Journal of Artificial Intelligence Research 82 (2025) 1691-1710 Submitted 09/2024; published 03/2025

MRC and Transfer Learning Framework for Document-level Event Factuality Identification with Heterogeneous Spectral Attention Networks

Zhong Qian QIANZHONG@SUDA.EDU.CN
(Corresponding Author)
Peifeng Li PFLI@SUDA.EDU.CN
Qiaoming Zhu QMZHU@SUDA.EDU.CN
Guodong Zhou GDZHOU@SUDA.EDU.CN
School of Computer Science and Technology, Soochow University
1st Shizi Street, Suzhou, Jiangsu 215006, China

Abstract

This paper concentrates on Document-level Event Factuality Identification (DEFI) that predicts event factuality values from the viewpoint of the document. At present, the shortcomings of previous studies are multi-fold, including data limitation and scarcity, coarse-grained interpretability without span-level factuality clues, no unified model for different datasets. This paper is devoted to address the above problems by building unified Machine Reading Comprehension (MRC) frameworks comprised of both span-extraction and multiple-choice styles, which exploit Heterogeneous Spectral Attention Networks (HSAN) with spectral networks and hypergraph attention networks as the fine-grained encoders, especially for span-level encoding. Moreover, we integrate Transfer Learning (TL) as cross-domain data augmentation to learn more span-level information from classical MRC datasets by source and target adapters. Experimental performance on ExDLEF corpus, which contains both English and Chinese documents, shows that our span-extraction MRC model is superior to several state-of-the-art baselines, and proves the effectiveness of transfer learning under MRC paradigms.

1. Introduction

1.1 Introduction for Our Study

Event Factuality Identification (EFI) aims to predict the factual nature of an event according to the texts in which it occurs. Currently, EFI can be classified as two sub-tasks by the granularity of texts: Sentence-level Event Factuality Identification (SEFI) and Document-level Event Factuality Identification (DEFI). Evolving from SEFI, DEFI has become the mainstream, and also depends on sentence-level event mentions, just as illustrated in Figure 1. Given the event EE1 "Ukraine's drones attacked Zaporizhzhia Nuclear Power Plant", the following observations can be made:

All the sentences contain the event mentions, including triggers or relevant arguments. The triggers "attack" appear as verbal forms in S1, S4, S5 and nominal forms in S2, S3, S6, S7, while arguments "Ukraine", "ZNPP" occur in S1, S3, S4, S5 and S1-S6, respectively. But sentences hold different stances for the factuality of EE1. According to S1 and S3, this event has not effected by any speculative or negative semantics, and is evaluated as CT+,

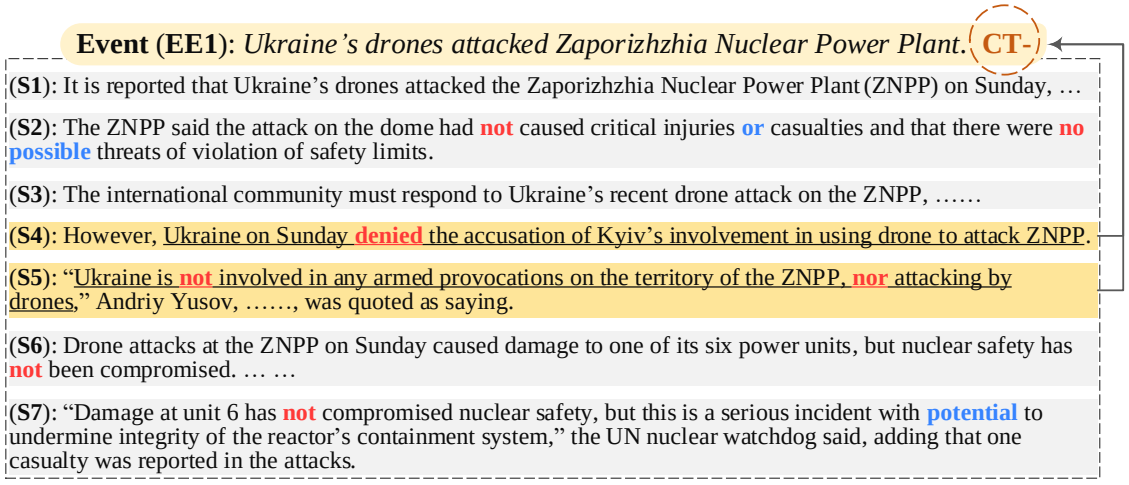

Figure 1: An example of document-level event factuality. **Speculative cues** are **blue**, and **negative cues** are **red**. Spans that contain the correct document-level event factuality are underlined.

which means it is certain that the event happens. However, the event has been negated by the negative cues "denied" in S4 and "not, nor" in S5, and is identified as CT- (it is certain that the event does not happen). It is worth noting that S4 and S5 contain the correct document-level semantics of EE1, whose the document-level factuality should be CT- as well. Besides, there are also other events, e.g., PS+ event "Integrity of the reactor's containment system is undermined", which is governed by speculative cue "potential" in S7. We should rule out the influence of these irrelevant events for EE1.

In the field of EFI, researchers have shifted to DEFI tasks, mainly involving with two definitions: 1) Sentence-level event mentions are known, and annotated by event triggers explicitly. Hence, methods are designed to capture syntactic and semantic features from interactions among event triggers, arguments, speculative/negative cues, e.g., attentional multi-layer RNN (Qian et al., 2019; Huang et al., 2019), local-to-global graph networks (Cao et al., 2021; Sheng et al., 2023), heterogeneous graph network (Zhang et al., 2022b, 2023), sentence-to-document inference network (Zhang et al., 2023). 2) Sentence-level event mentions are unknown, and only an event is given outside the document, which is the research focus of this paper. This task is still in the preliminary stage, and previous work is relatively limited, e.g., a reinforced hierarchical attention network for text selection (Qian et al., 2022a), and MRC framework with transfer learning to address data scarcity (Qian et al., 2022b).

On balance, DEFI is exposed to these problems: 1) *Data limitation.* Existing methods are confined in the domain-specific dataset, without data augmentation or external knowledge; 2) *Ignorance of span-level mentions.* Previous models employed coarse-grained fusion solution at sentence-level and document-level, ignoring the fine-grained span-level semantics for interpretability. 3) *Non-unified framework.* Related solutions cover various models, and

fail to construct a unified paradigm suitable for real-world application. These issues lead to low performance and robustness.

To address the aforementioned issues, we propose a novel model named Heterogeneous Spectral Attention Network (HSAN) as the end-to-end DEFI solution. In this model, we cast input texts into span-Extraction MRC (Ext-MRC) and Multiple-choice MRC (Mch-MRC) styles to build unified paradigm. Besides, this model can learn external knowledge from other general MRC datasets via Transfer Learning (TL) as cross-domain data augmentation.

1.2 Research Objectives

This paper is a significant extension of our conference version (Qian et al., 2022b), whose following aspects still need further improvement. One one hand, the backbone of the model is completely based on BERT, and does not utilize more fine-grained networks. On the other hand, it does not dive into span-level texts substantively for sentence-level or clause-level interpretability, leading to a coarse-grained transfer learning framework, although investigates spans when discarding questions in the case study. Thus, according to the above problems of previous work and our conference paper, our research objectives are comprised of these points: 1) We aim to learn more cross-domain semantics from other universal MRC datasets by transfer learning; 2) We are dedicated to build a unified MRC framework for DEFI; 3) We plan to design more fine-grained networks as adapters to capture high-level information for both source and target corpora. To sum up, innovations and contributions of this paper are summarized as follows:

1) We propose a unified MRC framework for DEFI task, where we consider both span-extraction and multiple-choice MRC paradigms. By applying MRC formulation, we can construct a unified solution architecture with specific input sequence that are appropriate in different datasets, and can overcome the problem of "non-unified framework".

2) We integrate transfer learning to learn meaningful semantics from several general MRC corpora as cross-domain data augmentation. Based on transfer learning mechanism, we can better capture sentence-level mentions that contain correct event factuality with help of more samples from external MRC datasets, and solve the problem of "data limitation".

3) We design Heterogeneous Spectral Attention Network (HSAN) for fine-grained encoding that also considers span-level information. According to the target adapter, we can furtherly extract span-level event mentions within sentences. This fine-grained encoding and span extraction mechanism can rule out irrelevant semantics and locate spans that contain correct event factuality. Therefore, we can address the problem of "ignorance of span-level mentions".

4) Extensive experiments on ExDLEF corpus containing both English and Chinese sub-corpora demonstrate that HSAN model performs better than other competitive baselines.

2. Related Work

This section presents a brief introduction to related work involving with document-level event factuality identification (Section 2.1), transfer learning (Section 2.2), and MRC-style tasks (Section 2.3).

2.1 Document-level Event Factuality Identification

As a basic research topic in the field of Natural Language Processing (NLP) and Artificial Intelligence (AI), Document-level Event Factuality Identification (DEFI) is firstly proposed and defined by (Qian et al., 2019), who also developed the first task-specific corpus named DLEF. Based on DLEF where event triggers are annotated explicitly, researchers designed various models started from multi-layer RNN with attentions working on syntactic paths and sentences (Qian et al., 2019; Huang et al., 2019). To list a few representative studies that bridge the gap between local and global factuality, Zhang et al. (2021) equipped attention networks with speculative and negative scopes to discriminate factual events from non-factual ones. Cao et al. (2021) applied local uncertainty and global structures on graph convolution network. Sheng et al. (2023) utilized uncertain relational hypergraph attention networks to globally consider local factuality features. Researchers also developed the family of heterogeneous graph networks, i.e., semantics-syntax fusion network (Zhang et al., 2022b) and graph-level contrastive learning (Zhang et al., 2023). In addition, Zhang et al. (2023) present a sentence-to-document inference network, which contains gated multi-layer interactive attentions.

Recently, researchers defined a novel DEFI task, where only the event (usually a sentence summarized from the document) and the document are given explicitly, without any other annotated information, nor sentence-level event triggers. The main difficulty lies in the extraction and aggregation of sentence-level mentions. Qian et al. (2022a) employed multi-granularity attention network with hierarchical reinforcement learning to select tokens and sentences. To learn more sentence-level semantics from external MRC datasets, Qian et al. (2022b) formulated DEFI as MRC tasks with transfer learning framework. This paper is an extended version of (Qian et al., 2022b), and dedicated to propose a more fine-grained model using heterogeneous spectral hypergraph attention networks integrated with both MRC and transfer learning.

2.2 Transfer Learning

Transfer Learning (TL) aims to improve the learning of the predictive model on the target domain using the knowledge in the source domain, and is a useful mechanism for cross-domain data augmentation, which is applied in MRC tasks. Wu et al. (2022) designed a multilingual MRC framework named siamese semantic disentanglement model to transfer semantic knowledge to the target language. Cao et al. (2023) considered sharing, teaching and aligning transferred knowledge for cross-lingual MRC.

Besides, TL can help to improve the performance of other event-related tasks. Zhang et al. (2022a) investigated transfer learning from semantic role labeling to event argument extraction facilitated by template-based slot querying strategies. Zhang et al. (2023) exploited cross-dataset transfer learning to extract overlap and specific knowledge for event argument extraction. He et al. (2024) introduced multi-granularity contrastive transfer learning for cross-lingual document-level event causality identification.

Based on these applications, this paper also applied TL that extracting fine-grained span-level information from typical MRC corpora to help to promote the results of DEFI.

Abbreviation	Full Name
DEFI	Document-level Event Factuality Identification
MRC	Machine Reading Comprehension
Ext-MRC	Span-Extraction-style Machine Reading Comprehension
Mch-MRC	Multiple-Choice-style Machine Reading Comprehension
TL	Transfer Learning

Table 1: Main terminologies used in this paper.

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

Table 2: Event factuality values in English and Chinese.

2.3 MRC-style Tasks

The main advantage of Machine Reading Comprehension (MRC) or Question Answering (QA) paradigms is re-framing NLP tasks as unified frameworks and capturing semantic level information. We mainly list several studies on event-related tasks. In the field of event extraction, Li et al. (2023) enabled a question generation model incorporating contextual information for event extraction, Liu et al. (2024) proposed question-context bridging to reconstruct the semantic relationship between templates and texts.

As for document-level event argument extraction, Liu et al. (2022a) applied back-translation based query generation and dependency-guided QA process for argument linking. Liu et al. (2022b) devised knowledge transfer and sample generation via MRC model. Uddin et al. (2024) presented a QA approach to extract document-level event-argument structures, and generated questions by templates for argument types.

Our MRC model is mainly motivated the above studies. By MRC formulation, we can not only build a unified framework for DEFI task, but learn more information from classical MRC datasets as cross-domain augmentation, especially those span-level semantics.

3. Task Definition

This section gives the formal definition of DEIF in detail, and then expounds span-extraction and multiple-choice MRC-style frameworks (Ext-MRC-style and Mch-MRC-style) for DEFI, respectively. This MRC solution for DEFI can build a unified paradigm with specific input, and can be applied to several datasets with different languages. Hence we can deal with the problem of "non-unified frameworks". For clear display and reference, we list abbreviations of related terminologies in Table 1.

3.1 Definition for DEFI

DEFI samples can be defined by a triple group $\{y, \mathbb{E}, \mathbb{D}\}$, where y is the annotated factuality of the event $\mathbb{E}$, and $\mathbb{D}$ is the document from which $\mathbb{E}$ is extracted. Then, DEFI requires to identify a factuality label $\hat{y}$ for the event $\mathbb{E}$ based on texts in $\mathbb{D}$.

To be more specific, y is defined as the combination of modality and polarity. Modality describes the certainty degree, including CerTain/一定（不）/CT and PoSsible/可能（不）/PS. Polarity denotes whether the event happens or not by Positive/正极性+/发生(+) and Negative/负极性/不发生(-). Therefore, the applicable factuality values can be textualized as:

- It is [**certain/possible**]$_{\text{modality}}$ that the event [**happens/does not happen**]$_{\text{polarity}}$.

Besides, Underspecified/未指定(U/u) is the reserved value representing that the factuality (either modality or polarity) is unknown or uncommitted. Factuality values are displayed by Table 2.

3.2 Ext-MRC-style Definition for DEFI

For DEFI, an Ext-MRC-style sample is denoted as a triple group $\mathbb{S} = \{\mathbb{Q}, \mathbb{C}, \mathbb{A}\}$. Concretely, $\mathbb{Q}$ is the Question containing the event and candidate applicable factuality values, whose English and Chinese versions are as follows:

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

and $\mathbb{C}$ is the Context, i.e., the document from whose perspective the factuality of the event $\mathbb{E}$ can be inferred. $\mathbb{A}$ is the Answer that is the sub-string of $\mathbb{Q}$, which is different from most previous classical MRC tasks that extracted $\mathbb{A}$ from $\mathbb{C}$. The annotated $\mathbb{A}$ is one of the applicable values mentioned in Section 3.1.

3.3 Mch-MRC-style Definition for DEFI

A Mch-MRC-style DEFI sample is formulated as a quad group $\{\mathbb{Q}, \mathbb{C}, \mathbb{O}, \mathbb{A}\}$. Specifically, the Question $\mathbb{Q}$ is designed as:

- What is the factuality of the event "$\mathbb{E}$"?
- 事件"$\mathbb{E}$"的事实性是什么?

The Context $\mathbb{C}$ still refers to the document. $\mathbb{O}$ is the set of Options that are applicable factuality values, and the Answer $\mathbb{A}$ is one of $\mathbb{O}$ ($\mathbb{A} \in \mathbb{O}$).

4. Heterogeneous Spectral Attention Network

This section describes the architecture of the proposed Heterogeneous Spectral Attention Network (HSAN) model, covering input layer (Section 4.1), shared adapter (Section 4.2), task-specific adapter (Section 4.3), output layer (Section 4.4), as shown in Figure 2. And then we present the transfer learning mechanism in the model in Section 4.5.

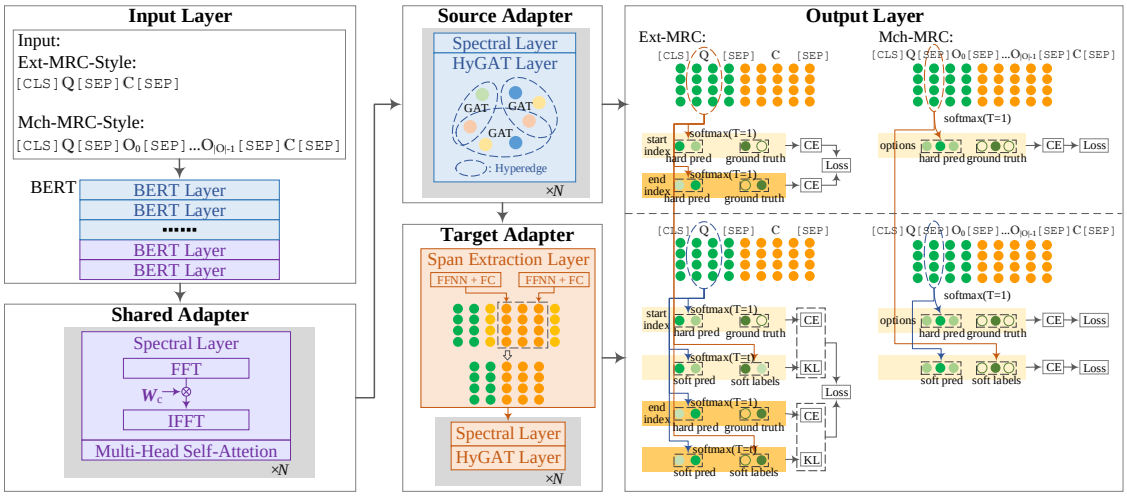

Figure 2: The architecture of the Heterogeneous Spectral Attention Network (HSAN).

4.1 Input Layer

BERT is selected as the backbone encoder. Based on the definition of Ext-MRC-style paradigm for DEFI, the input sequence $\mathbb{I}$ is comprised of the question $\mathbb{Q}$ and the context $\mathbb{C}$. While the input $\mathbb{I}$ of the Mch-MRC-style paradigm consists of $\mathbb{Q}$, the option set $\mathbb{O}$, and $\mathbb{C}$:

$$\mathbb{I} = \begin{cases} [\texttt{CLS}], \mathbb{Q}, [\texttt{SEP}], \mathbb{C}, [\texttt{SEP}]\}, \text{EXT} - \text{MRC} \\ [\texttt{CLS}], \mathbb{Q}, [\texttt{SEP}], \mathbb{O}_0, [\texttt{SEP}], \mathbb{O}_1, [\texttt{SEP}], \ldots, \mathbb{O}_{|\mathbb{O}|-1}, [\texttt{SEP}], \mathbb{C}, [\texttt{SEP}]\}, \text{Mch} - \text{MRC} \end{cases} \tag{1}$$

Then we can encode $\mathbb{I}$ as $\boldsymbol{H}_0$ using BERT, i.e., $\boldsymbol{H}_0 = \text{BERT}(\mathbb{I})$.

4.2 Shared Adapter

This network acts as a global encoder for both source (classical MRC) and target (DEFI) domain. Thus, this shared adapter is not only responsible for learning high-level abstract features for source and target tasks, but also extracting manifold semantics from global and local texts, which is implemented by stacks of multi-head version of spectral layers (Eq. (2)) and attention layers (Eq. (3)):

$$\boldsymbol{H}_{0,i}^{(1)} = \text{IFFT}(\boldsymbol{W}_{c,i}(\text{FFT}(\boldsymbol{H}_{0,i}))) \tag{2}$$

$$\boldsymbol{H}_{0}^{(2)} = \text{MHSA}(\text{Concat}(\{\boldsymbol{H}_{0,i}^{(1)}\})) \tag{3}$$

where $\boldsymbol{H}_{0,i}$ is sampled from $\boldsymbol{H}_0$ by linear layers, Concat is the concatenation operator. Spectral layer is composed of Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT), where $\boldsymbol{W}_{c,i}$ is a learnable parameter to determine the weight of each head to capture the most significant information flowing from FFT to IFFT. Therefore, the main function of spectral layer is to refine and capture different information, including event mentions (triggers, arguments, etc.), speculative and negative cues. MHSA is Multi-Head Self-Attention. To ensure the diversity and richness of features, we also consider the multi-view version of

spectral layers to obtain $\boldsymbol{H}_{0,j}^{(2)}$ according to the above equations, and the output $\boldsymbol{H}_1$ is the concatenation (denoted as the operator Concat) of them:

$$\boldsymbol{H}_1 = \text{Concat}(\{\boldsymbol{H}_{0,j}^{(2)}\}) \tag{4}$$

4.3 Task-Specific Adapter

This network contains Source Adapter and Target Adapter for the source and target domain. The main difference between them is that target adapter starts with a span extraction layer, but source adapter does not.

Source Adapter. Similar to shared adapter above, source adapter is made up of spectral layers and attention layers:

$$\boldsymbol{H}_{\text{SL}}^{\text{Src}} = \text{SpectLayer}(\boldsymbol{H}_1) \tag{5}$$

$$\boldsymbol{H}^{\text{Src}} = \text{HyGAT}(\boldsymbol{H}_{\text{SL}}^{\text{Src}}) \tag{6}$$

where SpectLayer is the spectral layer defined by Eq. (2). HyGAT is HyperGraph Attention neTwork computed below.

HyperGraph Attention neTwork (*HyGAT*). A hypergraph contains hyperedges and nodes, where each hyperedge can connect more than two nodes. HyGAT is able to learn more fine-grained local features by updating the states of tokens connected by hyperedges, which means we can aggregate the semantics of event triggers, arguments, speculative and negative cues, excluding the influence of unrelated texts. The first step is to determine the hyperedge, whose probability distribution is calculated as:

$$p(\text{HyE}) = \text{Sigmoid}(\boldsymbol{W}^{\text{HyE}}\boldsymbol{H}^{\text{Input}}) \tag{7}$$

then we can select the nodes on the hyperedges to form $\boldsymbol{H}^{\text{HyE}}$. To ensure semantic integrity of questions and options, their tokens are always nodes of hyperedges, and we mainly select tokens from the document $\mathbb{C}$. For any two tokens $\boldsymbol{q}, \boldsymbol{k} \in \boldsymbol{H}^{\text{HyE}}$ on the hyperedge, the multi-head attention weights $\boldsymbol{A}$ are computed as:

$$\boldsymbol{A}_{ij} = \frac{\exp(\sigma(\boldsymbol{\alpha}^\top \text{Concat}(\boldsymbol{k}_i, \boldsymbol{q}_j)))}{\sum_u \exp(\sigma(\boldsymbol{\alpha}^\top \text{Concat}(\boldsymbol{k}_u, \boldsymbol{q}_j)))} \tag{8}$$

and the states of nodes in the hyperedge are updated as below, where $\boldsymbol{H}^{\text{HyE},m}$ is one head sampled from the input:

$$\boldsymbol{H}^{\text{HyE}} = \text{Concat}_m(\{\boldsymbol{H}^{\text{HyE},m}\boldsymbol{A}^m\}) \tag{9}$$

then the states in $\boldsymbol{H}^{\text{Input}}$ is updated according to $\boldsymbol{H}^{\text{HyE}}$. We also consider several hyperedges to update the states for enough nodes.

Target Adapter. This network has one additional Span Extraction Layer (SpanLayer) than the source adapter, and directly utilizes the output of the source adapter $\boldsymbol{H}^{\text{Src}}$ as the input. We have noticed that document-level event factuality usually depends on span-level texts within sentence-level mentions, just as illustrated by Figure 1. Texts out of those spans may become noise and produce wrong semantics for the event factuality. Therefore,

the main motivation of span extraction is to filter irrelevant mentions and to obtain local span-level texts that contain the correct factuality for the event. Similar to hypergraph networks that ensure the completeness of questions and options, we apply span extraction on the representation of the document $\boldsymbol{H}_c^{\text{Src}} \in \boldsymbol{H}^{\text{Src}}$, and the distributions of start and end indices of one span is as follows:

$$\boldsymbol{p}_s^c = \text{softmax}(\boldsymbol{W}_s^c \text{FFN}(\boldsymbol{H}_c^{\text{Src}}) + \boldsymbol{b}_s^c) \tag{10}$$

$$\boldsymbol{p}_e^c = \text{softmax}(\boldsymbol{W}_e^c \text{FFN}(\boldsymbol{H}_c^{\text{Src}}) + \boldsymbol{b}_e^c) \tag{11}$$

where FFN is Feed Forward Network. Then we can obtain several spans $\{\boldsymbol{H}_{\text{sp},i}^{\text{Src}}\}$, which is concatenated to replace $\boldsymbol{H}_c^{\text{Src}}$ in $\boldsymbol{H}^{\text{Src}}$ to form an updated state $\boldsymbol{H}_{\text{sp}}^{\text{Tar}}$ as a high-level fine-grained representation. Next, $\boldsymbol{H}_{\text{sp}}^{\text{Tar}}$ is fed into another stack of spectral layer and hypergraph attention network. In general, the target adapter can be denoted as the follows:

$$\{\boldsymbol{H}_{\text{sp},i}^{\text{Src}}\} = \text{SpanLayer}(\boldsymbol{H}^{\text{Src}}) \tag{12}$$

$$\boldsymbol{H}_{\text{sp}}^{\text{Tar}} = \text{Concat}(\boldsymbol{H}_q^{\text{Src}}, \{\boldsymbol{H}_{\text{sp},i}^{\text{Src}}\}) \tag{13}$$

$$\boldsymbol{H}^{\text{Tar}} = \text{HyGAT}(\text{SpectLayer}(\boldsymbol{H}_{\text{sp}}^{\text{Tar}})) \tag{14}$$

where $\boldsymbol{H}_q^{\text{Src}}$ is the representation of the question $\mathbb{Q}$ in $\boldsymbol{H}^{\text{Src}}$.

4.4 Output Layer

For source and target adapter, we set output layers with the same structure in them. But the output layers of Ext-MRC and Mch-MRC should be discussed separately.

Ext-MRC. The answer is extracted from the question. Then, the probabilities of tokens in the question $\mathbb{Q}$ of being the start and end indices are denoted as:

$$\boldsymbol{p}_s^T = \text{softmax}((\boldsymbol{W}_s^o \boldsymbol{H}^o + \boldsymbol{b}_s^o)/T) \tag{15}$$

$$\boldsymbol{p}_e^T = \text{softmax}((\boldsymbol{W}_e^o \boldsymbol{H}^o + \boldsymbol{b}_e^o)/T) \tag{16}$$

where $\boldsymbol{H}^o \in \{\boldsymbol{H}^{\text{Src}}, \boldsymbol{H}^{\text{Tar}}\}$, T is the temperature. The objective function for the source domain (classical MRC) is:

$$\mathcal{L}_{\text{Ext}}^{\text{Src}} = \frac{1}{N}[\epsilon \text{CE}(\boldsymbol{p}_s^{T=1}, \boldsymbol{p}_s^g) + (1-\epsilon)\text{CE}(\boldsymbol{p}_e^{T=1}, \boldsymbol{p}_e^g)] \tag{17}$$

where CE is cross entropy loss, N is the number of samples, ϵ is the trade-off coefficient. $\boldsymbol{p}_{s/e}^g$ is the distribution of the annotated labels, i.e., the binary vector where only the start or end indices are 1.

For the target domain (DEFI task), we introduce soft labels from the source domain as the clues for knowledge distillation. The main motivation lies in the relevance of spans in source and target domain, since both typical MRC and Ext-MRC-style DEFI rely on span-level texts. Compared with the source dataset, spans of DEFI often require more elements, i.e., not only event triggers, arguments (similar to general MRC), but also speculation and

Algorithm 1 Training process of the HSAN model

Input: 1) Untrained HSAN model (Section 4); 2) MRC-style samples $\{\mathbb{S}\}$, where $\mathbb{S} = \{\mathbb{Q}, \mathbb{C}, \mathbb{A}\}$ for Ext-MRC, and $\mathbb{S} = \{\mathbb{Q}, \mathbb{C}, \mathbb{O}, \mathbb{A}\}$ for Mch-MRC

Output: Trained HSAN model.

1: Train HSAN on the source domain. These sub-networks are optimized: BERT, Shared Adapter, Source Adapter, Output Layer;

2: Fine-tune HSAN on the target domain. These sub-networks are optimized: last N_{ft} layers of BERT, Shared Adapter, Target Adapter, Output Layer.

negation. Therefore, the objective function is designed as:

$$\mathcal{L}_s^{\text{Tgt}} = \frac{1}{N}[\text{CE}(\boldsymbol{q}_s^{T=1}, \boldsymbol{p}_s^g) + \text{KL}(\boldsymbol{q}_s^T \| \boldsymbol{p}_s^T)] \tag{18}$$

$$\mathcal{L}_e^{\text{Tgt}} = \frac{1}{N}[\text{CE}(\boldsymbol{q}_e^{T=1}, \boldsymbol{p}_e^g) + \text{KL}(\boldsymbol{q}_e^T \| \boldsymbol{p}_e^T)] \tag{19}$$

$$\mathcal{L}_{\text{Ext}}^{\text{Tgt}} = \epsilon \mathcal{L}_s^{\text{Tgt}} + (1-\epsilon)\mathcal{L}_e^{\text{Tgt}} \tag{20}$$

where KL is Kullback-Leibler divergence, $\boldsymbol{p}$ and $\boldsymbol{q}$ are probability distribution learned by source and target adapter.

Mch-MRC. we adopt the state of the first `[SEP]` before the first option as the representation of the options, which is denoted as $\boldsymbol{h}^o$. The probability distribution of options are calculated as below, where T is the temperature:

$$\boldsymbol{p}^T = \text{softmax}((\boldsymbol{W}\boldsymbol{h}^o + \boldsymbol{b})/T) \tag{21}$$

The objective function for the source domain (classical Mch-MRC) is:

$$\mathcal{L}_{\text{Mch}}^{\text{Src}} = \frac{1}{N}\text{CE}(\boldsymbol{p}^{T=1}, \boldsymbol{p}^g) \tag{22}$$

where $\boldsymbol{p}^g$ is the distribution of annotated labels. Similar to Ext-MRC, we apply knowledge distillation for target domain to learn moderate information from the distribution predicted by source domain, and develop the objective function:

$$\mathcal{L}_{\text{Mch}}^{\text{Tar}} = \frac{1}{N}[\text{CE}(\boldsymbol{q}^{T=1}, \boldsymbol{p}^g) + \text{KL}(\boldsymbol{q}^T \| \boldsymbol{p}^T)] \tag{23}$$

4.5 Transfer Learning

The main principle of transfer learning is fine-tuning some portions of the model on the target domain after training the whole model and learning enough knowledge on the source domain. It is noticeable that this is also a data augmentation method, since transfer learning can study more information from samples of external classical MRC datasets. Therefore, our HSAN model can be divided into three parts, i.e., those sub-networks that are trained or fine-tuned 1) only on the source domain; 2) on both source and target domain; 3) only on the target domain. The process of training the model via transfer learning can be formulated by Algorithm 1.

	Uu	PS-	CT-	PS+	CT+	Total
English	42	51	745	660	3532	5030
Chinese	22	42	1504	953	2629	5150

Table 3: Statistics of ExDLEF corpus.

	Corpus	Task	Used	Total
English	SQuAD	Ext	15,000	130,217
	NewsQA	Ext	15,000	103,960
	RACE	Mch	15,000	87,866
	DREAM	Mch	10,197	10,197
Chinese	SQuAD-Ch	Ext	15,000	76,449
	CMRC2018	Ext	10,111	10,111
	C^3	Mch	6,013	11,869
	CMRC2017	Mch	15,000	310,000

Table 4: MRC corpora used as source datasets, where Ext/Mch mean Ext-MRC/Mch-MRC, and "Used" & "Total" means used & total samples in training sets.

5. Experimentation

In this section, we first introduce source and target datasets (Section 5.1), and then give experimental settings (Section 5.2), followed by baselines (Section 5.3). As for experimental analysis, overall results and analysis (Section 5.4) are presented, ablation study (Section 5.5) and case study (Section 5.6) are also considered as further detailed analysis.

5.1 Corpus

We take into account target and source corpora under the framework of transfer learning.

Target Corpus. We utilize ExDLEF corpus as the target dataset to evaluate the performance of our model. ExDLEF is an extended version of DLEF-v2 corpus (Qian et al., 2022a, 2022b), whose distributions of factuality values are displayed by Table 3. We can observe that CT-, PS+, CT+ events occupy the majority (98.15%/98.75% in the English and Chinese, respectively). Therefore, to be consistent with previous work (Qian et al., 2019, 2022a, 2022b; Zhang et al., 2022, 2022b, 2023, 2023), we mainly consider their performance, and neglect other minor values (i.e., Uu and PS-) with much smaller proportions.

Source Corpus. Under the framework of transfer learning as cross-domain data augmentation, we consider the following MRC corpora as the source datasets, whose statistics are given in Table 4:

English Ext-MRC: 1) **SQuAD** (Rajpurkar et al., 2018) is an extractive dataset whose documents are collected from Wikipedia, and we exploit version 2.0; 2) **NewsQA** (Trischler et al., 2017) contains news articles from CNN.

English Mch-MRC: 1) **RACE** (Lai et al., 2017) is a multiple-choice dataset where the passages are collected from English exams for middle and high school Chinese students; 2) **DREAM** (Sun et al., 2019) is a dialogue-based dataset, whose texts are collected from English as a foreign language examinations designed by human experts.

Chinese Ext-MRC: 1) **ChineseSQuAD**[1] (or **SQuAD-Ch** for short) is the Chinese version translated from the English SQuAD assisted by manual correction; 2) **CMRC2018** (Cui et al., 2019) is directly annotated on Chinese Wikipedia paragraphs.

Chinese Mch-MRC: 1) **C**^{3} (Sun et al., 2020), which contains dialogues or more formally written mixed-genre texts collected from Chinese-as-a-second-language examinations, is divided into C^3-Dialogue (C^3_D) and C^3-Mixed (C^3_M); 2) **CMRC2017** (Cui et al., 2018) is made up with texts from reading , and includes cloze track and user query track.

5.2 Experimental Settings

We consider 10-fold cross validation on ExDLEF corpus. For transfer learning model, It is worth noting that the samples in training sets include those from ExDLEF and other external MRC datasets. Therefore, in each fold, we report the average performance of the five rounds of experiments, where each round adopts a fixed-sized subset sampled from the source dataset randomly. F1-score is the main evaluation metrics for each category of factuality, and macro-/micro-averaged F1 (MacF/MicF) are also employed to obtain the performance of all the values.

5.3 Baselines

The following models are considered as competitive baselines for the fair comparison with our HSAN model:

Pipeline DEFI models. These methods are designed for previous DEFI tasks depending on event-related elements, e.g., event triggers, speculative and negative cues. Therefore, detecting these elements may produce cascaded errors. 1) **LSTM-Attn** (Qian et al., 2019) is a multi-layer LSTM model with hierarchical attentions. This model integrates syntactic (paths from cues to event triggers in dependency trees) and semantic (sentences with event triggers) features; 2) **ULGN** (Cao et al., 2021) is an uncertain local-to-global network that models the uncertainty of local information and leverages the global structure for integrating the local information; 3) **URHAT** (Sheng et al., 2023) is an uncertain relational hypergraph attention network, which reframes the document graph as a hypergraph and learns features of uncertain nodes to summarize document-level event factuality.

End-to-End DEFI model. This model is developed to address DEFI defined in this paper, i.e., only involving in the event and the document. Currently, only **RMAHN** (Qian et al., 2022a) has been published, which is a hierarchical attention network using reinforcement learning for token and sentence selection.

Light-weighted MRC models. These models are used to prove that DEFI models can get benefit from external knowledge introduced by transfer learning, rather than complicated structures only. 1) **BiDAF** (Seo et al., 2017) employs bidirectional attention flow to rep-

1. https://github.com/junzeng-pluto/ChineseSquad

Models	CT-	PS+	CT+	MacF	MicF
LSTM-Attn	44.19 | 62.07	43.55 | 53.24	81.34 | 78.28	56.36 | 64.53	70.96 | 68.72
ULGN	47.24 | 64.34	45.31 | 54.51	82.07 | 78.36	58.20 | 65.74	71.84 | 69.78
URHAT	50.68 | 66.38	46.65 | 55.14	82.80 | 78.73	60.04 | 66.75	73.33 | 70.56
BiDAF	51.49 | 68.15	49.11 | 59.04	81.32 | 80.47	60.64 | 69.22	72.11 | 73.20
+TL	54.89 | 73.48	52.78 | 63.51	82.94 | 83.55	63.54 | 73.51	74.55 | 76.96
QANet	52.83 | 70.86	50.55 | 61.52	82.55 | 81.68	61.98 | 71.35	73.42 | 74.61
+TL	55.03 | 73.54	53.35 | 64.38	83.26 | 83.29	63.88 | 73.74	74.87 | 77.09
RMHAN	57.17 | 74.76	55.39 | 65.19	84.51 | 82.53	65.69 | 74.16	76.73 | 77.10
HSAN					
Mch	56.29 | 73.77	52.78 | 64.43	84.40 | 82.42	64.49 | 73.54	75.87 | 76.58
+TL	60.22 | 76.48	56.07 | 67.35	85.52 | 84.56	67.27 | 76.13	77.58 | 78.71
Ext	58.38 | 75.71	54.76 | 65.56	85.25 | 83.68	66.13 | 74.99	77.26 | 77.84
+TL	**63.43** | **79.29**	**60.71** | **70.51**	**87.30** | **85.62**	**70.48** | **78.47**	**80.09** | **80.75**

Table 5: Performance of models on DEFI. Format: F1-scores for "English | Chinese" sub-corpus. "Ext" and "Mch" represent span-extraction and multiple choice style MRC of our HSAN model. "TL" denotes transfer learning. For English source datasets, we consider SQuAD and RACE for Ext-MRC and Mch-MRC. While for Chinese source datasets, we select SQuAD-Ch and CMRC2017.

resent contexts at different levels and learn query-aware representations; 2) **QANet** (Yu et al., 2018) consists of convolution and self-attention modeling local and global interactions.

5.4 Overall Results and Analysis

Table 5 exhibits the performance of several baselines and our HSAN model on DEFI task. We analyze the results according the following aspects:

1) *Encoder*. HSAN model is superior to other baselines that also utilize transfer learning, including not only BiDAF, QANet, but our conference model (Figure 3). One of the primary reasons is HSAN integrates several types of encoders outside BERT, e.g., spectral layer, multi-head self-attention, hypergraph attention networks, which can learn more useful high-level semantics for DEFI. But Qian et al. (2022b) only considered residual networks.

2) *English vs. Chinese*. The performance on Chinese sub-corpus is better than that on the English one, mainly due to the more balanced distribution of factuality values, i.e., Chinese CT+ events are fewer than the English CT+ ones, while Chinese CT- and PS+ events are more than corresponding English ones.

3) *Pipeline models*. LSTM-Attn, ULGN, URHAT are dependent on several elements, e.g., event triggers, speculative and negative cues. Currently, the mainstream methods still pursue using sentence-level information to infer document-level factuality. These models give concrete solutions that dive into lexical elements, and usually gain lower results than other methods, attributed to the cascade errors from upstream tasks.

4) *Text selection*. RMHAN integrates token and sentence selection to capture sentence-level factuality. Hence, RMHAN can achieve satisfactory performance compared with most models. Our HSAN model reach way down into continual spans within sentences, which encompass event triggers, arguments, speculative and negative cues, to obtain more fine-

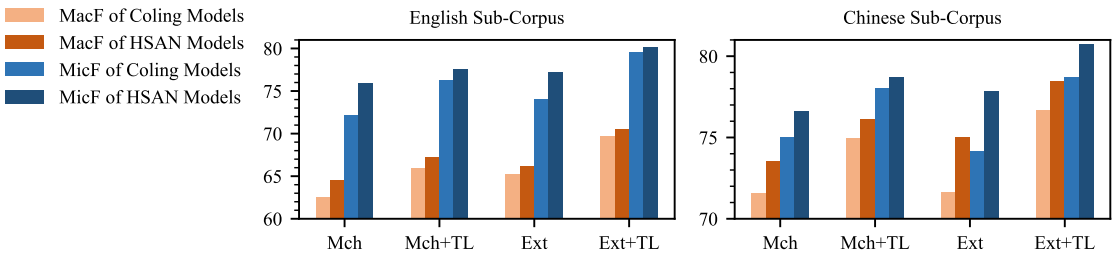

Figure 3: The performance of HSAN model in comparison with Qian et al. (2022b) on ExDLEF corpus.

Languages	**Models**	**Source Datasets**	**CT-**	**PS+**	**CT+**	**MacF**	**MicF**
English	Ext	SQuAD	63.43	60.71	87.30	70.48	80.09
		NewsQA	63.84	58.68	86.75	69.76	79.50
	Mch	RACE	60.22	56.07	85.52	67.27	77.58
		DREAM	57.37	54.83	84.49	65.55	75.97
Chinese	Ext	SQuAD-Ch	79.29	70.51	85.62	78.47	80.75
		CMRC2018	77.71	67.40	84.15	76.42	79.19
	Mch	CMRC2017	76.48	67.35	84.56	76.13	78.71
		C^3	72.05	62.19	83.74	72.64	76.26
		C^3-Mixed	74.36	64.33	84.47	74.39	77.82
		C^3-Dialogue	71.81	60.74	82.55	71.70	75.27

Table 6: Performance of our HSAN model on DEFI with difference source datasets. Evaluation metrics: F1-scores.

grained span-level semantics to infer document-level factuality. Enhanced by transferred semantics from other MRC datasets, HSAN can beat RMHAN on MacF and MicF.

5) *Transfer Learning.* With the application of transfer learning, the performance of corresponding MRC models (BiDAF, QANet, HSAN) can all be improved. This testifies that models can indeed learn useful transferred semantics from external classical MRC datasets. The key point of DEFI is extracting and integrating sentence-level factuality to deduce document-level factuality, which is similar to typical Ext-MRC and Mch-MRC whose answers can be inferred from or consistent with sentence/span-level texts. Hence, HSAN can beneficial from external MRC datasets.

6) *Ext-MRC vs Mch-MRC.* Compared with Mch-MRC datasets, Ext-MRC-style source datasets can usually bring higher performance for the HSAN model. We argue that there is more similarity between Ext-MRC and DEFI, since both of them require correct identification of span-level texts to infer the document-level query. But options in Mch-MRC datasets are sometimes rewritten forms of corresponding spans, and the textual inconsistency may lead to the difficulty of semantic understanding.

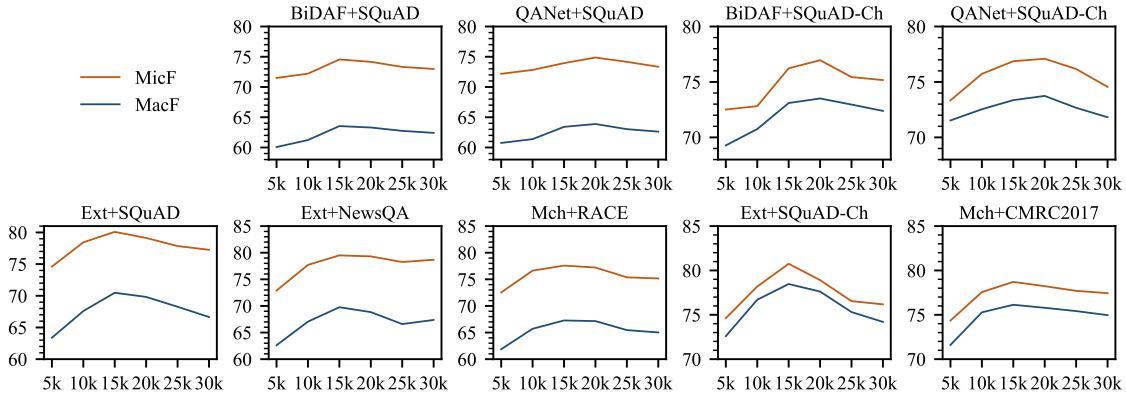

Figure 4: The performance (Y-axis) of some transfer learning models, i.e., HSAN (Ext&Mch-MRC-style), BiDAF and QANet, with regard to the scale of samples from source dataset (X-axis). Each sub-figure is titled as "Model+Source Dataset".

5.5 Ablation Study

To investigate the contributions of input, models, internal mechanism, we launch several ablation studies as below.

1) *Different source datasets.* Table 6 is used to explore the relation between the performance and different source datasets, where we consider several Ext-MRC and Mch-MRC corpora as the source datasets for transfer learning.

For English Ext-MRC, we can obtain excellent F1-scores when we use SQuAD and NewsQA as source datasets, which demonstrates that the transferred knowledge from these typical MRC corpora can indeed help to extract local span-level information that contains correct factuality for events in the target dataset. We also notice that higher F1-scores can be achieved if we select SQuAD than NewsQA, mainly attributed to the sentences with proper and regular grammar in Wikipedia documents of SQuAD. While news texts in NewsQA are a bit more free-style than that of SQuAD. In addition, the complicated semantics of NewsQA texts also brings multi-hop situations of MRC, which increases the difficulty of cross-domain adaptation.

For Chinese Ext-MRC-style source datasets, SQuAD-Ch can lead to higher F1-scores than CMRC2018, and both of them are collected from Wikipedia. It is obvious that SQuAD-Ch has more samples and can offer more span-level information with more training samples than CMRC2018 (Table 4).

In terms of Mch-MRC-style corpora, document-based datasets (RACE, CMRC2017) can be more useful than dialog-based ones (DREAM, C^3_D) for our HSAN model on the performance of DEFI. It is obviously due to the consistency of the text genre of document-based datasets and our ExDLEF corpus. C^3 is a compound dataset combined by dialogues (C^3_D) and exams (C^3_M), and we can observe that dialogues can result in performance degradation compared with C^3 and C^3_M, in which more formal texts are included. In addition to the

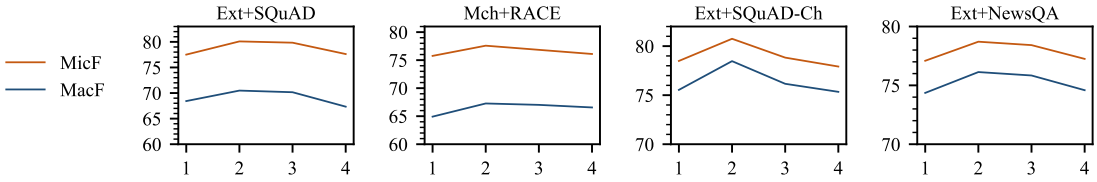

Figure 5: The performance (Y-axis) of HSAN model with regard to the number of extracted spans (X-axis). Each sub-figure is titled as "Model+Source Dataset".

inconsistency caused by dialogues, the average context length of C^3_D (76.31) is shorter than C^3_M (180.21), causing C^3_D to offer less semantics than C^3_M.

2) *Light-weighted MRC models.* The performance of light-weighted models (BiDAF and QANet) can be also facilitated after utilizing transfer learning mechanism. The significance of this group of models lies in the confirmation of the positive function of transfer learning when we exclude the influence of complicated structures in MRC models. Figure 4 shows that compared with our HSAN model, more samples (usually about 20,000) in the source dataset are fed into BiDAF and QANet to obtain the optimal MacF and MicF scores. This is mainly due to the lower learning capability caused by their simpler architectures.

3) *Size of samples from source datasets.* Since we consider transferred knowledge from external MRC datasets, we plan to investigate the relation between the scale of samples from source datasets and the performance of HSAN model. Thereby, we mainly study SQuAD/SQuAD-Ch, NewsQA, RACE, CMRC2017, since their scales are large enough. As displayed in Figure 4, after reaching the optimal performance with a moderate size of samples (usually about 15,000 samples), the MacF and MicF will not continue to improve even if we feed more source samples. Hence, we apply 15,000 samples to train the HSAN model on source datasets in Table 5. The principal reason is overfitting on source datasets, i.e., excessive knowledge transferred from source datasets will become the dominant in the model and interfere with the fine-tuning on the target dataset.

4) *Number of Extracted Spans.* In task-specific adapter (Section 4.3) of HSAN, we set a span extraction layer to capture spans of texts for events as dynamic and adaptive span selection mechanism. Fewer spans are likely to miss the correct sentence-level mentions that can infer to the document-level event factuality, but too many spans may cover the whole sentence, causing the span extraction to be meaningless. Hence, we plan to the upper bound of the number of extracted spans, and present Figure 5, which reveals the performance of HSAN when numbers of span vary. We can observe that appropriate number of spans can indeed contribute to high results, because extracted spans can cover correct sentence-level event mentions, including event triggers & arguments, speculative & negative semantics. Thus, we set the number of sampled spans as 2. Nevertheless, more spans do not bring better performance. Since the extracted spans are dominated by event triggers and arguments, texts that are unrelated to the factuality, or wrong sentence-level event factuality may also be involved, which are not helpful for the performance of our model.

EE1: Ukraine's drones attacked Zaporizhzhia Nuclear Power Plant.
A: certain negative (CT-) / **P**: certain negative (CT-)
… … However, Ukraine on Sunday denied the accusation of Kyiv's involvement in using drone to attack ZNPP. "Ukraine is not involved in any armed provocations on the territory of the ZNPP, nor attacking by drones," Andriy Yusov, a spokesman for Ukraine's military intelligence, … ...

EE2: U.S. involves in Israel-Gaza war.
A: possible positive (PS+) / **P**: possible positive (PS+)
… … While Biden has not yet sent troops, the possibility that the United States involves in the Israel-Gaza war is now "higher than most people realize," according to Michael DiMino, "sometimes intent isn't enough to prevent things from spiraling out of control. And that may involve direct U.S. involvement in some cases" … …

EE3: NASA returns humans to the Moon.
A: possible positive (PS+) / **P**: certain positive (CT+)
… … NASA is possibly gambling that commercial partners may help it to return to the Moon again by taking over some crucial tasks that it handled during the Apollo era. In May, NASA announced that it had signed contracts with three companies that will each carry as many as 14 experiments to the Moon aboard small robotic landers … …

EE4: US and Iran reach a formal agreement.
A: certain negative (CT-) / **P**: possible positive (PS+)
… … Iran has reportedly slowed the pace at which it is enriching uranium to nearly weapons-grade levels, which means an agreement has been likely reached between the US and Iran. However, asked whether US is pursuing a formal agreement with Iran at the press conference, Secretary of State Blinken denied there is agreement in the offing. "… … whether Iran takes actions to reach a deal not only between us, but with other countries, we will see by their actions", said Blinken … …

CE1: 法国从尼日尔撤军。(France withdraws its troops from Niger.)
A: 一定发生(CT+) / **P**: 一定不发生 (CT-)
……马克龙称，法国驻尼日尔的军队应邀协助打击恐怖主义，如今“尼日尔事实上的政权”不再愿意打击恐怖主义，因此结束军事合作，并不得不撤军，同时将与尼日尔政变军人协商，以确保撤军平稳完成……不过法国仍拒绝承认尼日尔军政府是合法政权。……
(… …Macron claimed that French troops in Niger were invited to help to fight terrorism. Now that the "Niger's de facto authorities" no longer want to fight terrorism, France ends military cooperation and has to withdraw its troops. At the same time, France will negotiate with the Niger coup soldiers to ensure a smooth withdrawal... But France still refuses to recognize Niger's military government as legitimate… …)

CE2: 沙特承认以色列。(Saudi Arabia recognizes Israel.)
A: 可能不发生 (PS-) / **P**: 可能发生 (PS+)
……过去两年的美沙关系是近代史上最糟糕的时期，而中国斡旋沙伊和解的举动，则为中东地区带来了罕见的和解潮。在这样的大背景下，沙特很可能继续拒绝承认以色列，同样，美国也不愿意见到沙特和伊朗、叙利亚等国的和解。……
(... … the US-Saudi relationship in the past two years was the worst in modern history, and China's mediation of Saudi-Iranian reconciliation has brought a rare wave of reconciliation to the Middle East. Against this backdrop, Saudi Arabia is likely to continue to refuse to recognize Israel, and the United States is also unwilling to see Saudi Arabia reconcile with Iran, Syria and other countries ... …)

Figure 6: Several samples whose factuality and spans identified by our HSAN model, where "A/P" denotes "Annotated/Predicted" labels. Green/Red means correct/wrong predicted labels. Extracted spans are highlighted by Orange, and are merged into one if there are overlaps among spans.

5.6 Case Study

To illustrate the interpretability of our model and the predicted results more convincingly, we give several English and Chinese cases, which include correct and error samples identified by HSAN model, in Figure 6, respectively. To reveal the effectiveness of span extraction mechanism, we also highlight the spans extracted by our model. It should be noted that we aim to infer the document-level factuality based on the scope of the text that contains the correct factuality of the event, instead of detecting the precise spans. The performance of CT+ is quite high due to their majority. Therefore, error cases are usually non-factual

because of wrong identification of speculation and negation, which will be discussed in detail below. Additionally, there are also other errors owing to non-applicable texts (e.g., "positive, certain positive", which cannot be mapped into applicable factuality values) extracted from the question of Ext-MRC, which is not the point of discussion here due to their minority.

In Figure 6, our Ext-MRC-style HSAN predicts the correct results for the CT- event EE1 and PS+ event EE2. For EE1, the extracted span contains the negative cues "denied, not", while for EE2, the span includes the speculative cue "possibility". And our model has correctly determined that these non-factual semantics govern the corresponding event.

We also analyze two error cases EE3 and EE4. EE3 is a PS+ event according to speculative cues "possibly, may". But our model only focuses on the trigger "return" and the argument "the Moon", and fails to include corresponding speculative cues in the span. EE4 is annotated as CT-, and its truth is negated by "denied", which is also excluded from the predicted span. Our model evaluates EE4 as PS+ mistakenly, mainly effected by the speculative word "likely" in the span.

Chinese samples are closely related to speculation and negation, and most correct and error cases are similar to the English cases in Figure 6. Here, we focus on the analysis of two error cases, which presents the intrinsic property of Chinese speculative and negative cues. The event CE1 is a fact, but is judged as CT- mistakenly, probably caused by the character "不(not)" due to the character-level encoding. Actually, the phrase "不得不" is inseparable and means "have to", rather than negation. In terms of another PS- event CE2, although the predicted span contains "可能(likely)" and "拒绝(refuse)", which denotes incomplete negation, our model still commits to CE2 as PS+. This means our model fails to identify negation from "拒绝(refuse)", leading to the absence of negative semantics for CE2.

6. Conclusion

We investigate DEFI task, which only considers an event and a corresponding document. To solve the shortcomings in existing work, such as data limitation, ignorance of span-level information, coarse-grained, lack of unified framework, we develop a novel MRC-style model called Heterogeneous Spectral Attention Network (HSAN), which formulates DEFI as span extraction and multiple choice MRC tasks, and can learn span-level semantics. In particular, to address data scarcity, especially the deficiency of span-level information, we employ transfer learning to study more local semantics from external MRC datasets. Experiments on ExDLEF corpus demonstrate the effectiveness of our model that can beat several strong baselines. In the future, we plan to explore more methods of data augmentation, including cross-lingual, cross-domain, or multi-modal solutions. Moreover, we will also study more fine-grained DEFI tasks, e.g., identifying factuality for different participants of an event.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 62276177, 62006167 and 62376181), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 24KJB520036), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Cao, P., Chen, Y., Yang, Y., Liu, K., & Zhao, J. (2021). Uncertain local-to-global networks for document-level event factuality identification. In *Proceedings of EMNLP 2021*, pp. 2636–2645.

Cao, T., Wang, C., Tan, C., Huang, J., & Zhu, J. (2023). Sharing, teaching and aligning: Knowledgeable transfer learning for cross-lingual machine reading comprehension. In *Findings of EMNLP 2023*, pp. 455–467.

Cui, Y., Liu, T., Che, W., Xiao, L., Chen, Z., Ma, W., Wang, S., & Hu, G. (2019). A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of EMNLP 2019*, pp. 5882–5888.

Cui, Y., Liu, T., Chen, Z., Ma, W., Wang, S., & Hu, G. (2018). Dataset for the first evaluation on chinese machine reading comprehension. In *Proceedings of LREC 2018*.

He, Z., Cao, P., Jin, Z., Chen, Y., Liu, K., Zhang, Z., Sun, M., & Zhao, J. (2024). Zero-shot cross-lingual document-level event causality identification with heterogeneous graph contrastive transfer learning. In *Proceedings of COLING 2024*, pp. 17833–17850.

Huang, R., Zou, B., Wang, H., Li, P., & Zhou, G. (2019). Event factuality detection in discourse. In *Proceedings of NLPCC 2019*, Vol. 11839 of *Lecture Notes in Computer Science*, pp. 404–414.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. H. (2017). RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of EMNLP 2017*, pp. 785–794.

Liu, J., Chen, Y., & Xu, J. (2022a). Document-level event argument linking as machine reading comprehension. *Neurocomputing*, *488*, 414–423.

Liu, J., Chen, Y., & Xu, J. (2022b). Mrcaug: Data augmentation via machine reading comprehension for document-level event argument extraction. *IEEE ACM Transactions on Audio, Speech and Language Processing*, *30*, 3160–3172.

Liu, L., Liu, M., Liu, S., & Ding, K. (2024). Event extraction as machine reading comprehension with question-context bridging. *Knowledge-Based Systems*, *299*, 112041.

Lu, D., Ran, S., Tetreault, J. R., & Jaimes, A. (2023). Event extraction as question generation and answering. In *Proceedings of ACL 2023*, pp. 1666–1688.

Qian, Z., Li, P., Zhu, Q., & Zhou, G. (2019). Document-level event factuality identification via adversarial neural network. In *Proceedings of NAACL-HLT 2019*, pp. 2799–2809.

Qian, Z., Li, P., Zhu, Q., & Zhou, G. (2022a). Document-level event factuality identification via reinforced multi-granularity hierarchical attention networks. In *Proceedings of IJCAI 2022*, pp. 4338–4345.

Qian, Z., Zhang, H., Li, P., Zhu, Q., & Zhou, G. (2022b). Document-level event factuality identification via machine reading comprehension frameworks with transfer learning. In *Proceedings of COLING 2022*, pp. 2622–2632.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. In *Proceedings of ACL 2018*, pp. 784–789.

Seo, M. J., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. In *Proceedings of ICLR 2017*.

Sheng, J., Cong, X., Cao, J., Guo, S., Li, C., Wang, L., Liu, T., & Xu, H. (2023). Uncertain relational hypergraph attention networks for document-level event factuality identification. In *Proceedings of ECAI 2023*, Vol. 372, pp. 2129–2137.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., & Cardie, C. (2019). DREAM: A challenge dataset and models for dialogue-based reading comprehension. *TACL*, *7*, 217–231.

Sun, K., Yu, D., Yu, D., & Cardie, C. (2020). Investigating prior knowledge for challenging chinese machine reading comprehension. *TACL*, *8*, 141–155.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2017). Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP (Rep4NLP@ACL 2017*, pp. 191–200.

Uddin, M. N., George, E. R., Blanco, E., & Corman, S. R. (2024). Asking and answering questions to extract event-argument structures. In *Proceedings of COLING 2024*, pp. 1609–1626.

Wu, L., Wu, S., Zhang, X., Xiong, D., Chen, S., Zhuang, Z., & Feng, Z. (2022). Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. In *Proceedings of ACL 2022*, pp. 991–1000.

Yu, A. W., Dohan, D., Luong, M., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of ICLR 2018*.

Zhang, H., Li, P., Qian, Z., & Zhu, X. (2023). Incorporating factuality inference to identify document-level event factuality. In *Findings of ACL 2023*, pp. 13990–14002.

Zhang, H., Qian, Z., Li, P., & Zhu, X. (2022). Evidence-based document-level event factuality identification. In *Proceedings of PRICAI 2022*, Vol. 13630 of *Lecture Notes in Computer Science*, pp. 240–254.

Zhang, H., Qian, Z., Zhu, X., & Li, P. (2021). Document-level event factuality identification using negation and speculation scope. In *Proceedings of ICONIP 2021*, Vol. 13108 of *Lecture Notes in Computer Science*, pp. 414–425.

Zhang, K., Shuang, K., Yang, X., Yao, X., & Guo, J. (2023). What is overlap knowledge in event argument extraction? APE: A cross-datasets transfer learning model for EAE. In *Proceedings of ACL 2023*, pp. 393–409.

Zhang, Z., Strubell, E., & Hovy, E. H. (2022a). Transfer learning from semantic role labeling to event argument extraction with template-based slot querying. In *Proceedings of EMNLP 2022*, pp. 2627–2647.

Zhang, Z., Liu, C., Qian, Z., Zhu, X., & Li, P. (2022b). Hs^{2}n: Heterogeneous semantics-syntax fusion network for document-level event factuality identification. In *Proceedings of PRICAI 2022*, Vol. 13630 of *Lecture Notes in Computer Science*, pp. 309–320.

Zhang, Z., Qian, Z., Zhu, X., & Li, P. (2023). Code: Contrastive learning method for document-level event factuality identification. In *Proceedings of DASFAA 2023*, Vol. 13945 of *Lecture Notes in Computer Science*, pp. 497–512.