Document-level Event Factuality Identification via Machine Reading Comprehension Frameworks with Transfer Learning

Zhong Qian[1*], **Heng Zhang**[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
qianzhong@suda.edu.cn, 20204227055@stu.suda.edu.cn,
{pfli, qmzhu, gdzhou}@suda.edu.cn

Abstract

Document-level Event Factuality Identification (DEFI) predicts the factuality of a specific event based on a document from which the event can be derived, which is a fundamental and crucial task in Natural Language Processing (NLP). However, most previous studies only considered sentence-level task and did not adopt document-level knowledge. Moreover, they modelled DEFI as a typical text classification task depending on annotated information heavily, and limited to the task-specific corpus only, which resulted in data scarcity. To tackle these issues, we propose a new framework formulating DEFI as Machine Reading Comprehension (MRC) tasks considering both Span-Extraction (Ext) and Multiple-Choice (Mch). Our model does not employ any other explicit annotated information, and utilizes Transfer Learning (TL) to extract knowledge from universal large-scale MRC corpora for cross-domain data augmentation. The empirical results on DLEFM corpus demonstrate that the proposed model outperforms several state-of-the-arts.

1 Introduction

This paper focuses on Document-level Event Factuality Identification (DEFI) task, which is defined as identifying the factuality of a specific event based on a document from which the event is derived. As a sub-task in Event Factuality Identification (EFI), different from Sentence-level Event Factuality Identification (SEFI) focusing on just a single sentence, DEFI requires comprehensive understanding documents with regard to events.

Figure 1 illustrates an example of document-level event factuality. The current event is E1, i.e., "Barack Obama joins Joe Biden's cabinet", and the sentences S2-S6 contain the event mentions referring to E1. From Figure 1 we can know that: 1) Factuality of event mentions vary among sentences. S4, S5 and S6 negate E1 by the negative cues "not" and "denied", and commit to E1 as "certain negative/CT-". But some other sentences express different factuality with regard to E1. S2 evaluates E1 as "possible positive/PS+" according to the speculative cue "speculation", and S3 commits to E1 as "Underspecified/Uu" since the event mention is in the clause led by "if". However, the document-level factuality of E1 is unique, i.e., CT-. 2) In addition to E1, there are irrelevant mentions of other events in the document as well, e.g., the PS+ event "Susan Rice and Michèle Flournoy were considering for key posts under Biden"in S6, and the CT+ event "Barack Obama was elected president in 2008" in S7 , which may cause E1 to be identified as PS+ or CT+ falsely.

Event (E1): Barack Obama joins Joe Biden's cabinet.

Document: **(S1)** "I will help him in any ways that I can," Obama, said of his former vice president, in a new interview with CBS Sunday Morning's Gayle King. **(S2)** With the victory of former US Vice President Biden in the presidential election, there are speculations about **whether** Obama will return to the White House and serve in Biden's cabinet. **(S3)** King also asked **if** he would join Biden's cabinet. **(S4)** "I'm **not planning** to suddenly work on the White House staff **or** something," said Obama. **(S5)** He also jokingly responded, "There are **probably** some things I would **not** be doing, because Michelle would leave me. ... ..."**(S6)** Although Obama **denied** that he would take a position in Biden's cabinet, Susan Rice and Michèle Flournoy were among Obama administration veterans reportedly being **probably** considered for key posts under Biden. ... ... **(S7)** When Barack Obama was elected president in 2008, he became the first African American to hold the office.

Document-level Factuality: certain negative / CT-

Figure 1: An example of document-level event factuality. **Speculative cues** are **blue**, and **negative cues** are **red**.

Currently, most EFI studies limited to SEFI (Saurí and Pustejovsky, 2012; Rudinger et al., 2018; Qian et al., 2018a; Veyseh et al., 2019). While DEFI is still in its early stage, and previous work (Qian et al., 2019; Huang et al., 2019; Cao et al., 2021) usually regarded DEFI as a typical text classification task. Currently, DEFI is mainly faced

Proceedings of the 29th International Conference on Computational Linguistics, pages 2622–2632
October 12–17, 2022.

	+	-	u
CT/一定	CT+/一定发生	CT-/一定不发生	CTu/知道是否发生
PS/可能	PS+/可能发生	PS-/可能不发生	(NA)
U/未指定	(NA)	(NA)	Uu/未指定

Table 1: Event factuality values in English and Chinese.

with these limitations, i.e., 1) Most previous work on EFI only considered sentence-level task, i.e., SEFI, which means DEFI catches much less attention and has been in the preliminary phase; 2) Related models on DEFI required various annotated information, e.g., event triggers, speculative and negative cues, and cannot be applied to real world directly; 3) The performance of DEFI is limited by the scale of dataset, i.e., DLEF (Qian et al., 2019), the only available DEFI corpus, and related work did not adopt any form of data augmentation.

To address the above issues, we propose a new end-to-end paradigm for DEFI, i.e., Document-level Event Factuality identification via Machine Reading Comprehension Frameworks with Transfer Learning (DEFI-MRC-TL), casting DEFI into MRC tasks, and considering both Span-Extraction MRC (Ext-MRC) and Multiple-Choice MRC (Mch-MRC). To address the problem of data insufficiency, we adopt Transfer Learning (TL) as *Cross-Domain Data Augmentation*, which learns information from large-scale source datasets and applies it to the target dataset. Therefore, our model is comprised of two sub-models that can be denoted as Ext-TL and Mch-TL, respectively. Our MRC formulation is mainly inspired by recent studies formulating NLP tasks into MRC problems (McCann et al., 2018; Li et al., 2019; Du and Cardie, 2020; Li et al., 2020b; Liu et al., 2020). To sum up, the major contributions of our paper can be summarized as follows:

1) We propose a new framework for DEFI by formulating it as MRC tasks, and we consider both span-extraction and multiple-choice MRC.

2) We consider a transfer learning mechanism that trains our MRC model on large-scale MRC corpora (e.g., SQuAD2.0, RACE) and fine-tunes on the target dataset (i.e., DLEFM). To the best of our knowledge, this is the first DEFI model considering both MRC framework and transfer learning.

3) We construct the first MRC-style DEFI corpus, i.e., DLEFM, annotating both events and document-level event factuality. Empirical evaluations on DLEFM can prove the generalization and effectiveness of our MRC framework for end-to-end DEFI.

2 Approach

This section introduces our DEFI-MRC-TL model, where **Overview** (§2.1) gives the definitions of DEFI task, even factuality values, and transfer learning used by our model, then §2.2 and §2.3 present the data formalization and detailed structure of Ext-TL and Mch-TL model, respectively.

2.1 Overview

Document-level Event Factuality Identification can be defined as to identify a label $y \in Y$ for the event $\mathbb{E}$ (usually a sentence) based on a document $\mathbb{D}$, where Y is the set of event factuality values defined in Table 1 (Qian et al., 2019). Therefore, one sample $\mathbb{S}$ can be denoted as $\mathbb{S} = (y, \mathbb{E}, \mathbb{D})$.

Event Factuality Values are composed of modality and polarity (Saurí, 2008; Saurí and Pustejovsky, 2012). Modality conveys the certainty degree of events, including these values:

- **Certain/CT/**一定(不): It is *certain* that the event happens / does not happen.
- **Probable/PR/**很可能(不): It is *probable* that the event happens / does not happen.
- **Possible/PS/**可能(不): It is *possible* that the event happens / does not happen.
- **Underspecified/Uu/**未指定: The degree of certainty of the event is *unknown* or *uncommitted*.

while polarity expresses whether the event happens by the following values.

- **Positive / + /**正极性**/**发生：It is *certain* / *probable* / *possible* that the event happens.
- **Negative / - /**负极性**/**不发生：It is *certain* / *probable* / *possible* that the event does not happen.
- **Underspecified / u /**未指定: The polarity of the event is *unknown* or *uncommitted*.

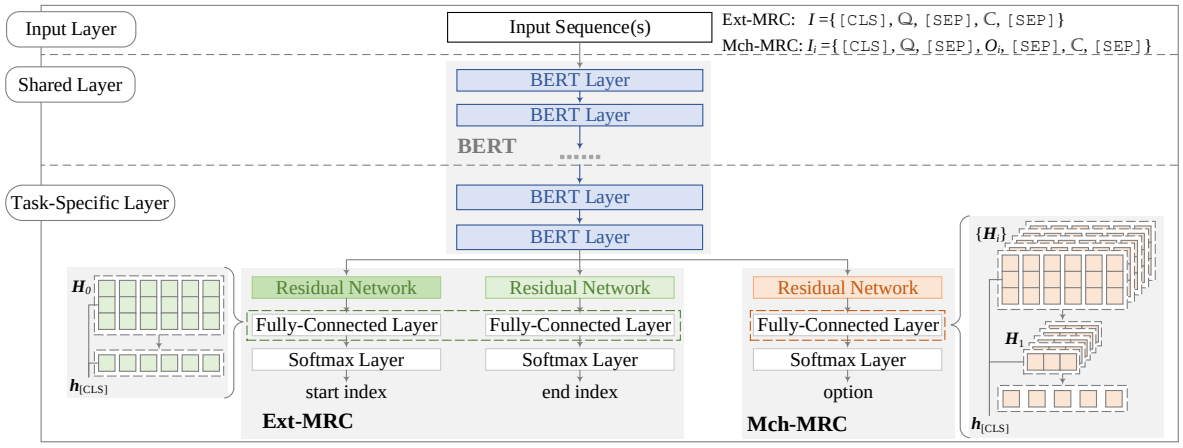

Figure 2: Overall architecture of our DEFI-MRC-TL models.

This paper uses the factuality values in Table 1 (Qian et al., 2019), where PSu and U+/- are not applicable (NA) semantically (Saurí and Pustejovsky, 2012). Compared with (Saurí and Pustejovsky, 2012), PS and PR are merged into PS due to similar semantics. Moreover, no event is annotated as CTu although applicable. Therefore, there are five applicable factuality values in DLEFM, i.e., Uu, PS-, CT-, PS+, CT+.

Transfer Learning is adopted as *Cross-Domain Data Augmentation* for DEFI, i.e., firstly learning knowledge from large-scale source datasets, and then applying it to the target datasets for fine-tuning. The architecture of DEFI-MRC-TL model is shown in Figure 2. With BERT (Devlin et al., 2019) as the backbone, our models consists of 1) **Shared Layer**, which is comprised of the first N_s layers of BERT, and optimized when training on the source datasets only; and 2) **Task-Specific Layers**, which contains the remaining layers of BERT and all the other networks in DEFI-MRC-TL, and is optimized when both training on the source datasets and fine-tuning on the target ones.

2.2 Ext-MRC Transfer Learning Model

We first give the data formalization of Ext-MRC model, and then describe the structure of Ext-TL.

Data Formalization for Ext-MRC. A sample $\mathbb{S}$ defined in §2.1 can be reformulated as a triple sample $\mathbb{S} = \{\mathbb{Q}, \mathbb{C}, \mathbb{A}\}$, where $\mathbb{Q} = \{q_0, \ldots, q_{|\mathbb{Q}|-1}\}$ is the *Question* that integrates the information of both event and candidate factuality values, $\mathbb{C} = \{c_0, \ldots, c_{|\mathbb{C}|-1}\}$ is the *Context* that refers to the document text $\mathbb{D}$ from which the event is derived, and $\mathbb{A} = \{a_0, \ldots, a_{|\mathbb{A}|-1}\} \subsetneq \mathbb{Q}$ is the *Answer* that is a sub-string of $\mathbb{Q}$, and each a_i belongs to $\mathbb{Q}$. To be specific, a $\mathbb{Q}$ in Ext-MRC can be denoted as:

- `What is the factuality of the event "`$\mathbb{E}$`", underspecified underspecified, possible negative, certain negative, possible positive, or certain positive?`

While for Chinese samples, $\mathbb{Q}$ is denoted as:

- `事件"`$\mathbb{E}$`"的事实性是未指定，可能不发生，一定不发生，可能发生，还是一定发生?`

where $\mathbb{E}$ is the event, and all the applicable factuality values are integrated into $\mathbb{Q}$. Therefore, Ext-MRC model extracts event factuality values from questions, rather than contexts.

Ext-TL Model. MRC-style input data is fed into MRC models defined below. To be in line with BERT, we concatenate the question $\mathbb{Q}$, the context $\mathbb{C}$, and the special token `[CLS]/[SEP]` as the input sequence I, which is fed into BERT:

$$I = \{[\texttt{CLS}], \mathbb{Q}, [\texttt{SEP}], \mathbb{C}, [\texttt{SEP}]\} \quad (1)$$

$$\boldsymbol{H}_0 = \text{BERT}(I) \quad (2)$$

where $\boldsymbol{H}_0 \in \mathbb{R}^{d \times N_L}$, d is the dimension of hidden states in BERT, and N_L is the length of I. During the phase of training on the source dataset, $\boldsymbol{H}_0$ is directly fed into softmax layer to compute the probability distributions of start and end indices.

In terms of fine-tuning on the target dataset, we have noticed several differences between typical MRC task (e.g., SQuAD) and our MRC framework for DEFI, i.e., 1) Instead of extracting answers from contexts $\mathbb{C}$, we extract answers from questions $\mathbb{Q}$,

and 2) Answers in typical MRC vary among documents. While in DEFI, the answer can only be one of the factuality values defined in Table 1.

Therefore, when fine-tuning the model, we utilize a variant of Residual Network (ResNet) (He et al., 2016) as the Adapter Network to encode $\boldsymbol{H}_0$ in order to bridge the information learned from source and target datasets. Actually, we use two ResNets to encode the information for the start and end indices of the answer separately:

$$\boldsymbol{H}_s = \text{ResNet}_s(\boldsymbol{H}_0) \quad (3)$$

$$\boldsymbol{H}_e = \text{ResNet}_e(\boldsymbol{H}_0) \quad (4)$$

where ResNet is composed of a stack of several residual layers, i.e., ResNet $= \{\text{ResLayer}\}$. For any input $\boldsymbol{U}_0$, each ResLayer is computed as:

$$\boldsymbol{U}_1 = \text{LN}(\text{Gelu}(\boldsymbol{W}_{r1}\boldsymbol{U}_0 + \boldsymbol{b}_{r1})) \quad (5)$$

$$\boldsymbol{U}_2 = \text{LN}(\text{Gelu}(\boldsymbol{W}_{r2}\boldsymbol{U}_1 + \boldsymbol{b}_{r2})) \quad (6)$$

$$\boldsymbol{U}_r = \boldsymbol{U}_2 + \boldsymbol{U}_0 \quad (7)$$

where $\boldsymbol{W}_{r1} \in \mathbb{R}^{h\times d}$, $\boldsymbol{b}_{r1} \in \mathbb{R}^h$, $\boldsymbol{W}_{r2} \in \mathbb{R}^{d\times h}$, and $\boldsymbol{b}_{r2} \in \mathbb{R}^d$ are parameters, h is the dimension of the hidden states in ResLayer, and LN is the Layer Normalization. Then, the probability of start and end index can be computed as:

$$\boldsymbol{p}_s = \text{softmax}(\boldsymbol{W}_s\boldsymbol{H}_s + \boldsymbol{b}_s) \quad (8)$$

$$\boldsymbol{p}_e = \text{softmax}(\boldsymbol{W}_e\boldsymbol{H}_e + \boldsymbol{b}_e) \quad (9)$$

Finally, the predicted start and end indices are obtained as:

$$\hat{i}_s = \{i|\text{argmax}(\boldsymbol{p}_s(i))\} \quad (10)$$

$$\hat{i}_e = \{i|\text{argmax}(\boldsymbol{p}_e(i))\} \quad (11)$$

where $i = 0, 1, \ldots, N_L$.

To ensure the generalization capability of Ext-MRC model on both source and target datasets, we do not discard the question $\mathbb{Q}$ or the context $\mathbb{C}$ when computing start and end index. The objective function $\mathcal{L}(\theta)$ of Ext-MRC is designed as:

$$\mathcal{L}_s(\theta) = -\frac{1}{N}\sum_{i=0}^{N-1}\log p_s(y_s^i|\theta) \quad (12)$$

$$\mathcal{L}_e(\theta) = -\frac{1}{N}\sum_{i=0}^{N-1}\log p_e(y_e^i|\theta) \quad (13)$$

$$\mathcal{L}(\theta) = \epsilon\mathcal{L}_s(\theta) + (1-\epsilon)\mathcal{L}_e(\theta) \quad (14)$$

where N is the number of samples, y_s^i and y_e^i are annotated start and end indices of the i-th sample, ϵ is the trade-off coefficient, and we set $\epsilon = 0.5$.

2.3 Mch-MRC Transfer Learning Model

Data Formalization for Mch-MRC. While in Mch-MRC, a sample $\mathbb{S}$ containing the event $\mathbb{E}$ can be represented as a quad sample, i.e., $\mathbb{S} = \{\mathbb{Q}, \mathbb{C}, \mathbb{O}, a\}$, where $\mathbb{Q} = \{q_0, \ldots, q_{|\mathbb{Q}|-1}\}$ is the *Question* that integrates the information of event, $\mathbb{C} = \{c_0, \ldots, c_{|\mathbb{C}|-1}\}$ is *Context* referring to the document $\mathbb{D}$ from which the event $\mathbb{E}$ is derived, $\mathbb{O} = \{O_0, \ldots, O_{|\mathbb{O}|-1}\}$ is the set of *Options*, and a is the *Answer* that is one of the options, i.e., $a \in \mathbb{O}$. Specifically, a $\mathbb{Q}$ in our Mch-MRC task can be denoted as:

- `What is the factuality of the event "`$\mathbb{E}$`"?`

For Chinese samples, a $\mathbb{Q}$ is denoted as:

- 事件"$\mathbb{E}$"的事实性是什么?

where $\mathbb{E}$ is the event, and $\mathbb{O}$ is the option set. For English samples, $\mathbb{O}$={Uu, PS-, CT-, PS+, CT+}, and for Chinese samples, $\mathbb{O}$={未指定/Uu, 可能不发生/PS-, 一定不发生/CT-, 可能发生/PS+, 一定发生/CT+} (Table 1).

Mch-TL Model. Similarly, our Mch-MRC model encodes each option O_i with the question $\mathbb{Q}$ and context $\mathbb{C}$ as previous work (Jin et al., 2020; Gu et al., 2021). Formally, given each option $O_i \in \mathbb{O}$, where $i = 0, \ldots, |\mathbb{O}| - 1$, we can obtain a set of input sequence $I = \{I_i\}_{i=0}^{|\mathbb{O}|-1}$ for BERT. Each I_i and its matrix representation are denoted as:

$$I_i = \{[\texttt{CLS}], \mathbb{Q}, [\texttt{SEP}], O_i, [\texttt{SEP}], \mathbb{C}, [\texttt{SEP}]\} \quad (15)$$

$$\boldsymbol{H}_i = \text{BERT}(I_i) \quad (16)$$

where BERT encodes $\{I_i\}$ as $\{\boldsymbol{H}_i\}$. For each $\boldsymbol{H}_i$, the state $\boldsymbol{h}^i_{[\texttt{CLS}]}$ of [CLS] is selected as the vector representation to make up $\boldsymbol{H}_1$:

$$\boldsymbol{H}_1 = \left\{\boldsymbol{h}^i_{[\texttt{CLS}]}\right\}_{i=0}^{|\mathbb{O}|-1} \quad (17)$$

where $\boldsymbol{H}_1 \in \mathbb{R}^{d\times N_o}$, and is fed into the softmax when training on the source datasets.

Similar to Ext-MRC model, during the phase of fine-tuning on the target dataset, we exploit residual networks as adapters to encode each $\boldsymbol{H}_i$ in Eq. 16 as well, since the text genre of the target dataset is different from that of the source dataset:

$$\tilde{\boldsymbol{H}}_i = \text{ResNet}(\boldsymbol{H}_i) \quad (18)$$

Then, we also use the hidden state of [CLS] $\tilde{\boldsymbol{h}}^i_{[\texttt{CLS}]}$ to denote each option O_i, and $\boldsymbol{H}_1$ is comprised of $\tilde{\boldsymbol{h}}^i_{[\texttt{CLS}]}$, where $i = 0, \ldots, |\mathbb{O}| - 1$ (Eq. 17).

Finally, we feed $\boldsymbol{H}_1$ into softmax layer to compute the probability distribution of each option O_i.

$$\boldsymbol{p}_o = \text{softmax}(\boldsymbol{W}_o \boldsymbol{H}_1 + \boldsymbol{b}_o) \quad (19)$$

where $\boldsymbol{W}_o \in \mathbb{R}^{1\times d}$ and $\boldsymbol{b}_o \in \mathbb{R}^1$ are parameters. The objective function is defined as:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=0}^{N-1} \log p_o(y_o^i|\theta) \quad (20)$$

where N is the number of samples, y_o^i is the annotated label of i-th sample.

3 Experimentation

This section first introduces **Experiment Settings** (§3.1), including target and source datasets, evaluation metrics, implementation details, and baselines. Then, experimental analysis focuses on the following aspects, i.e., 1) **Results and Analysis** (§3.2) discusses the comparisons of our model with baselines; 2) **Ablation** (§3.3) inspects the effectiveness of source datasets and networks in DEFI-MRC-TL model; 3) **Case Study** (§3.4) illustrates what our Ext-TL model can learn from several samples to reveal the internal mechanism of Ext-TL.

3.1 Experimental Settings

Target Dataset. DLEFM, whose formalization is defined in §2, is the target dataset to verify our DEFI-MRC-TL model, including two sub-corpora, i.e., English (DLEFM-E) and Chinese (DLEFM-C), whose statistics are shown in Table 2. The main differences between DLEFM and previous DLEF corpus lie in the following aspects:

Size. The sizes of Chinese sub-corpora are nearly the same in them. But DLEF contains more Chinese documents than English ones (4649 vs. 1727), which is less suitable to evaluate the performance on English texts. Hence, DLEFM annotates more English documents than DLEF.

Annotation. DLEF annotates not only event triggers and their sentence-level & document-level factuality, but also speculative and negative cues. And DLEFM further annotates ONE document-level event $\mathbb{E}$ based on event triggers in each document. Due to MRC framework, DLEFM corpus annotates questions for Ext-MRC (§2.2), while annotates questions and options for Mch-MRC (§2.3).

Task. Since DLEF can offer various annotated information, previous work (Qian et al., 2019; Huang et al., 2019) usually utilize annotated event triggers, sentence-level factuality, speculative and negative cues directly. While this paper aims to model DEFI as an end-to-end task, i.e., only considers questions, contexts, and options without any other explicit annotated information. Therefore, our model can apply to real-world applications directly.

	Uu	PS-	CT-	PS+	CT+	Total
English	38	46	671	594	3181	4530
Train	21	31	404	357	1905	2718
Develop	8	7	140	122	629	906
Test	9	8	127	115	647	906
Chinese	20	38	1358	860	2374	4650
Train	14	24	824	513	1415	2790
Develop	4	7	257	164	498	930
Test	2	7	277	183	461	930

Table 2: Dataset statistics of DLEFM.

Corpus	Language	Task	Used	Total
SQuAD2.0	English	Ext	20,000	130,217
NewsQA	English	Ext	20,000	103,960
RACE	English	Mch	20,000	87,866
CMRC2018	Chinese	Ext	10,111	10,111
C^3	Chinese	Mch	6,013	11,869

Table 3: MRC corpora used as source datasets, where Ext/Mch mean Ext-MRC/Mch-MRC, and "Used" & "Total" means used & total samples in training sets.

Source Datasets. For cross-domain data augmentation, the following corpora are selected as source datasets whose statistics are shown in Table 3: 1) **SQuAD2.0** (Rajpurkar et al., 2018) contains the existing SQuAD (Rajpurkar et al., 2016) collected from Wikipedia. 2) **NewsQA** (Trischler et al., 2017) collects news articles and highlights from CNN. 3) **RACE** (Lai et al., 2017) contains documents collected from the English exams for students. 4) **CMRC2018** (Cui et al., 2019) is composed of human-annotated questions on Chinese Wikipedia paragraphs. 5) $\mathbf{C}^3$ (Sun et al., 2020) is the first free-form Chinese multiple-choice MRC dataset sampled from Chinese examinations. According to the types, C^3 can be divided into C^3-Dialogue (C^3_D) and C^3-Mixed (C^3_M).

Evaluation. We focus on the performance of the three main categories of factuality values, i.e., CT-, PS+, CT+, since they occupy 98.15%/98.75% in DLEFM-E/DLEFM-C, respectively, and we do not consider the minor values (i.e., Uu and PS-) due to their small proportions as previous work (Saurí and Pustejovsky, 2012; Qian et al., 2018a, 2019). F1-score is used as the main evaluation metrics for each category of factuality values. To obtain the performance of all the values, macro- and micro-averaging F1 is also employed.

Models	CT-	PS+	CT+	Macro-Ave	Micro-Ave
LSTM-A	42.05 / 59.08	41.07 / 54.68	78.43 / 77.04	53.85 / 63.60	67.23 / 67.53
ULGN	45.87 / 61.07	43.05 / 49.58	81.87 / 76.49	56.93 / 62.38	70.55 / 66.27
BiDAF	50.66 / 67.84	49.75 / 60.94	81.06 / 81.21	60.49 / 69.99	72.86 / 73.23
BiDAF-TL	55.65 / 72.82	53.59 / 64.47	83.28 / 83.74	64.17 / 73.68	75.43 / 76.66
QANet	51.62 / 68.07	51.05 / 62.32	81.68 / 81.46	61.45 / 70.61	73.33 / 73.63
QANet-TL	55.81 / 72.91	53.25 / 65.05	83.38 / 83.39	64.14 / 73.78	75.61 / 76.58
BERT-B	54.22 / 71.11	53.11 / 63.11	81.30 / 81.55	62.88 / 71.92	74.18 / 74.84
Mch	54.09 / 71.86	52.39 / 63.66	82.01 / 81.48	62.83 / 72.33	74.42 / 75.16
Mch-TL	58.16 / 74.83	56.53 / 65.78	83.68 / 83.84	66.12 / 74.81	76.63 / 77.46
Ext	56.50 / 73.06	54.87 / 65.36	83.44 / 82.71	64.93 / 73.71	76.34 / 76.39
Ext-TL	**61.85 / 77.20**	**58.91 / 69.92**	**85.23 / 84.91**	**68.66 / 77.34**	**78.09 / 79.43**

Table 4: Performance of models on DEFI. Format: F1-scores for "English / Chinese" sub-corpus.

Models	Source Datasets	CT-	PS+	CT+	Macro-Ave	Micro-Ave
Ext-TL	SQuAD2.0	**61.85**	58.91	**85.23**	**68.66**	**78.09**
	NewsQA	60.68	**59.22**	84.85	68.25	77.87
Mch-TL	C^3-Dialogue	71.74	62.24	81.94	71.98	75.20
	C^3-Mixed	**74.83**	**65.78**	83.84	**74.81**	**77.46**
	C^3	74.04	64.03	**83.81**	73.96	76.62

Table 5: Performance of Ext-TL and Mch-TL on DEFI with difference source datasets. Format: F1-scores.

Implementation Details. BERT-Base version is chosen as the backbone of DEFI-MRC-TL model. We set 2 residual layers in the residual networks. The dimension of the hidden states of residual networks is set as 768. Adam (Kingma and Ba, 2015) is applied to optimize our model.

To take full advantage of knowledge learned from source datasets, we fine-tune as few BERT layers as possible on the target dataset. We observe that the performance on CT- and PS+ is very low (F1-score<10), or even no results can be obtained if only fine-tuning the last layer of BERT. Therefore, for those BERT-based transfer learning models, we fine-tune the last two layers of BERT (i.e., the shared layers contain $N_s = 10$ BERT layers) and all the layers of residual networks, and freeze other layers.

For all the models, we report the average evaluation metrics of the five rounds of experiments. For the training of TL models, each round adopts a subset that has a fixed size and is sampled from the source dataset randomly.

Baselines. For fair comparison with our DEFI-MRC-TL models, we employ the following models as baselines:

1) **LSTM-A** (Qian et al., 2019) uses dependency paths from cues to event triggers as syntactic features, and the sentences with event triggers as semantic features;

2) **ULGN** (Cao et al., 2021) is based on graph convolutional networks relying on event triggers;

3) **BiDAF** (Seo et al., 2017) employs bidirectional attention flow to obtain query-aware context representations;

4) **QANet** (Yu et al., 2018) adopts encoders consisting exclusively of convolution and self-attention;

5) **BERT-B** (BERT-Base) directly uses the event $\mathbb{E}$ and document $\mathbb{D}$ as the input sequence;

6) **Ext** & **Ext-TL** are Ext-MRC models, and **Mch** & **Mch-TL** are Mch-MRC models for DEFI. TL mean Transfer Learning is considered.

3.2 Results and Analysis

The comparisons of the performance of various models with our DEFI-MRC-TL model are summarized in Table 4. LSTM-A and ULGN get relatively lower results than other approaches, mainly due to the cascade errors, since we use raw texts as input for fair comparison, i.e., first extracting event triggers (F1=83.19%/79.65% for English/Chinese sub-corpus), speculative and negative cues (F1=68.80%/75.42%), then identifying document-level factuality.

BERT-B is a strong baseline compared with lightweighted models with simple structures, which can

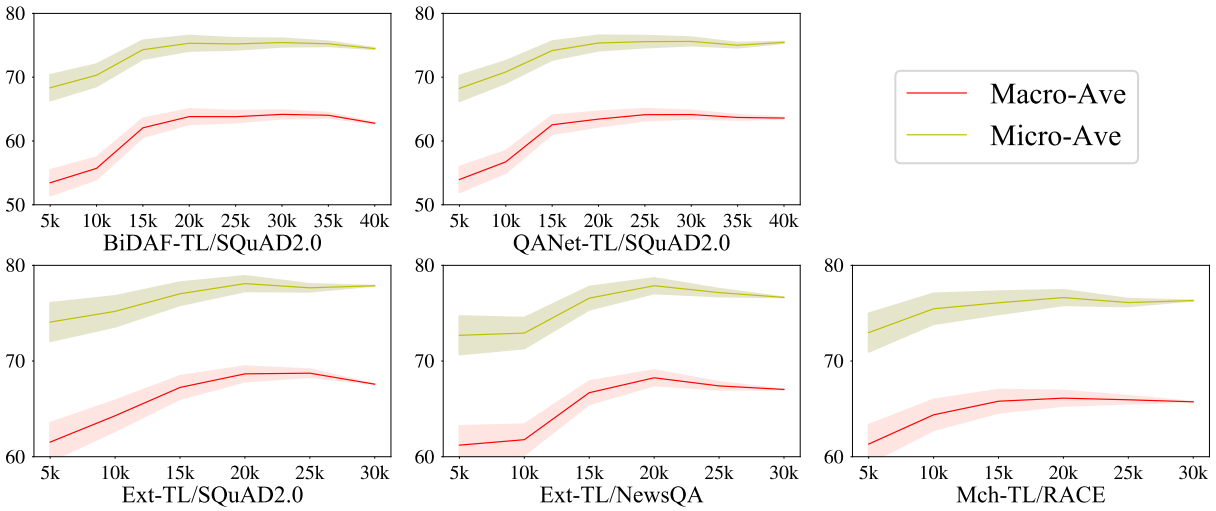

Figure 3: The test performance (F1-scores) w.r.t. the size of questions sampled from the source datasets for several transfer learning models on the English sub-corpus of DLEFM (i.e., DLEFM-E). For each sub-figure, the format of the label is "model/source dataset".

yield better results than LSTM-A and BiDAF when only fine-tuning on DLEFM, and achieve similar results with Mch.

Mch is slightly better than BERT-B. We argue that Mch can be regarded as a variant of text classification model integrating context knowledge with each option, whose input contains more information than BERT-B. Ext can get better results than BERT-B and Mch on both English and Chinese sub-corpora, which can validate the advantages of span-extraction MRC formulation for DEFI.

Table 4 demonstrates that all the TL models can achieve better performance than their corresponding original models that are only fine-tuned on target dataset, which can manifest the significance of cross-domain data augmentation using TL. For transfer learning models, BiDAF-TL and QANet-TL obtain lower performance than both Ext-TL and Mch-TL, since BiDAF and QANet have much fewer parameters than those BERT-based models. Ext-TL is superior to Mch-TL on DEFI. The main reason is that Ext-MRC models can extract meaningful and evidential texts for identifying document-level event factuality implicitly on both source and target dataset. Samples and analysis will be presented in §3.4 below.

Furthermore, Table 4 also shows that the performance on the Chinese sub-corpus is better than that on the English one, especial on CT- and PS+. The reason is that the factuality value distribution on the Chinese sub-corpus is more balanced than that on the English one.

3.3 Ablation

Text Genre of Source Dataset. We investigate the effects of source datasets with different text genres on the test performance on DLEFM, and present the performance of Ext-TL and Mch-TL w.r.t. difference source datasets in Table 5. Ext-TL achieves satisfactory performance employing either SQuAD2.0 or NewsQA as the source dataset, proving both of them can offer meaningful knowledge transferred to DEFI task. Moreover, using SQuAD2.0 obtains higher results than NewsQA, mainly owing to the high quality of Wikipedia articles with correct grammar and semantics.

Size of Source Dataset. Since the size of SQuAD2.0, NewsQA and RACE are quite large, we explore the relationship between the performance of DEFI on DLEFM-E and the scale of source dataset, and give the results in Figure 3. We can conclude that using too many samples from the source dataset can not lead to higher performance on the target dataset. For Ext-TL and Mch-TL, results can not be improved using more than 20k samples, or even degrades. It is mainly attributed to overfitting on source datasets when samples selected from them occupy too large quantities. Therefore, we adopt suitable amount of training samples from source datasets as shown in Table 3. For CMRC2018 and C^3, we leverage the whole training sets due to their limitations of the sizes.

In addition, we evaluate Mch-TL on Chinese sub-corpus. The source dataset C^3 consists of two sub-corpora, i.e, C^3_D and C^3_M. Different from all

Event (E1): Barack Obama joins Joe Biden's cabinet. [certain negative (CT-)/certain negative (CT-)]
Document: … …**whether** Obama will return to the White House and serve in Biden's cabinet. King also asked **if** he would join Biden's cabinet. "I'm **not planning** to suddenly work on the White House staff **or** something," said Obama. He also jokingly responded, "There are **probably** some things I would **not** be doing, ..."

Event (E2): New York City installs security barriers. [possible positive (PS+)/possible positive (PS+)]
Document: … … City Mayor Bill de Blasio announced the **plan** at a press conference held in Times Square on Tuesday. He said New York City **plans** to install 1,500 new security barriers in high-profile locations to guard against vehicle attacks and other terror-related incidents. ... ...

Event (E3): White House imposes new restrictions. [certain positive (CT+)/certain negative (CT-)]
Document: … …"Under **no** circumstances during a government shutdown will any government owned, rented, leased **or** chartered aircraft support any Congressional delegation, **without** the express written approval of the White House Chief of Staff," Russell Vought, … …

Event (E4): Hurricane Michael hits Florida. [possible positive (PS+)/positive, certain positive (non-applicable value, NA)]
Document: … … Hurricane Michael is **forecast** to strike Florida Panhandle in southeastern United States on Wednesday … … Hurricane Michael is currently centered about 360 miles (about 579 km) south of Panama City, Florida, and is moving north. … …

Figure 4: Spans extracted by Ext-TL are underlined. Questions are neglected when extracting spans in the documents to investigate the interpretability of Ext-TL. **Speculative cues** are **blue**, and **negative cues** are **red**. Format of labels: [Annotated/Predicted].

the other datasets considered in this paper, C^3_D is made up of dialogs. Moreover, the average context length of C^3_D (76.31) is obviously shorter than that of C^3_M (180.21) and Chinese sub-corpus in DLEFM (664.80). Hence, Mch-TL is not able to get higher results when employing the whole C^3_D, but can achieve the best performance on Chinese sub-corpus only using C^3_M as the source dataset.

Light-weighted solutions. Furtherly, in order to verify that TL models can benefit from adequate samples in the source datasets rather than only BERT-based models with large-scale and complicated structures, we also consider the light-weighted model, i.e., BiDAF and QANet based TL models, in Figure 3. QANet-TL outperforms BiDAF-TL, attributed to the more complicated attention in QANet. Compared with Ext-TL and Mch-TL, both BiDAF-TL and QANet-TL need more samples when training on the source dataset (i.e., SQuAD2.0) to reach the optimal performance, mainly due to the simpler structure of BiDAF and QANet than those BERT-based models. Both BiDAF-TL and QANet-TL are superior to BiDAF, BERT-B, and Mch, which can manifest the usefulness of transfer learning.

3.4 Case Study

As mentioned in §2, Ext-TL model discards neither questions nor contexts, and extracts answers (i.e., event factuality values) from the whole input sequence. To explore the interpretability of Ext-TL, we discard the questions in Equation 8 and 9, and extract text spans from the contexts.

As shown in Figure 4, Ext-TL identifies the correct values for events E1 and E2. In E1, the extracted span contains the mention "I'm not planning to suddenly work on the White House" that evaluates E1 as CT- according to the negative cue "not". While in E2, the extracted span commits to the event "New York City installs security barriers" as PS+ according to the speculative cue "plans".

However, E3 and E4 get wrong results. In term of E3, the extracted span contains no mention w.r.t. "White House imposes new restrictions", and another CT- event "Government aircraft support Congressional delegation" negated by "no" leads to the mistaken value of E3. While for E4, to identify it as PS+ correctly, we need to extract the event mention with speculative semantics, e.g., "Hurricane Michael is forecast to strike Florida Panhandle". But the actual span contains neither speculative nor negative semantics, extracting a non-applicable value that does not include "possible".

Therefore, these cases illustrate that correct identification of document-level event factuality relies on event mentions, speculative and negative information that governs the event.

4 Related Work

Event Factuality Identification started with SEFI, whose early work adopted rule-based models (Saurí and Pustejovsky, 2012), traditional machine learning models (de Marneffe et al., 2012; Lee et al., 2015), and hybrid models of them (Qian et al., 2018b). Recently, with the successful applications of neural networks in NLP, researchers focused on SEFI via neural networks, and captured information from sentences (He et al., 2017; Sheng et al.,

2019), sequential (Rudinger et al., 2018; Qian et al., 2018a) and graph-based (Veyseh et al., 2019) information from dependency trees, and furthurly more syntactic knowledge produced by generative adversarial networks (Qian et al., 2018a).

Compared with SEFI, DEFI remains at an initial stage, and limits to DLEF corpus (Qian et al., 2019). Previous studies (Qian et al., 2019; Huang et al., 2019) utilized multi-layer LSTM networks with attention to extract knowledge from dependency paths and sentences. Cao et al. (2021) learned local and global information of events by graph convolution networks. They relied on annotated information, e.g., event triggers, speculative and negative cues, and ignored data augmentation.

MRC/QA-Style Formulation has been widely utilized in NLP tasks over the past years, e.g, relation extraction (Li et al., 2019), named entity recognition (Li et al., 2020b), event extraction (Du and Cardie, 2020; Liu et al., 2020; Li et al., 2020a). To be specific, Li et al. (2020b) proposed a unified framework MRC handling both flat and nested NER tasks. Li et al. (2020a) designed MQAEE model casting event extraction into MRC problems to extract triggers and arguments successively. McCann et al. (2018) investigated MRC paradigms for ten tasks, including machine translation, sentiment analysis, semantic role labeling, etc.

Transfer Learning, or TL for short, is an effective technique for domain adaptation, and has achieved satisfactory results on various NLP applications, e.g., text classification (Houlsby et al., 2019; Stickland and Murray, 2019), sentiment classification (Fei and Li, 2020), neural machine translation (Aji et al., 2020), dialog system (Lin et al., 2020). Particularly, researchers also investigated TL for MRC/QA tasks. Kung et al. (2020) leveraged transfer learning to extract rationales through QA for zero-shot task transfer. Chung et al. (2018) explored both supervised and unsupervised transferability of knowledge learned among multiple-choice QA. Furthermore, some studies considered other TL paradigms, i.e., continual domain adaptation for domain drift in MRC (Su et al., 2020), and multi-task learning for QA (Wang et al., 2021; Lin et al., 2021) that is not dependent on specific domain of data.

5 Conclusion

This paper designs a novel framework formalizing Document-level Event Factuality Identification as MRC tasks, and considers both span-extraction and multiple-choice MRC. Furthermore, our model takes into account transfer learning as cross-domain data augmentation capturing extra knowledge from large-scale corpus in typical MRC. Experiments on DLEFM corpus demonstrate that our model can achieve state-of-the-art performance. In the future, we will explore cross-document event factuality identification and apply more effective data augmentation method.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62006167, 61836007 and 62276177), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of ACL 2020*, pages 7701–7710. Association for Computational Linguistics.

Pengfei Cao, Yubo Chen, Yuqing Yang, Kang Liu, and Jun Zhao. 2021. Uncertain local-to-global networks for document-level event factuality identification. In *Proceedings of EMNLP 2021*, pages 2636–2645. Association for Computational Linguistics.

Yu-An Chung, Hung-yi Lee, and James R. Glass. 2018. Supervised and unsupervised transfer learning for question answering. In *Proceedings of NAACL-HLT 2018*, pages 1585–1594. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5882–5888. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of EMNLP 2020*, pages 671–683. Association for Computational Linguistics.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of ACL 2020*, pages 5759–5771. Association for Computational Linguistics.

Yingjie Gu, Xiaoye Qu, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Xiaolin Gui. 2021. Read, retrospect, select: An MRC framework to short text entity linking. *CoRR*, abs/2101.02394.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR 2016*, pages 770–778. IEEE Computer Society.

Tianxiong He, Peifeng Li, and Qiaoming Zhu. 2017. Identifying chinese event factuality with convolutional neural networks. In *Proceedings of CLSW 2017*, volume 10709 of *Lecture Notes in Computer Science*, pages 284–292. Springer.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Rongtao Huang, Bowei Zou, Hongling Wang, Peifeng Li, and Guodong Zhou. 2019. Event factuality detection in discourse. In *Proceedings of NLPCC 2019*, volume 11839 of *Lecture Notes in Computer Science*, pages 404–414. Springer.

Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-Tür. 2020. MMM: multi-stage multi-task learning for multi-choice reading comprehension. In *Proceedings of AAAI 2020*, pages 8010–8017. AAAI Press.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Po-Nien Kung, Tse-Hsuan Yang, Yi-Cheng Chen, Sheng-Siang Yin, and Yun-Nung Chen. 2020. Zero-shot rationalization by multi-task transfer learning from question answering. In *Proceedings of EMNLP 2020*, pages 2187–2197. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of EMNLP 2017*, pages 785–794. Association for Computational Linguistics.

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. The Association for Computational Linguistics.

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

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020b. A unified MRC framework for named entity recognition. In *Proceedings of ACL 2020*, pages 5849–5859. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of ACL 2019*, pages 1340–1350. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of EMNLP 2020*, pages 3391–3405. Association for Computational Linguistics.

Zizheng Lin, Haowen Ke, Ngo-Yin Wong, Jiaxin Bai, Yangqiu Song, Huan Zhao, and Junpeng Ye. 2021. Multi-relational graph based heterogeneous multi-task learning in community question answering. In *Proceedings of CIKM 2021*.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of EMNLP 2020*, pages 1641–1651. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

Zhong Qian, Peifeng Li, Yue Zhang, Guodong Zhou, and Qiaoming Zhu. 2018a. Event factuality identification via generative adversarial networks with auxiliary classification. In *Proceedings of IJCAI 2018*, pages 4293–4300. ijcai.org.

Zhong Qian, Peifeng Li, Guodong Zhou, and Qiaoming Zhu. 2018b. Event factuality identification via hybrid neural networks. In *Proceedings of ICONIP 2018*, volume 11305 of *Lecture Notes in Computer Science*, pages 335–347. Springer.

Zhong Qian, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2019. Document-level event factuality identification via adversarial neural network. In *Proceedings of NAACL-HLT 2019*, pages 2799–2809. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of ACL 2018*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of EMNLP 2016*, pages 2383–2392. The Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of NAACL-HLT 2018*, 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. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Comput. Linguistics*, 38(2):261–299.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of ICLR 2017*. OpenReview.net.

Jiaxuan Sheng, Bowei Zou, Zhengxian Gong, Yu Hong, and Guodong Zhou. 2019. Chinese event factuality detection. In *Proceedings of NLPCC 2019*, volume 11839 of *Lecture Notes in Computer Science*, pages 486–496. Springer.

Asa Cooper Stickland and Iain Murray. 2019. BERT and pals: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

Lixin Su, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Continual domain adaptation for machine reading comprehension. In *Proceedings of CIKM 2020*, pages 1395–1404. ACM.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging chinese machine reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:141–155.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017*, pages 191–200. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of ACL 2019*, pages 4393–4399. Association for Computational Linguistics.

Zhiguo Wang, Patrick Ng, Ramesh Nallapati, and Bing Xiang. 2021. Retrieval, re-ranking and multi-task learning for knowledge-base question answering. In *Proceedings of EACL 2021*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of ICLR 2018*. OpenReview.net.