Uu	5	6	1
	CT+	1022	119	9
	Total	1284	385	58
Chinese	CT-	491	612	239
	PS+	321	425	102
	PS-	9	11	16
	Uu	8	5	7
	CT+	2061	290	52
	Total	2890	1343	416

Table 2: Statistics of the documents in DLEF corpus with n types of sentence-level event factuality values.

Table 2 indicates that sentence-level factuality usually agrees with document-level factuality in CT+ documents, making them straightforward to be identified. However, in those non-CT+ documents with non-factual document-level values, sentence-level factuality is likely to have different values from documents, making them more difficult to be identified. In general, English and Chinese corpus have 25.64% and 37.84% documents with different sentence-level event factuality values, indicating this corpus is suitable for the document-level event factuality identification.

Table 3 shows the statistics of the DLEF corpus. CT+ document-level events are in the majority, because information reported by news texts is usually real.

Kappa (Cohen, 1960) is employed to measure the inter-annotator agreement of annotating document- and sentence-level event factuality between the two independent annotators who annotate the entire corpus, just as shown in Table 4. These two annotators are postgraduate stu-

Corpus	Statistics		
English	Documents	CT-	279/16.16%
		PS+	274/15.87%
		PS-	12/0.69%
		Uu	12/0.69%
		CT+	1150/66.59%
		Total	1727
	Sentence-Level Events	CT-	662/11.52%
		PS+	574/9.99%
		PS-	37/6.44%
		Uu	71/1.24%
		CT+	4401/76.61%
		Total	5745
	Avg. Len. of Sentences		14.73
	Avg. Len. of Documents		467.25
Chinese	Documents	CT-	1342/28.87%
		PS+	848/18.24%
		PS-	36/0.77%
		Uu	20/0.43%
		CT+	2403/51.69%
		Total	4649
	Sentence-Level Events	CT-	3923/20.69%
		PS+	2879/15.18%
		PS-	123/0.65%
		Uu	555/2.93%
		CT+	11482/60.55%
		Total	18962
	Avg. Len. of Sentences		29.00
	Avg. Len. of Documents		716.38

Table 3: Statistics of DLEF corpus. The units of length of English and Chinese texts are tokens and Chinese characters, respectively.

Corpus	Value	Sent-Level	Doc-Level
English	All	0.81	0.91
	CT-	0.82	0.89
	PS+	0.77	0.87
	CT+	0.84	0.93
Chinese	All	0.82	0.81
	CT-	0.83	0.82
	PS+	0.79	0.78
	CT+	0.83	0.84

Table 4: Inter-annotator agreement of event factuality.

dents who major in NLP. In addition, the Kappa of events on English and Chinese corpus are 0.83 and 0.85, respectively. All the Kappa values are larger than 0.75, proving the effectiveness and meaningfulness of our DLEF corpus.

3 Adversarial Neural Network for Document-Level Event Factuality Identification

This section describes the LSTM neural network for document-level event factuality identification in detail. As shown in Figure 2, to extract feature representations of events from the view of documents, we consider both intra- and inter-sequence attention for dependency paths and sentences. In addition, due to the diversity of contents of documents in DLEF corpus, we consider adversarial training to ensure the robustness of our model.

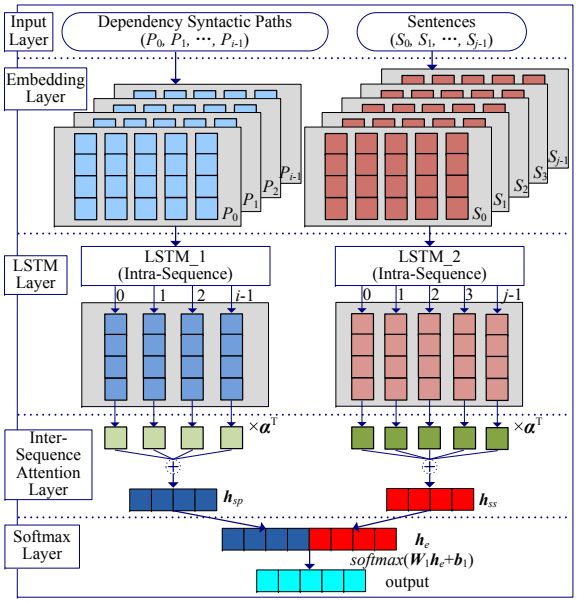

Figure 2: Neural network architecture for document-level event factuality identification.

3.1 Input Features

For our task, we use the specified events that have been annotated, and utilize the Chinese cues in CNeUn corpus (Zou et al., 2015) and the English cues in BioScope corpus (Vincze et al., 2008) that also considers multi-word cues, e.g., *rule out*. We do not use any annotated sentence-level event factuality. For one event, we consider all the sentences containing it, and mainly employ the following two features in our model:

1) **Syntactic Features**: Previous studies (Saurí and Pustejovsky, 2012; de Marneffe et al., 2012) have proved the effectiveness of dependency trees on event factuality identification tasks. Hence, we employ the dependency paths from speculative or negative cues to the event as syntactic features.

2) **Semantic Features**: We use the sentences containing the event as semantic features.

In addition, we also consider the above features in contexts of each sentence containing the event as the input, and set the windows size as 3, i.e., one sentence before and after the current one. If adjacent sentences contain speculative or negative cues, the dependency path is the concatenation of the path from the cue to the root and the path from the root to the event (Quirk and Poon, 2017).

3.2 LSTM with Two Attention Layers

A dependency path or sentence can be represented as $\boldsymbol{X}_0$ according to the embedding table. We employ LSTM with hidden units n_h to model the sequences from both directions to produce the forward hidden sequence $\overrightarrow{\boldsymbol{H}}$, the backward hidden sequence $\overleftarrow{\boldsymbol{H}}$, and the output sequence $\boldsymbol{H} = \overrightarrow{\boldsymbol{H}} + \overleftarrow{\boldsymbol{H}}$. We adopt the attention mechanism to capture the most important information from $\boldsymbol{H}$, and obtain the output $\boldsymbol{h}$:

$$\boldsymbol{H}_m = \tanh(\boldsymbol{H}) \quad (1)$$

$$\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{v}^T \boldsymbol{H}) \quad (2)$$

$$\boldsymbol{h} = \tanh(\boldsymbol{H}\boldsymbol{\alpha}^T) \quad (3)$$

where $\boldsymbol{v} \in \mathbb{R}^{n_h}$ is the parameter. One event can have k sequences $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}$, whose representation is $\boldsymbol{H}_s = \boldsymbol{h}_0, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_{k-1}$ according to the above equations, where $\boldsymbol{H}_s \in \mathbb{R}^{k \times n_h}$. To extract the feature representation $\boldsymbol{h}_s \in \mathbb{R}^{n_h}$ from the k sequences, we utilize an inter-sequence attention mechanism that is computed as:

$$\boldsymbol{H}_{ms} = \tanh(\boldsymbol{H}_s) \quad (4)$$

$$\boldsymbol{\alpha}_s = \mathrm{softmax}(\boldsymbol{v}_s^T \boldsymbol{H}_{ms}) \quad (5)$$

$$\boldsymbol{h}_s = \tanh(\boldsymbol{H}_s \boldsymbol{\alpha}_s^T) \quad (6)$$

where $\boldsymbol{v}_s \in \mathbb{R}^{n_h}$ is the parameter. Suppose that an event has i dependency paths $\boldsymbol{P}_0, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_{i-1}$, and appears in j sentences $\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_{j-1}$. Considering that dependency paths and sentences contain syntactic and semantic information, respectively, we employ two LSTM neural networks defined above to learn vector representations $\boldsymbol{h}_{sp}$ and $\boldsymbol{h}_{ss}$ of dependency paths and sentences, and concatenate them into the feature representation of the event $\boldsymbol{h}_e$:

$$\boldsymbol{h}_e = \boldsymbol{h}_{sp} \oplus \boldsymbol{h}_{ss} \quad (7)$$

where $\oplus$ is the concatenation operator. Finally, $\boldsymbol{h}_e$ is fed into the softmax layer to compute the probability of the factuality values of the event:

$$\boldsymbol{o} = \mathrm{softmax}(\boldsymbol{W}_1 \boldsymbol{h}_e + \boldsymbol{b}_1) \quad (8)$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{c \times \dim(\boldsymbol{h}_e)}$ and $\boldsymbol{b}_1 \in \mathbb{R}^c$ are parameters, and $c = 5$ is the number of categories of factuality values (CT+, CT-, PS+, PS-, Uu). The objective function of the proposed neural network is designed as:

$$L_D(\theta) = -\frac{1}{m} \sum_{i=0}^{m-1} \log p(y_j^{(i)} | x^{(i)}, \theta) \quad (9)$$

where $y^{(i)}$ is the golden label of the instance $x^{(i)}$ and $p(y_j^{(i)}|x^{(i)})$ is the probability, m is the number of instances, and θ is the parameter set to learn. This model with **TWO** attention layers is denoted as **Att_2** in the next section.

3.3 Adversarial Training

As described in Section 2, documents in DLEF corpus cover various topics. To improve the robustness of our model, we consider Adversarial Training. Similar to previous work (Miyato et al., 2016; Wu et al., 2017), we add a small adversarial perturbation $\boldsymbol{e}_{adv}$ to word embeddings, and employ the following objective function:

$$L_{adv}(\boldsymbol{X}|\theta) = L(\boldsymbol{X} + \boldsymbol{e}_{adv}|\theta) \quad (10)$$

$$\boldsymbol{e}_{adv} = \underset{\|\boldsymbol{e}\| \leqslant \epsilon}{\arg\max}\, L(\boldsymbol{X} + \boldsymbol{e}|\hat{\theta}) \quad (11)$$

where $\hat{\theta}$ is a fixed copy value of the current θ and $\boldsymbol{X}$ is the input. Due to the intractable nature in the computation of Eq. (11), Goodfellow et al. (2014) proposed Eq. (12) to linear $L(\boldsymbol{X}|\theta)$ near $\boldsymbol{X}$ to approximate Eq. (11):

$$\boldsymbol{e}_{adv} = \epsilon \boldsymbol{g} / \|\boldsymbol{g}\| \quad (12)$$

$$\boldsymbol{g} = \nabla_{\boldsymbol{T}} L(\boldsymbol{X}|\hat{\theta}) \quad (13)$$

where $\boldsymbol{T}$ is the embedding table.

4 Experiments

We introduce the experimental settings and the baselines, finally presenting the experimental results and analysis in detail.

4.1 Experimental Settings

The PS- and Uu documents only cover 1.39% and 1.20% in our English and Chinese corpus, respectively. Therefore, we mainly focus on the performance of CT+, CT-, and PS+.

For fair comparison, we perform 10-fold cross-validation on English and Chinese corpora, respectively. In addition to Precision, Recall, and F1-measure for each category of factuality value, we consider macro- and micro-averaging to obtain the overall performance of all the categories of factuality values. The hidden units of LSTM are set as $n_h = 50$. We initialize word embeddings via Word2Vec (Mikolov et al., 2013), setting the dimensions as $d_0 = 100$, and fine-tuning them during training. SGD with momentum is applied to optimize our models.

Att_2 and **Att_2+AT** are the models proposed in Section 3 that consider the contexts, i.e., one sentence before and after the current sentence containing the event as the input. Compared to Att_2, Att_2+AT considers Adversarial Training (AT, the same below). We also consider the following baselines for the comparison with our models:

MaxEntVote is a maximum entropy model that only considers the view of *AUTHOR* (de Marneffe et al., 2012). We use maximum entropy model to identify sentence-level event factuality, and consider voting mechanism, i.e., choose the value committed by the most sentences as the document-level factuality value. We also consider other machine learning models, e.g. Lee et al. (2015), but obtain lower micro-/macro-averaged F1 on English (59.38/33.36) and Chinese corpus (53.91/43.20).

SentVote identifies sentence-level event factuality, and does not consider inter-sequence attention in the model proposed in Section 3. Similar to MaxEntVote model, voting mechanism is used to identify document-level event factuality in this SentVote model.

MP_2 considers Max-Pooling instead of attention compared with Att_2.

Att_1 considers only intra-sequence attention, but not the inter-sequence attention. For an event, we concatenate its i dependency paths and j sentences into one path and one sentence as the input, respectively.

4.2 Results and Analysis

4.2.1 Architecture of Neural Networks

Table 5 presents the performances of our models and baselines. MaxEntVote gives relatively lower results than other models, especially on CT- and PS+. SentVote models are better than MaxEntVote, but still obtain lower results than Att_2, which can prove that inter-sequence attention is more useful than voting. Max-pooling only selects the most active information for each dimension of features, While attention takes into account all the features and assigns weights for them according their degrees of importance. Hence, Att_2 gets better results than MP_L2. Att_1 only considers the intra-sequence attention and obtains lower results than Att_2, which proves the effectiveness of inter-sequence attention. Att_2 and Att_2+AT achieve better results than other baselines. Compared to Att_2, Att_2+AT considers the adversarial perturbation and training that can alleviate overfitting. Therefore, Att_2+AT is superior to Att_2, which can prove the effectiveness of adversarial training.

On both English and Chinese corpora, the performance of CT+ is better than those of PS+ and CT-. On one hand, it is easier to identify CT+ documents due to their majority. On the other hand, most news texts hardly contain bogus and false contents. Therefore, in most CT+ documents, sentence-level factuality values are consistent with the document-level value, just as in S10. However, in PS+ and CT- documents with non-CT+ document-level values, sentence-level factuality values have different viewpoints with the corresponding document, varying among CT-, PS+, and CT+, making the task more difficult, e.g., S11.

(S10) *India successfully* ***tested***(CT+) *a supersonic missile, capable of destroying an incoming ballistic missile at low altitude. ..... The* ***test***(CT+) *was carried out from a test range in Odisha , official sources said.*

(S11) *Argentine navy said it had not* ***contacted***(CT-) *the SAN Juan submarine. ... ... Some media previously said the navy may have received signals from the submarine and* ***contacted***(PS+) *it.*

4.2.2 Input of Neural Networks

For Att_2+AT, we also investigate the effects of contexts of the sentences containing events as the input on the performance. The results is given in Table 6, which shows that contexts can improve the performance more significantly on the Chinese corpus than the English corpus. We find that in the Chinese corpus these sentences are commonly in the same paragraph and have a strong semantic coherence. Therefore, information in adjacent sentences can contribute to the identification of the document-level factuality of the events in the current sentences.

Sentences S12 and S13 are adjacent sentences in one paragraph. The document-level factuality value of the event "*provided*" in S12 is CT-. However, the sentence-level value of "*provided*" is PS+. If we consider S13, the negative cue "*denied*" can lead to the correct document-level factuality value of "*provided*". While in the English corpus, similar sentences are much fewer, because paragraphs in most English news texts only contain one or two sentences, and sentences in different paragraphs share less semantic correlation

Corpus	Systems	CT-	PS+	CT+	Micro-A	Macro-A
English	MaxEntVote	58.17	35.89	75.14	68.42	56.40
	SentVote	70.22	57.85	83.98	78.06	70.68
	MP_2	70.57	56.39	83.72	77.65	70.23
	Att_1	65.25	53.65	79.18	73.23	66.03
	Att_2	73.88	59.29	88.59	81.84	73.92
	Att_2+AT	**76.87**	**62.14**	**89.84**	**83.56**	**76.28**
Chinese	MaxEntVote	62.44	58.29	72.22	67.72	64.32
	SentVote	72.66	58.39	80.68	74.70	70.58
	MP_2	74.34	65.17	78.91	75.22	72.81
	Att_1	68.82	49.78	81.89	71.12	67.28
	Att_2	81.41	73.35	86.58	82.79	80.45
	Att_2+AT	**83.35**	**74.06**	**87.52**	**84.03**	**81.64**

Table 5: F1-measures of baselines and our model.

Corpus	Systems	CT-	PS+	CT+	Micro-A	Macro-A
English	Att_2+AT	76.87	62.14	89.84	83.56	76.28
	w/o CTX	+0.95	+0.47	-0.80	-0.31	+0.21
	w/o Dpath (Only Sent)	-16.49	-8.28	-4.88	-7.08	-9.88
	w/o Sent (Only Dpath)	-20.79	-12.63	-19.25	-19.20	-17.56
Chinese	Att_2+AT	83.35	74.06	87.52	84.03	81.64
	w/o CTX	-3.05	-2.62	-1.03	-1.84	-2.23
	w/o Dpath (Only Sent)	-10.31	-7.23	-7.80	-8.49	-8.44
	w/o Sent (Only Dpath)	-17.53	-11.02	-14.79	-15.19	-14.44

Table 6: F1-measures of Att_2+AT with different input features.

than those in the same paragraph. Hence, performance improvement is less when considering adjacent sentences in the English corpus.

(S12) 外界质疑在竞标过程中，墨西哥政府为相关企业**提供**了“有利位置”。(*It is doubted that the Mexican government **provided** "vantage points" for the enterprises involved during the bidding process.*)

(S13) 墨西哥外交部在7日对此予以回应，否认了这种说法。(*The Mexican Foreign Ministry responded and denied the rumor on 7th.*)

If we consider more adjacent sentences, e.g., two sentences before and after the current sentence, however, the results will be a bit lower. The micro-/macro-averaged F1 on English and Chinese corpus are 81.20/75.65 and 82.57/80.91, respectively. We think the reason is that some sentences are far away from the current sentence and have little effect on the current event, and considering more contexts may also lead to overfitting.

Moreover, we explore the effects of considering only dependency path (Dpath) and only sentence (Sent) in Table 6. Att_2+AT achieves the best results when considering both paths and sentences as input, proving that both of them are effective features for our model. Att_2+AT obtains higher performance with only sentences than only paths as input, meaning that Att_2+AT is mainly beneficial from sentences that can offer semantic information. Error analysis shows that documents with incorrect identified values contains sentences with more speculative or negative cues:

(S14) *When asked if it might be **arson**, authorities said that no fire raiser has been found now, but the possibility of artificial arson should not been ruled out.*

S14 contains speculative cues “*if*”, “*might*”, “*possibility*” and negative cues “*no*”, “*not*”, “*ruled out*”. It is difficult to identify whether the events are governed by the cues when only considering the dependency paths and ignoring the semantic information offered by sentences. S14 can demonstrate the importance of semantic features.

4.2.3 Documents with Different Sentence-Level Event Factuality Values

As mentioned in Section 2.4, the document-level task becomes trivial if most documents have only one category of sentence-level factuality value that

Corpus	*n*	CT-	PS+	CT+	Micro-A	Macro-A
English	*n*=1	**85.63**	64.46	**94.91**	**91.36**	**81.67**
	n⩾2	58.33	**65.79**	56.01	60.91	60.04
Chinese	*n*=1	**84.45**	**76.73**	**93.44**	**89.85**	**84.87**
	n⩾2	82.68	69.91	58.51	73.22	70.37

Table 7: F1-measures of Att_2+AT on the documents with *n* types of sentence-level factuality values.

Corpus	Level	CT-	PS+	CT+	Micro-A	Macro-A
English	Sentence	72.05	59.68	91.14	85.96	74.29
	Document (with Joint Opt)	75.46	**62.80**	88.65	82.89	75.64
	Document (w/o Joint Opt)	**76.87**	62.14	**89.84**	**83.56**	**76.28**
Chinese	Sentence	74.20	68.88	87.73	81.98	76.94
	Document (with Joint Opt)	83.30	73.74	87.40	83.83	81.48
	Document (w/o Joint Opt)	**83.35**	**74.06**	**87.52**	**84.03**	**81.64**

Table 8: F1-measures of Att_2+AT with joint optimization.

is the same as document-level value. Table 7 shows the performance of Att_2+AT on the documents with n different types of sentence-level factuality values. The micro- and macro-averaged F1 of $n\geqslant 2$ are lower than those of n=1, indicating that the factuality of documents that have different types of sentence-level factuality are more difficult to identify due to the interference from sentence-level values.

We notice that in the Chinese corpus, the performance of CT- is much higher than that of PS+ and CT+ when $n\geqslant 2$. According to the analysis on the Chinese corpus, we find that most CT- documents are usually used to deny the rumors, i.e., those sentence-level events whose factuality values are not CT-. Therefore, the sentence-level CT- events are often in the topic sentences of the documents and dominate among sentences, which can contribute to the better results of document-level CT- events in Chinese corpus.

4.2.4 Joint Optimization Model

Because document-level event factuality is related with sentence-level factuality information, we also consider the joint optimization model for them. For sentence-level task, we use the LSTM neural network in Section 3 and only consider the current sentence, i.e., do not consider information in adjacent sentences and the inter-sequence attention layer. The objective of document- and sentence-level task are denoted as $L_D(\theta)$ and $L_S(\theta)$, and the objective of our joint optimization model is:

$$L_J(\theta) = \varepsilon L_D(\theta) + (1-\varepsilon)L_S(\theta) \qquad (14)$$

where ε=0.6 is the trade-off. The performance of both sentence-level and document-level event factuality identification is shown in Table 8. The micro-/macro-averaged F1 of joint optimization model on English and Chinese corpus are 82.89/75.64 and 83.83/81.48, respectively. Although document-level event factuality is based on the factuality information in sentences, sentence-level factuality value of an event only depends on the current sentence, and is likely to have a different value compared to the current document-level factuality. Therefore, the joint model can not improve the performance of document-level task.

5 Related Work

Researchers have studied document-level tasks in many NLP applications, e.g., sentiment analysis (Xu et al., 2016; Dou, 2017), named entity recognition (Luo et al., 2018), and machine translation (Born et al., 2017). But related studies on event factuality are limited to the sentence-level task. Diab et al. (2009) and Prabhakaran et al. (2010) presented studies of belief annotation and tagging, and classified predicate events into Committed Belief (CB), Non-CB or Not Applicable using a supervised framework. For factuality assessment, Lee et al. (2015) employed dependency features, while Stanovsky et al. (2017) considered deep linguistic information, such as modality classes, syntactic re-ordering with PropS tree annotation structure (Lotan et al., 2013). Baly et al. (2018) considered a set of features and predicted the factuality of reporting and bias of news media.

Saurí (2008) and Saurí and Pustejovsky (2012) proposed a rule-based model to identify event factuality on FactBank. de Marneffe et al. (2012) used a machine learning model and Qian et al. (2015) utilized a two-step framework combining machine learning and rule-based approaches on FactBank. In addition to FactBank, Prabhakaran et al. (2015) proposed a ongoing framework for a larger corpus based on LU, and Cao et al. (2013) constructed a Chinese corpus annotated with event factuality based on ACE 2005. However, no previous work annotated a document-level corpus. We construct DLEF corpus with document-level event factuality for the first time.

Some studies focused on document-level event identification task. Choubey et al. (2018) designed a rule-based classifier to identify central events according to event coreference relations. Liu et al. (2018) utilized a kernel-based neural model that captured semantic relations between discourse units for event salience identification. However, they did not consider the document-level event factuality. To our best knowledge, this paper is the first work on document-level event factuality identification task.

Previous studies (He et al., 2017; Rudinger et al., 2018; Qian et al., 2018) have tried neural network models on sentence-level factuality identification. Recent research has shown that neural networks with multi-level attention can extract meaningful information from heterogeneous input and improve the performance of NLP tasks, e.g., discourse relation (Liu and Li, 2016), relation classification (Wang et al., 2016), and question answering (Yu et al., 2017). Moreover, to improve the robustness of neural networks, related studies considered adversarial perturbation and training on text classification (Miyato et al., 2016) and relation extraction (Wu et al., 2017). This paper is in line in proposing an adversarial neural network with both intra- and inter-sequence attention.

6 Conclusion

We investigated document-level event factuality identification task by constructing a corpus annotated with document- and sentence-level event factuality based on both English and Chinese texts. To identify document-level event factuality, we proposed an LSTM neural network with both intra- and inter-sequence attention, and consider adversarial training to improve the robustness. Experimental results showed that document-level event identification on our DLEF corpus is useful, and our adversarial training model outperforms several baselines. To our knowledge, this is the first paper for the document-level event factuality identification.

In the future work, we will consider to detect events and their sentence-level and document-level factuality with a joint framework, and we will also continue to expand the scale of our DLEF corpus.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments on this paper. This work was partially supported by national Natural Science Foundation of China (NSFC) via Grant Nos. 61836007, 6177235, 61773276, 61673290.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James R. Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3528–3539.

Leo Born, Mohsen Mesgar, and Michael Strube. 2017. Using a graph-based coherence model in document-level machine translation. In *Proceedings of DiscoMT@EMNLP 2017*, pages 26–35.

Yuan Cao, Qiaoming Zhu, and Peifeng Li. 2013. The construction of chinese event factuality corpus. *Journal of Chinese Information Processing*, 27(6):38–45.

Prafulla Kumar Choubey, Kaushik Raju, and Ruihong Huang. 2018. Identifying the most dominant event in a news article by mining event coreference relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 340–345.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, 20(1):37–46.

Mona T. Diab, Lori S. Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop, LAW 2009*, pages 68–73.

Zi-Yi Dou. 2017. Capturing user and product information for document level sentiment analysis with deep memory network. In *Proceedings of EMNLP 2017*, pages 521–526.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

Tianxiong He, Peifeng Li, and Qiaoming Zhu. 2017. Identifying chinese event factuality with convolutional neural networks. In *Proceedings of CLSW 2017*, pages 284–292.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of EMNLP 2015*, pages 1643–1648.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of EMNLP 2016*, pages 1224–1233.

Zhengzhong Liu, Chenyan Xiong, Teruko Mitamura, and Eduard H. Hovy. 2018. Automatic event salience identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1226–1236.

Amnon Lotan, Asher Stern, and Ido Dagan. 2013. Truthteller: Annotating predicate truth. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 752–757.

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

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS 2013*, pages 3111–3119.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2016. Virtual adversarial training for semi-supervised text classification. *CoRR*, abs/1605.07725.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona T. Diab. 2010. Automatic committed belief tagging. In *COLING 2010*, pages 1014–1022.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering, Papers from 2003 AAAI Spring Symposium*, pages 28–34, Stanford University, Stanford, CA, USA.

Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Event factuality identification via generative adversarial networks with auxiliary classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4293–4300.

Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2015. A two-step approach for event factuality identification. In *IALP 2015*, pages 103–106.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of EACL 2017*, pages 1171–1182.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. 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 731–744. Association for Computational Linguistics.

Roser Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University, Waltham, MA, USA.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of ACL 2017*, pages 352–357.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of ACL 2016*, pages 1298–1307.

Gregory Werner, Vinodkumar Prabhakaran, Mona Diab, and Owen Rambow. 2015. Committed belief tagging on the factbank and lu corpora: A comparative study. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 32–40, Denver, Colorado. Association for Computational Linguistics.

Yi Wu, David Bamman, and Stuart J. Russell. 2017. Adversarial training for relation extraction. In *Proceedings of EMNLP 2017*, pages 1778–1783.

Jiacheng Xu, Danlu Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Cached long short-term memory neural networks for document-level sentiment classification. In *Proceedings of EMNLP 2016*, pages 1660–1669.

Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. 2017. Multi-level attention networks for visual question answering. In *Proceedings of CVPR 2017*, pages 4187–4195.

Bowei Zou, Qiaoming Zhu, and Guodong Zhou. 2015. Negation and speculation identification in chinese language. In *Proceedings of ACL-IJCNLP 2015*, pages 656–665.