Event Factuality Identification via Generative Adversarial Networks with Auxiliary Classification

Zhong Qian[1], **Peifeng Li**[1], **Yue Zhang**[2], **Guodong Zhou**[1], **Qiaoming Zhu**[1] *
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
[2]Singapore University of Technology and Design, Singapore
qianzhongqz@163.com, pfli@suda.edu.cn, yue_zhang@sutd.edu.sg, {gdzhou, qmzhu}@suda.edu.cn

Abstract

Event factuality identification is an important semantic task in NLP. Traditional research heavily relies on annotated texts. This paper proposes a two-step framework, first extracting essential factors related with event factuality from raw texts as the input, and then identifying the factuality of events via a Generative Adversarial Network with Auxiliary Classification (AC-GAN). The use of AC-GAN allows the model to learn more syntactic information and address the imbalance among factuality values. Experimental results on FactBank show that our method significantly outperforms several state-of-the-art baselines, particularly on events with embedded sources, speculative and negative factuality values.

1 Introduction

Event factuality expresses the commitment of relevant sources towards the factual nature of events, conveying whether an event is characterized as a fact, a possibility, or an impossible situation. Event factuality identification is useful for deep NLP applications, such as opinion detection, temporal ordering of events, textual entailment, and rumor identification. In principle, event factuality is related to various factors, including predicates, speculative and negative cues. Two examples are given below:

(S1) *McCulley, a famous economist,* ***doubts*** *that the tax rate will* ***increase*** *soon.*

(S2) *He* ***knows*** *they are not able to* ***go*** *to the village due to the flood.*

In this paper, events are marked in ***bold*** and sources are underlined in example sentences. In S1, the event ***increase*** is possible (PS+) according to the predicate ***doubts***, while in S2, the event ***go*** is impossible (CT-) due to the predicate ***knows*** and the negation word *not*. Table 1 shows that factuality can be characterized by the combination of epistemic modality and polarity. Modality conveys the certainty degree of events, such as *certain* (CT), *probable* (PR), and *possible* (PS), while polarity expresses whether the event has happened, including *positive* (+) and *negative* (-). In addition, U/u means *underspecified*. Some combined values are not applicable (NA) grammatically (e.g., PRu, PSu, and U+/- [Saurí, 2008; Saurí and Pustejovsky, 2012]), and are not considered.

	+ (positive)	- (negative)	u (underspecified)
CT(certain)	CT+	CT-	CTu
PR(probable)	PR+	PR-	(NA)
PS(possible)	PS+	PS-	(NA)
U(underspecified)	(NA)	(NA)	Uu

Table 1: Various values of event factuality.

Previous methods employed rules [Saurí, 2008; Saurí and Pustejovsky, 2012], machine learning models [de Marneffe *et al.*, 2012; Lee *et al.*, 2015], or a combination of the two [Qian *et al.*, 2015; Stanovsky *et al.*, 2017]. These approaches rely on annotated information, such as predicates, sources, speculative and negative cues, which are limited and can be costly to obtain. In addition, the performance of previous work is imbalanced on different values of event factuality. On one hand, the performance of speculative values is low due to their scarcity (4.36%) (e.g., [Saurí and Pustejovsky, 2012] achieved much lower performance of PR+ and PS+ compared to that of CT+ and Uu (F1: 45.71, 59.46 vs 84.85, 74.61) on Aquaint TimeML in FactBank using annotated data). On the other hand, events embedded in other predicates and sources (i.e., embedded events (31.04%)), which can have complicated syntactic structures (e.g., the event ***increase*** is embedded in the predicate ***doubts*** in S1), gave lower macro-averaged F1 than that of those events only with *AUTHOR* as sources (F1 67 vs 73 [Saurí and Pustejovsky, 2012]).

This paper proposes a two-step supervised framework to identify event factuality in raw texts, in which we first extract basic factors related with factuality (i.e., events, predicates, sources, and cues), and then utilize an **G**enerative **A**dversarial **N**etwork with **A**uxiliary **C**lassification for **E**vent **F**actuality identification (EF-AC-GAN). To automatically produce more syntactic paths and improve the performance of embedded events with complicated syntactic structures, we utilize the generator in EF-AC-GAN to generate syntactic paths that are

close to the distribution of real ones. In addition, to address the imbalance among factuality values and improve the performance of factuality values that are in minority (i.e., CT-, PR+, and PS+), we design two auxiliary classification tasks in EF-AC-GAN, one output deciding whether the event is Uu or Non-Uu, and the other indicating whether the event is modified by a cue and determining the modality and polarity of the event. **Shortest Dependency Paths (SDP)** are the main syntactic features for EF-AC-GAN.

Experimental results on a standard benchmark show that EF-AC-GAN outperforms the baselines significantly, especially on embedded events, speculative and negative factuality values. The code of this paper is released at `https://github.com/qz011/ef_ac_gan`.

2 Basic Factor Extraction

This section introduces the basic factors related with factuality, namely events, SIPs, sources and cues, and presents the methods to identify them.

Events in FactBank are defined by TimeML [Pustejovsky *et al.*, 2003]. For event detection we utilize the maximum entropy classification model [Chambers, 2013].

Source Introducing Predicates (SIPs) are events that can not only introduce additional sources but influence event factuality. For example, in S1 above, the SIP ***doubts*** introduces *McCulley* as a new source, and *McCulley* evaluates the event ***increase*** as PS+ according to ***doubts***. We consider both **lexical level features** and **sentence level features** to detect SIPs. Similar to event detection, we consider the *token*, *part-of-speech (POS)* and *hypernym* in WordNet[1] of the event as **lexical level features** and concatenate them into vector $\boldsymbol{l}$.

A SIP has at least one embedded event [Saurí, 2008]. Hence, we propose **Pruned Sentence (PSen)** as a **sentence level feature** (i.e., a clause of an event is replaced by $\langle$*event*$\rangle$ if containing other embedded events; Nouns, pronouns and the current event are unchanged, while other tokens are replaced by $\langle O \rangle$). S4 is the PSen of the event ***says*** in S3.

(S3) *Tom, who is the secretary of the manager,* ***says*** *the manager will* ***attend*** *a meeting later .*

(S4) *Tom* $\langle O \rangle$ *who* $\langle O \rangle$ $\langle O \rangle$ *secretary* $\langle O \rangle$ $\langle O \rangle$ *manager* $\langle O \rangle$ ***says*** $\langle$*event*$\rangle$ $\langle O \rangle$

Because PSen has a simplified structure, we extract sentence level features $\boldsymbol{c}$ from PSen $\boldsymbol{X}_0 \in \mathbb{R}^{d_0 \times n}$ through an attention-based CNN instead of an RNN. The CNN and its objective function is defined as follows, where $\oplus$ is the concatenation operator, and $\boldsymbol{W}_c, \boldsymbol{b}_c, \boldsymbol{v}_c, \boldsymbol{W}_{s0}$ and $\boldsymbol{b}_{s0}$ are model parameters:

$$\boldsymbol{Y}_0 = \boldsymbol{W}_c \boldsymbol{X}_0 + \boldsymbol{b}_c \tag{1}$$

$$\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{v}_c^T \tanh(\boldsymbol{Y}_0)) \tag{2}$$

$$\boldsymbol{c} = \tanh\left(\boldsymbol{Y}_0 \boldsymbol{\alpha}^T\right) \tag{3}$$

$$\boldsymbol{f} = \boldsymbol{l} \oplus \boldsymbol{c} \tag{4}$$

$$\boldsymbol{o} = \mathrm{softmax}(\boldsymbol{W}_{s0} \boldsymbol{f} + \boldsymbol{b}_{s0}) \tag{5}$$

$$J(\theta) = -\frac{1}{m} \sum_{i=0}^{m-1} \log p(y^{(i)} | x^{(i)}, \theta) + \frac{\lambda}{2} \|\theta\|^2 \tag{6}$$

[1] http://wordnet.princeton.edu

A **Relevant Source** is the participant of an event holding a specific stance with regard to the factuality. *AUTHOR* is always the source by default, and further sources (e.g. *McCulley* in S1) are represented in chain form [Saurí and Pustejovsky, 2012]: *McCulley_AUTHOR*, which means that we know about *McCulley*'s perspective only according to *AUTHOR* and *McCulley* is an **Embedded Source** in *AUTHOR*. The grammatical subjects of SIPs are chosen as the introduced **new sources**. After we have identified events and SIPs, we employ the recursive algorithm of [Saurí, 2008] to identify relevant sources.

Cues include speculative and negative words. PR/PS events are modified by PR/PS cues, while events can be negated by negative (NEG) cues (e.g., the factuality of event ***go*** in S2 is CT- due to the NEG cue *not*). [Velldal *et al.*, 2012] concluded that lexical sequence-oriented n-gram features can achieve excellent results on cue detection. Hence, we employ the *lexical features* developed by [Velldal *et al.*, 2012] to classify each token as *PR/PS/NEG cue*, or *not cue*.

3 AC-GAN for Event Factuality Identification

3.1 Overall Structure

GAN [Goodfellow *et al.*, 2014] involves a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$, which are trained in opposition to one another. Due to the game-theoretic formulation, $\mathcal{G}$ produces samples that looks real, and $\mathcal{D}$ discriminates between generated samples and real ones. On top of GAN, AC-GAN consider auxiliary classification for class labels. In AC-GAN, $\mathcal{G}$ synthesizes samples conditioned on class labels, and $\mathcal{D}$ discriminates not only real and generated samples but assigns class labels for them. The objective function has two parts, i.e., the log-likelihood of the real samples L_S and the correct class L_C:

$$\begin{aligned} L_S =& \mathbb{E}[\log P(S = real | X_{real})] + \\ & \mathbb{E}[\log P(S = generated | X_{generated})] \end{aligned} \tag{7}$$

$$\begin{aligned} L_C =& \mathbb{E}[\log P(C = c | X_{real})] + \\ & \mathbb{E}[\log P(C = c | X_{generated})] \end{aligned} \tag{8}$$

where $\mathcal{D}$ is trained to maximize $L_C + L_S$, and $\mathcal{G}$ is trained to maximize $L_C - L_S$.

We develop lexical and syntactic features according to the basic factors defined above, and consider **Shortest Dependency Paths (SDP)** from basic factors to events as syntactic features. As mentioned above, embedded and speculative/negative events are in the minority. In particular, events only with *AUTHOR* as sources are nearer to the root of the dependency tree, and their SDPs are simpler than those of embedded events. Hence, we design EF-AC-GAN for event factuality identification shown in Figure 1, where $\mathcal{G}$ generates SDPs conditioned on class labels, and $\mathcal{D}$ discriminates whether SDPs are generated and the auxiliary classification in $\mathcal{D}$ determines the class labels of events. We assign two class labels for each event: $label_u$, which represents whether the event is Uu, Non-Uu or other, and $label_{cue}$, which indicates whether the event is modified by a cue and further classifies Non-Uu events as CT+/-, PR+/-, PS+/-. Event factuality is determined directly by these two labels, which are demonstrated in detail below.

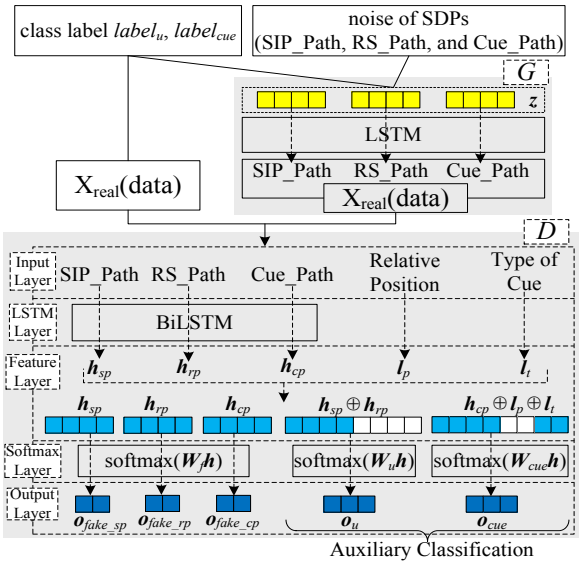

Figure 1: Architecture of AC-GAN for event factuality identification (EF-AC-GAN).

3.2 Features

Due to the success of dependency parse trees in previous work [Saurí, 2008; Saurí and Pustejovsky, 2012], we develop the following SDPs as syntactic features:

SIP_Path is extracted from the ancestor SIP that introduces the sources to the event.

Relevant Source Path (RS_Path) is extracted from the root of the dependency tree to the relevant sources of the event, and contains all the sources in the chain form.

Cue_Path is extracted from the cue to the event.

SIP_Path and RS_Path are used to judge whether the event is Un, Non-Uu or other. In addition to Cue_Path, we also consider the following cue-related lexical features to decide whether the event is governed by the cue:

Relative Position is the surface distance from the cue to the event, and is mapped into $\boldsymbol{l}_p$ with dimensions d_p.

Type of Cue includes PR, PS, and NEG, and is mapped into $\boldsymbol{l}_t$ with dimensions d_t.

If there is more than one cue in the sentence, we consider whether the current event is modified by each cue separately. An example sentence and its features for our EF-AC-GAN are shown in Figure 2.

3.3 The Discriminator

We utilize LSTM [Hochreiter and Schmidhuber, 1997] with hidden units n_{lstm} in $\mathcal{D}$ to model the sequences. Bidirectional LSTM (BiLSTM) is used to access the future as well as past context of SDPs, producing forward/backward hidden sequences $\overrightarrow{\boldsymbol{H}}/\overleftarrow{\boldsymbol{H}}$ and the output sequence $\boldsymbol{H}_p = \overrightarrow{\boldsymbol{H}} + \overleftarrow{\boldsymbol{H}}$. To capture the most important sources of information from the syntactic path, we adopt the attention model and obtain the output $\boldsymbol{h}_p$:

$$\boldsymbol{\alpha} = \text{softmax}(\boldsymbol{v}^T \tanh(\boldsymbol{H}_p)) \tag{9}$$

$$\boldsymbol{h}_p = \tanh(\boldsymbol{H}_p \boldsymbol{\alpha}^T) \tag{10}$$

where p=SIP_Path(sp), RS_Path(rp), Cue_Path(cp). We concatenate $\boldsymbol{h}_{sp}$ and $\boldsymbol{h}_{rp}$ into $\boldsymbol{f}_u$ to judge whether the event is Uu, Non-Uu or other:

$$\boldsymbol{f}_u = \boldsymbol{h}_{sp} \oplus \boldsymbol{h}_{rp} \tag{11}$$

where $\oplus$ is the concatenation operator. To determine whether the event is governed by a cue, we consider not only $\boldsymbol{h}_{cp}$ but the lexical features $\boldsymbol{l}_p$ (Relative Position) and $\boldsymbol{l}_t$ (Type of Cue) described above:

$$\boldsymbol{f}_{cue} = \boldsymbol{l}_p \oplus \boldsymbol{l}_t \oplus \boldsymbol{h}_{cp} \tag{12}$$

For the auxiliary classification of class labels, $\boldsymbol{f}_u$ and $\boldsymbol{f}_{cue}$ are fed into a softmax layer:

$$\boldsymbol{o}_u = \text{softmax}(\boldsymbol{W}_{s1} \boldsymbol{f}_u) \tag{13}$$

$$\boldsymbol{o}_{cue} = \text{softmax}(\boldsymbol{W}_{s2} \boldsymbol{f}_{cue}) \tag{14}$$

where $\boldsymbol{W}_{s1}$, $\boldsymbol{W}_{s2}$ are model parameters. $\boldsymbol{o}_u$ represents whether the event is Uu, Non-Uu or other ($label_u$), and $\boldsymbol{o}_{cue}$ is used to determine whether the event is governed by the cue ($label_{cue}$), and classify Non-Uu events as CT+/-, PR+/-, or PS+/-. We have two main reasons for the design of the two class labels. First, we can identify speculative and negative values (e.g., CT-, PR+/-, PS+/-) more precisely with the cues. Second, we can address imbalance among instances because speculative and negative values are typically in the minority.

In GAN it is essential to consider whether SIP_Path (sp), RS_Path (rp), and Cue_Path (cp) are generated:

$$\boldsymbol{o}_{fake\_p} = \text{softmax}(\boldsymbol{W}_f \boldsymbol{h}_p) \tag{15}$$

where $p = sp, rp$, and cp. The objective function of each output above is defined as:

$$L_D(j) = -\frac{1}{m} \sum_{i=0}^{m-1} \log p(y_j^{(i)} | x^{(i)}, \theta) \tag{16}$$

where $j = label_u, label_{cue}, fake\_sp, fake\_rp, fake\_cp$, and $y_j^{(i)}$ is the golden label of the corresponding output. The final objective function of $\mathcal{D}$ is :

$$L_D = \frac{1}{3}[L_D(fake\_sp) + L_D(fake\_rp) + L_D(fake\_cp)] + \frac{1}{2}[L_D(label_u) + L_D(label_{cue})] \tag{17}$$

3.4 The Generator

To produce more syntactic information and improve the performance of embedded and speculative/negative events, we generate the SDP $S = x_0, \cdots, x_t$ by feeding $label_u$ and $label_{cue}$ of events to the noise vector $\boldsymbol{z}$, i.e., $S = G(\boldsymbol{z}, label_u, label_{cue})$. LSTM is employed as the generator and generates a sequence of hidden states $\boldsymbol{h}_0, \ldots, \boldsymbol{h}_t$:

$$\boldsymbol{h}_t = \text{LSTM}(\boldsymbol{h}_{t-1}, \boldsymbol{x}_{t-1}) \tag{18}$$

where $\boldsymbol{x}_0$ is the input vector related to the noise vector $\boldsymbol{z}$ and the class labels:

$$\boldsymbol{x}_0 = \boldsymbol{z} \odot \boldsymbol{v}_u \odot \boldsymbol{v}_{cue} \tag{19}$$

where $\odot$ is the element-wise multiplication, and $\boldsymbol{v}_u$ and $\boldsymbol{v}_{cue}$ are the embeddings of $label_u$ and $label_{cue}$, respectively. SIP_Path, RS_Path, and Cue_Path have their respective noise vectors, which follow the normal distribution and are initialized randomly. Finally, a softmax layer maps the hidden states into the output token distribution:

$$p(x_t | x_{<t}) = \text{softmax}(\boldsymbol{W}_g \boldsymbol{h}_t) \tag{20}$$

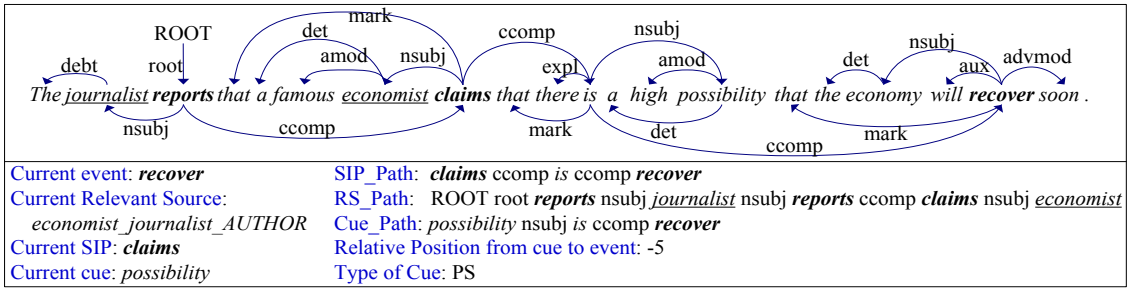

Figure 2: Example sentence and features of an event for EF-AC-GAN.

3.5 Training

When updating the discriminator $\mathcal{D}$ in each batch, we generate samples as the same number as the real samples, and the generated samples share the same Relative Position and Type of Cue with the corresponding real samples. $\mathcal{D}$(or $\mathcal{G}$) can become too strong and result in a gradient that cannot be used to improve $\mathcal{G}$(or $\mathcal{D}$). Therefore, we stop updating $\mathcal{D}$ when its training loss is less than 80% that of $\mathcal{G}$, and also take the corresponding strategy when $\mathcal{G}$ is too strong.

To prevent $\mathcal{G}$ over-training on $\mathcal{D}$, we utilize *Feature Matching* [Salimans *et al.*, 2016] to encourage greater variance in $\mathcal{G}$. Instead of maximizing the error of the output of $\mathcal{D}$, the new objective requires $\mathcal{G}$ to generate data that match the real data. More specifically, we train $\mathcal{G}$ to match the feature representations on the penultimate layer of $\mathcal{D}$. For each SDP, the objective is redefined as:

$$L_G(p) = \frac{1}{m}\sum_{i=0}^{m-1}(R(\boldsymbol{x}_p^{(i)}) - R(G(\boldsymbol{z}_p^{(i)})))^2 \qquad (21)$$

where R is the output of the BiLSTM layer before the final softmax layer, $\boldsymbol{x}_p$ are SDPs, and p=SIP_Path(sp), RS_Path(rp), Cue_Path(cp). The objective of $\mathcal{G}$ is defined as:

$$L_G = \frac{1}{3}[L_G(sp) + L_G(rp) + L_G(cp)] \qquad (22)$$

4 Experiments

4.1 Experimental Settings

We evaluate our models on FactBank [Saurí and Pustejovsky, 2009], which contains 3864 sentences and 13506 event factuality values. Table 2 presents the distribution of the values in FactBank. Following previous studies [Saurí and Pustejovsky, 2012; de Marneffe *et al.*, 2012], we only consider the five main categories of values, i.e., CT+, CT-, PR+, PS+, and Uu, which make up 99.05% of all the instances.

For fair comparison, we perform 10-fold cross-validation on FactBank. In addition to Precision, Recall, and F1-measure, macro- and micro-averaging are also applied to obtain the performance of all the factuality values. For SIP detection we set the dimensions of the *POS* and *hypernym* embeddings as 50 and $\lambda = 10^{-4}$. For event factuality identification, we set $n_{lstm} = 50$ and $d_p = d_t = 10$. We initialize word embeddings via Word2Vec [Mikolov *et al.*, 2013], setting the dimensions as $d_0 = 100$, and fine-tuning them during model training. SGD with momentum is applied to optimize our models.

Source	All	Author	Embed
CT+	7749/57.37%	5412/57.05%	2337/58.15%
CT-	433/3.21%	206/2.17%	227/5.65%
PR+	363/2.69%	108/1.14%	255/6.34%
PS+	226/1.67%	89/0.94%	137/3.41%
Uu+	4607/34.11%	3643/38.40%	964/23.99%
other	128/0.95%	29/0.31%	99/2.46%
Total	13506/100%	9487/100%	4019/100%

Table 2: Distribution of factuality values in FactBank.

	P(%)	R(%)	F1
Event	86.67	82.86	84.68
SIP	74.58	72.91	73.66
Source of events	80.70	77.44	78.99
Cue	64.78	70.13	67.05

Table 3: Performance of basic factor extraction.

4.2 Results on Basic Factor Extraction

Table 3 presents the performance of basic factor extraction. It is worth noting that a correctly identified SIP means that both the SIP and the new source introduced by it are correctly detected. For the SIP detection task, we also employ the model of [Chambers, 2013], obtaining F1=72.56, while our CNN achieves a higher F1=73.66 ($p < 0.05$ on *two-sample two-tailed t-test*). We argue that one SIP can determine **ALL** the sources of events embedded in it. Therefore, our CNN based on the PSen structures is effective.

4.3 Results on Event Factuality Identification

We employ the following baselines, whose features are developed according to the outputs of basic factor extraction, for fair comparison with our model:

Rules are developed by [Saurí, 2008] and [Saurí and Pustejovsky, 2012]. Instead of using annotated data directly, we obtain the performance of Rules using identified information according to basic factor extraction.

SVM is developed by [Saurí and Pustejovsky, 2012]. Besides, [de Marneffe *et al.*, 2012] and [Lee *et al.*, 2015] only considered *AUTHOR* as the source and employed traditional machine learning models. We re-implement them and obtain lower results than the SVM model on *AUTHOR* (macro-averaged F1 are 46.29 and 48.42, respectively).

ME+Rules: A two-step model combining a maximum entropy classification model and a simple rule-based model [Qian *et al.*, 2015].

Systems	Sources	CT+	CT-	PR+	PS+	Uu	Micro-A	Macro-A
Rules (Saurí et al.[2008; 2012])	All	61.83	54.52	20.75	39.89	26.08	50.71	40.62
	Author	64.83	48.83	13.35	31.93	26.17	53.49	37.02
	Embed	**53.19**	**61.94**	25.11	47.01	25.95	44.20	42.64
SVM (Saurí et al. [2012])	All	64.94	44.80	26.54	25.90	57.68	60.78	43.97
	Author	71.39	42.61	34.67	35.00	65.64	68.14	52.59
	Embed	50.78	45.60	28.58	27.66	25.45	43.33	35.40
ME_Rules ([Qian *et al.*, 2015])	All	61.55	43.52	17.65	41.49	53.58	56.89	43.56
	Author	67.75	44.21	11.55	40.99	60.95	63.80	43.75
	Embed	46.39	40.40	22.22	40.78	**30.46**	40.73	36.05
CNN-D	All	62.97	50.99	37.41	39.98	55.90	59.25	49.45
	Author	68.92	52.41	40.62	42.96	63.80	65.98	53.25
	Embed	50.53	49.00	36.99	42.93	18.57	43.26	39.18
L1	All	64.88	45.83	32.52	34.92	58.06	60.78	47.24
	Author	71.21	47.20	36.36	40.82	65.92	68.19	52.84
	Embed	50.76	44.22	31.59	33.67	24.88	43.22	36.62
BiLSTM	All	64.81	53.24	41.36	42.19	56.56	60.93	51.63
	Author	70.85	**56.88**	45.45	**46.63**	64.72	67.94	56.31
	Embed	51.89	49.24	38.89	40.08	18.76	44.37	39.31
EF-AC-GAN	All	**65.65**	**54.81**	**44.37**	**46.17**	**59.29**	**62.47**	**54.06**
	Author	**71.93**	56.45	**49.90**	45.35	**66.85**	**69.35**	**57.52**
	Embed	52.11	52.33	**41.47**	**47.99**	24.97	**46.14**	**43.56**

Table 4: Performances (F1-measures) of various models on event factuality identification.

CNN-D is a variant of the EF-AC-GAN whose BiLSTM in the discriminator $\mathcal{D}$ is replaced by the CNN with the hidden units $n_c = 50$.

L1 is a variant of the EF-AC-GAN, which concatenates $\boldsymbol{f}_u$ and $\boldsymbol{f}_{cue}$ in Equations (11) and (12) into ONE vector and consider only ONE class label of factuality.

BiLSTM utilizes BiLSTM in $\mathcal{D}$ of the EF-AC-GAN and does NOT consider generative model.

Table 4 shows the performance of various models on event factuality identification. Rules have low performance on Uu due to the error propagation from the basic factor extraction. SVM gets low results on CT-, PR+, and PS+ with minority events. For the model combining rules and machine learning classifiers, we employed the method of [Qian *et al.*, 2015] and achieved the performance between Rules and SVM (micro- and macro-averaged F1 are 56.89 and 43.56, respectively). Compared to the traditional models that utilized complicated algorithms or features (19 features in SVM and 15 features by [Qian *et al.*, 2015]), the features of our EF-AC-GAN are much fewer and easier to access.

Among all the models, EF-AC-GAN achieves the best results not only on CT+ and Uu but on CT-, PR+ and PS+ with All sources. The performance gaps among different factuality values illustrate that it is challenging to identify CT-, PR+ and PS+, which only cover 7.57% of all the instances. Compared to SVM, EF-AC-GAN improves the F1 of CT-, PR+ and PS+ by 10.01, 17.83 and 20.27, respectively. All the improvements are significant with $p < 0.05$ applying *two-sample two-tailed t-test* (the same below). The improved performance on CT+ and Uu means that EF-AC-GAN can effectively discriminate Non-Uu from Uu events, which can contribute to the outstanding results of CT+/-, PR+/- and PS+/-.

For both *AUTHOR* and Embedded Sources, EF-AC-GAN significantly outperforms the other methods on the micro-averaged and macro-averaged F1. The performance of CT-, PR+, PS+, and embedded sources shows that the generator $\mathcal{G}$ in EF-AC-GAN can produce meaningful dependency paths and EF-AC-GAN can benefit from $\mathcal{G}$ effectively. Besides, the results of events with embedded sources are lower than those with *AUTHOR* in EF-AC-GAN, mainly due to the complex syntactic structures of the events embedded in other sources. Similar to other models, the F1 of Uu for embedded sources is quite low (24.97) indicating that it is difficult to discriminate Uu from Non-Uu for embedded events.

To further verify the advantages of our model, we implement two variants of EF-AC-GAN, namely CNN-D and L1, as baselines. CNN ignores the context, while EF-AC-GAN considers both future and past context in syntactic paths and is superior to CNN-D. Compared with L1, both BiLSTM and EF-AC-GAN can obtain much better performance on CT-, PR+, and PS+, which can demonstrate that the design of the two class labels in auxiliary classification can address the problem of data imbalance. EF-AC-GAN also outperforms BiLSTM that does not consider the generative model, especially on PR+, PS+, and embedded sources, which illustrates that EF-AC-GAN can benefit from those generated samples. Moreover, if we only consider whether the syntactic paths are generated and neglect the auxiliary classification for the class labels of event factuality, we obtain poor performance (i.e., micro- and macro-averaged F1 are 17.85 and 20.79, respectively), which proves that class labels of event factuality are important supervised information for the training of EF-AC-GAN, and the auxiliary classification tasks are effective.

In conclusion, the experimental results of the neural network models above show that:

1) The design of the two class labels in the auxiliary classification task can improve the performance of speculative and negative factuality values (i.e., CT-, PR+, PS+) compared to the use of one class label.

2) The generator in EF-AC-GAN can produce useful and

Systems	Source	CT+	CT-	PR+	PS+	Uu	Micro-A	Macro-A
Rules	All	**86.69**	73.72	57.83	55.64	**76.13**	**81.52**	70.00
	Author	**88.32**	68.59	53.93	56.18	**81.18**	**84.56**	69.29
	Embed	**82.62**	**77.64**	60.44	54.16	**58.47**	**74.19**	66.67
BiLSTM	All	85.25	74.03	58.21	61.32	73.35	80.08	70.43
	Author	86.95	**73.09**	58.45	58.79	78.69	83.16	71.19
	Embed	81.48	74.78	57.80	63.92	48.67	72.67	65.33
EF-AC-GAN	All	85.46	**74.12**	**63.07**	**65.40**	75.10	80.81	**72.63**
	Author	87.21	72.49	**62.50**	**58.84**	79.96	83.76	**72.20**
	Embed	81.49	74.60	**64.47**	**66.66**	53.35	73.79	**68.11**

Table 5: Performances (F1-measures) on event factuality identification with annotated information.

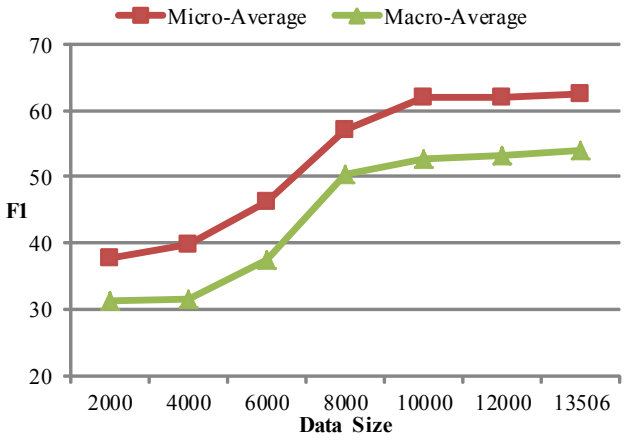

Figure 3: Performances of EF-AC-GAN with different data sizes of FactBank.

meaningful dependency paths to offer more syntactic information, and can improve the performance of the speculative/negative and embedded events in minority.

We investigate the relationship between the data size of FactBank and the performance of EF-AC-GAN, showing the results in Figure 3. Both the micro- and macro-averaged F1 of EF-AC-GAN become steady when we use more than 10000 samples. Particularly, in term of the performance of EF-AC-GAN on the whole FactBank, standard deviations of the micro- and macro-averaged F1 in 10 folds are less than 3, indicating that we obtain steady results.

To explore the upper bound of the performance of EF-AC-GAN, Table 5 shows the performance of Rules, BiLSTM and EF-AC-GAN with annotated information. Comparing Table 4 and 5, the micro- and macro-averaged F1 of Rules are improved by 30.81 and 29.38, respectively, and those of EF-AC-GAN by 18.34 and 18.57, respectively, which illustrates that the performance of Rules relies much more heavily on annotated information compared to EF-AC-GAN. Compared to Rules and BiLSTM, EF-AC-GAN achieves higher F1 of CT-, PR+, and PS+, indicating that our model is better at identifying speculative and negative factuality values. For embedded events, EF-AC-GAN is superior to BiLSTM on both micro- (73.79>72.67) and macro-averaged F1 (68.11>65.33). Therefore, the EF-AC-GAN model is more effective than BiLSTM, which does not consider GAN.

4.4 Error Analysis

We analyze the errors produced by the EF-AC-GAN model, which can be classified into the following main categories:

Incorrect relevant sources. Our model fails to identify correct relevant sources for events due to the error propagation from the basic factor extraction tasks (72.64%). These errors prove the significance of the SIP and relevant source detection task.

(S5) *Parliamentary president Rita Suessmuth **said** the People's Union correctly **listed** Jacques de Mathan as having **donated** about 274,500 marks in 1995, but **gave** a false address for him.*

(Event: ***donated***, Source: *Union_AUTHOR*, Uu)

For example, event ***donated*** is Uu according to *Union* in S8, but we fail to identify *Union* as the new source introduced by the SIP ***listed*** and miss the source *Union_AUTHOR* for ***donated***.

Incorrect Non-Uu/Uu. Whether an event is Non-Uu or Uu is mistakenly identified (EF-AC-GAN: 22.57%, BiLSTM: 25.08%). In S6, the BiLSTM model evaluates the event ***change*** as Uu incorrectly, and cannot assign PR+ to ***change*** even if determines that ***change*** is governed by the PR cue *appears*. In S7, BiLSTM classified the event ***take*** as Non-Uu and assign CT- to it according to the negative cue *not* incorrectly. While EF-AC-GAN identifies ***change*** as PR+ and ***take*** as Uu correctly, proving the usefulness of the auxiliary classification and the generated syntactic paths that can offer more syntactic information.

(S6) *Everyone **appears** to **believe** that somehow Cuba is going to **change**.*

(Event: ***change***, Source: *Everyone_AUTHOR*, PR+)

(S7) *He **added**, "This **has** nothing to do with Marty Ackerman and it is not **designed**, particularly, to **take** the company private."*

(Event: ***take***, Source: *He_AUTHOR*, Uu)

Incorrect Modality/Polarity. The modality or polarity of an event cannot be identified correctly because the model fails to determine whether the event is governed by a cue (EF-AC-GAN: 4.14%, BiLSTM: 4.39%).

(S7) *He **indicated** that some assets might be **sold** off to **service** the debt.*

(Event: ***service***, Source: *He_AUTHOR*, PS+)

(S8) *There was no **hint** of trouble in the last **conversation** between controllers and TWA pilot Steven Snyder.*

(Event: ***conversation***, Source: *AUTHOR*, CT+)

In S7, the event ***service*** is PS+ according to *He* due to the PS cue *might*. The BiLSTM model did not identify *might* for the event ***service***, while EF-AC-GAN gave the correct result. In S8, the event ***conversation*** is not governed by the negative cue *no* and is annotated as CT+. However, EF-AC-GAN regards *no* as the NEG cue of ***conversation*** and evaluates it as CT- mistakenly, indicating that EF-AC-GAN may overfit to the generated paths.

5 Related Work

Event factuality identification is a challenging task. Many studies limited the sources to the reader or *AUTHOR*. [Diab *et al.*, 2009] and [Prabhakaran *et al.*, 2010] presented a study of belief annotation and classified predicates into Committed Belief (CB), Non-CB or NA under a supervised framework. Recently, researchers presented scalable annotation schemes [Lee *et al.*, 2015; Stanovsky *et al.*, 2017], and developed new corpus [Soni *et al.*, 2014; Prabhakaran *et al.*, 2015]. To predict the factuality, [Lee *et al.*, 2015] employed lexical and dependency features, and [Stanovsky *et al.*, 2017] developed deep linguistic features.

FactBank [Saurí and Pustejovsky, 2009] considers both *AUTHOR* and embedded sources. Previous studies [Saurí, 2008; Saurí and Pustejovsky, 2012] proposed a complicated rule-based method to identify factuality of events on FactBank. [de Marneffe *et al.*, 2012] proposed a new annotation framework and identified the factuality of events in some sentences of FactBank. [Qian *et al.*, 2015] utilized a two-step framework combining machine learning and simple rule-based approaches.

Previous work also employed neural networks for factuality identification, but only considering the coarse-grained sentence-level factuality, e.g., uncertainty detection [Adel and Schütze, 2017] and modal sense classification [Marasović and Frank, 2016] for sentences, while this paper focuses on event factuality identification via AC-GAN instead of sentence factuality.

Syntactic paths [Roth and Lapata, 2016] and attention [Chen *et al.*, 2016; Zhou *et al.*, 2016] are helpful for many neural network-based NLP applications. Hence, we consider BiLSTM with attention to learn features from dependency paths in the discriminator of AC-GAN.

Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014] aim at fitting generative models to the distribution of realistic data, and have been proven successful in various AI and NLP applications, such as speech recognition [Chang and Scherer, 2017], image synthesis [Ghosh *et al.*, 2017], and text generation [Yu *et al.*, 2017]. Compared with GAN, AC-GAN [Odena *et al.*, 2017] considers both class labels of samples and the synthesis of sequences, and sets them as different outputs. AC-GAN is mainly applied on image synthesis [Dash *et al.*, 2017; Zhang *et al.*, 2017] instead of factuality-related NLP tasks. This paper applies AC-GAN to event factuality identification.

6 Conclusion

We presented a two-step framework for event factuality identification, which first extracts various basic factors, such as events, source introducing predicates, relevant sources and cues from the texts, and then employs EF-AC-GAN to identify event factuality. The design of auxiliary classification tasks in the discriminator can address the data imbalance among factuality values, and the generator can produce more syntactic information to improve the performance of speculative/negative and embedded events in minority. Experimental results show that EF-AC-GAN is superior to several state-of-the-art methods, especially on embedded events, and speculative and negative factuality values. To our best knowledge, this is the first work to apply AC-GAN to event factuality identification.

Acknowledgments

The authors would like to thank the four anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant Nos.61751206, 61772354 and 61773276, and was also supported by the Strategic Pioneer Research Projects of Defense Science and Technology under Grant No. 17-ZLXD-XX-02-06-02-04.

References

[Adel and Schütze, 2017] Heike Adel and Hinrich Schütze. Exploring different dimensions of attention for uncertainty detection. In *Proceedings of EACL 2017*, pages 22–34, 2017.

[Chambers, 2013] Nate Chambers. Navytime: Event and time ordering from raw text. In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013*, pages 73–77, 2013.

[Chang and Scherer, 2017] Jonathan Chang and Stefan Scherer. Learning representations of emotional speech with deep convolutional generative adversarial networks. In *Proceedings of ICASSP 2017*, pages 2746–2750, 2017.

[Chen *et al.*, 2016] Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. Neural sentiment classification with user and product attention. In *Proceedings of EMNLP 2016*, pages 1650–1659, 2016.

[Dash *et al.*, 2017] Ayushman Dash, John Cristian Borges Gamboa, Sheraz Ahmed, Marcus Liwicki, and Muhammad Zeshan Afzal. TAC-GAN - text conditioned auxiliary classifier generative adversarial network. *CoRR*, abs/1703.06412, 2017.

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

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

[Ghosh *et al.*, 2017] Arna Ghosh, Biswarup Bhattacharya, and Somnath Basu Roy Chowdhury. Handwriting profiling using generative adversarial networks. In *Proceedings of AAAI 2017*, pages 4927–4928, 2017.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS 2014*, pages 2672–2680, 2014.

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

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

[Marasović and Frank, 2016] Ana Marasović and Anette Frank. Multilingual modal sense classification using a convolutional neural network. In *Proceedings of the 1st Workshop on Representation Learning for NLP, Rep4NLP@ACL 2016*, pages 111–120, 2016.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013.*, pages 3111–3119, 2013.

[Odena *et al.*, 2017] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 2642–2651, 2017.

[Prabhakaran *et al.*, 2010] Vinodkumar Prabhakaran, Owen Rambow, and Mona T. Diab. Automatic committed belief tagging. In *Proceedings of COLING 2010*, pages 1014–1022, 2010.

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

[Pustejovsky *et al.*, 2003] James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 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, 2003.

[Qian *et al.*, 2015] Zhong Qian, Peifeng Li, and Qiaoming Zhu. A two-step approach for event factuality identification. In *2015 International Conference on Asian Language Processing, IALP 2015*, pages 103–106, 2015.

[Roth and Lapata, 2016] Michael Roth and Mirella Lapata. Neural semantic role labeling with dependency path embeddings. In *Proceedings of ACL 2016*, pages 1192–1202, 2016.

[Salimans *et al.*, 2016] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS 2016*, pages 2226–2234, 2016.

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

[Saurí and Pustejovsky, 2012] Roser Saurí and James Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):1–39, 2012.

[Saurí, 2008] Roser Saurí. *A Factuality Profiler for Eventualities in Text*. PhD thesis, Waltham, MA, USA, 2008.

[Soni *et al.*, 2014] Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. Modeling factuality judgments in social media text. In *Proceedings ACL 2014*, pages 415–420, 2014.

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

[Velldal *et al.*, 2012] Erik Velldal, Lilja Øvrelid, Jonathon Read, and Stephan Oepen. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38(2):369–410, 2012.

[Yu *et al.*, 2017] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of AAAI 2017*, pages 2852–2858, 2017.

[Zhang *et al.*, 2017] Lvmin Zhang, Yi Ji, and Xin Lin. Style transfer for anime sketches with enhanced residual u-net and auxiliary classifier GAN. *CoRR*, abs/1706.03319, 2017.

[Zhou *et al.*, 2016] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of ACL 2016*, pages 207–212, 2016.