

IMCI: Integrate Multi-view Contextual Information for Fact Extraction and Verification

Hao Wang[†], Yangguang Li[‡], Zhen Huang^{†*}, Yong Dou[†]

[†] National University of Defense Technology, Changsha

[‡] SenseTime, Beijing

{hao.wang, huangzhen, yongdou}@nudt.edu.cn

liyangguang@sensetime.com

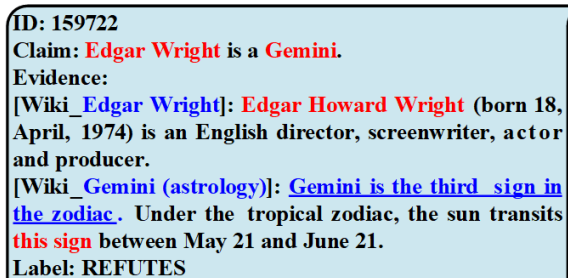
Abstract

With the rapid development of automatic fake news detection technology, fact extraction and verification (FEVER) has been attracting more attention. The task aims to extract the most related fact evidences from millions of open-domain Wikipedia documents and then verify the credibility of corresponding claims. Although several strong models have been proposed for the task and they have made great progress, we argue that they fail to utilize multi-view contextual information and thus cannot obtain better performance. In this paper, we propose to integrate multi-view contextual information (IMCI) for fact extraction and verification. For each evidence sentence, we define two kinds of context, i.e. **intra-document context** and **inter-document context**. Intra-document context consists of the document title and all the other sentences from the same document. Inter-document context consists of all other evidences which may come from different documents. Then we integrate the multi-view contextual information to encode the evidence sentences to handle the task. Our experimental results on FEVER 1.0 shared task show that our IMCI framework makes great progress on both fact extraction and verification, and achieves state-of-the-art performance with a winning FEVER score of 73.96% and label accuracy of 77.25% on the online blind test set. We also conduct ablation study to detect the impact of multi-view contextual information. Our codes will be released at <https://github.com/phoenixsecularbird/IMCI>.

1 Introduction

Fake news propagation is a severe social problem, which may cause great loss and lead to serious consequence, e.g. panic, quarrel, opposition and even war. The situation has become a general concern since Brexit and the U.S. President Campaign in

2016 and gets far more intense due to COVID-19 pandemic (Martino et al., 2020). In this condition, automatic fake news detection has been developing rapidly. According to Ruffo et al. (2021), automatic fake news detection mainly include textual-content based methods (Giachanou et al., 2019; Ghanem et al., 2020; Kaliyar et al., 2021), user-role based methods (Vo and Lee, 2019; Giachanou et al., 2020), multi-modal approaches (Zlatkova et al., 2019; Fung et al., 2021) and detection of bots and trolls (Stella et al., 2018; Sayyadiharikandeh et al., 2020). Among textual-content based methods, fact extraction and verification (FEVER) (Thorne et al., 2018) has been attracting developing attention. As shown in Figure 1, for a given claim, the task aims to select at most 5 most related sentences as evidences from millions of open-domain Wikipedia documents for fact extraction, and combine the selected evidences to judge the claim as SUPPORTS, REFUTES or NOT ENOUGH INFO (NEI) for fact verification.



ID: 159722
Claim: **Edgar Wright is a Gemini.**
Evidence:
[Wiki_Edgar Wright]: **Edgar Howard Wright** (born 18, April, 1974) is an English director, screenwriter, actor and producer.
[Wiki_Gemini (astrology)]: **Gemini is the third sign in the zodiac.** Under the tropical zodiac, the sun transits **this sign** between May 21 and June 21.
Label: REFUTES

Figure 1: An example from FEVER 1.0 shared task. (**Underlined sentence** is the **unlabeled intra-document context**. Words in red involve alias name, coreference and multi-hop reasoning, which may lead to model confusion. Words in blue help to handle these issues.)

Recently, several strong models (Nie et al., 2019b,a; Zhou et al., 2019; Liu et al., 2020; Hidey et al., 2020; Subramanian and Lee, 2020) have

*Corresponding author

been proposed for the task. Although they have made great progress and obtained excellent performance on the task, we argue that they fail to utilize multi-view contextual information and thus cannot obtain better performance. Specifically, we define two kinds of context for each evidence sentence, i.e. **intra-document context** and **inter-document context**. Intra-document context consists of the document title and all the other sentences from the same document. Inter-document context consists of all other evidences which may come from different documents. Multi-view contextual information is of great importance for fact extraction and verification. For instance, as shown in Figure 1, intra-document context information can help to clarify the relationship between different entities, e.g. “*Edar Wright*” and its alias name “*Edar Howard Wright*” in the first evidence, and “*Gemini*” and its coreference “*this sign*” in the second sentence. Besides, the two evidence sentences can be regarded as inter-document context of each other, and the information interaction and fusion between them is essential to verify the claim in this multi-hop sample.

To this end, we propose to integrate multi-view contextual information (IMCI) for fact extraction and fact verification, where we introduce the multi-view contextual information to encode the evidence sentences to handle the task. In summary, our contributions are as follows:

- We propose an iterative multi-view fact extraction model. It retrieves related documents and extracts related evidence sentences in two iterations, with multi-view context information joined.
- We propose a multi-view fact verification model. Each evidence sentence is encoded from two views, and a dual evidence fusion graph is adopted to fuse the information from diverse views and different evidences.
- Our IMCI framework makes great progress on both fact extraction and verification, and achieves state-of-the-art performance with a winning FEVER score of 73.96% on the online blind test set.

2 Iterative Multi-view Fact Extraction

Our fact extraction model iteratively conducts document retrieval and sentence retrieval in two iterations to obtain corresponding candidate evidence sentences, and then reranks the candidates of different iterations for better performance.

2.1 Document Retrieval

Document retrieval includes *coarse document retrieval* in iteration 1 and *refined document retrieval* in iteration 2.

Coarse document retrieval aims to quickly obtain most related documents from millions of open-domain Wikipedia documents with as high as possible recall and acceptable precision. Inspired by UKP-Athene (Hanselowski et al., 2018) and SR-MRS (Nie et al., 2019b), coarse document retrieval is a combination of constituency-based Wikipedia search and TF-IDF retrieval. These two respectively utilize search engine power and statistical word frequency information. For constituency-based Wikipedia search, we also conduct mention filtering like UKP-Athene (Hanselowski et al., 2018). That is, if the title of a document is not explicitly mentioned in the claim, then we consider it as weakly related and remove it.

Refined document retrieval aims to retrieve documents with improved performance than coarse retrieval, namely higher recall and also higher precision and F1 score. It adopts dense semantic retrieval and utilizes Wikipedia hyperlinks. Specifically, in iteration 2, we decide refined candidate documents according to corresponding candidate evidences from iteration 1. That is, for each claim, all documents which contain at least one candidate evidence will be taken into account. Furthermore, as top one candidate evidences show pretty high precision (86.11%, in Table 5), we regard them as golden evidence, and take all documents which have hyperlinks with them as refined candidate documents to process multi-hop problem.

2.2 Sentence Retrieval

Sentence retrieval aims at selecting most related sentences as evidences from candidate documents. In previous models, during sentence retrieval, it is required to design sampling strategy to obtain negative samples for neural retrieval model training. Besides, these models respectively encode and score each claim-sentence pair.

Differently, in our framework, to avoid sampling strategy design and also utilize multi-view contextual information, we encode each sentence within its corresponding intra-document context. Moreover, as mentioned, top one candidate evidences of iteration 1 show pretty high precision (86.11%, in Table 5). Therefore, in iteration 2 we take them as inter-document context, and insert them into the

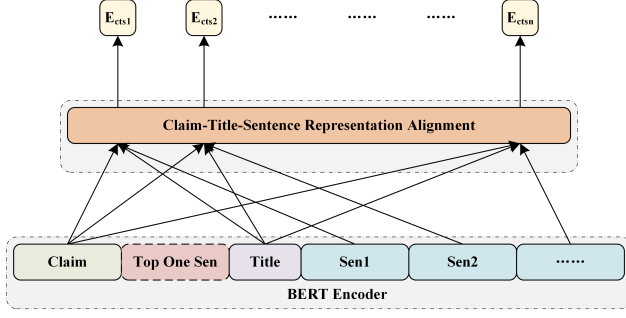


Figure 2: Sentence retrieval model. Sentences are encoded within intra-document context. In iteration 2, we insert top one candidate evidence (red dashed box) into the input sequence as inter-document context.

input sequence to process multi-hop problem.

Formally, as shown in Figure 2, the claim, the document title and all the sentences in the document are concatenated:

$$[CLS] \text{ claim } [SEP] \text{ sen}^* [SEP] \text{ title} \\ [SEP] \text{ sen1 } \text{ sen2 } \dots [SEP] \quad (1)$$

where sen^* denotes top one candidate evidence of iteration 1, which are taken as inter-document context in iteration 2. The sequence is encoded by BERT encoder. For the claim, we take the hidden state of the first claim token as claim representation E_c . For the title, we take the hidden state of the first title token as title representation E_t . For each sentence, we take the hidden state of the first sentence token as the sentence representation E_s . The sentence representation is enhanced through alignment with the title representation:

$$E_{ts} = W_a[E_t, E_s, E_t - E_s, E_t \odot E_s] \quad (2)$$

and the claim representation:

$$E_{cts} = W'_a[E_c, E_{ts}, E_c - E_{ts}, E_c \odot E_{ts}] \quad (3)$$

where \odot means element-wise Hadamard product. Then, the score of sentence \hat{y} is obtained through a Multi Layer Perceptron (MLP) with sigmoid activation function:

$$\hat{y} = \text{Sigmoid}(\text{MLP}(E_{cts})) \quad (4)$$

The training objective of sentence retrieval is defined as binary cross entropy loss, to maximize the probability of groundtruth evidence sentences:

$$\mathcal{L}_E = -\frac{1}{\sum_{i=1}^m n_i} \sum_{i=1}^m \sum_{j=1}^{n_i} [y_{ij} \cdot \log(\hat{y}_{ij}) \\ + (1 - y_{ij}) \cdot \log(1 - \hat{y}_{ij})] \quad (5)$$

where m is the batch size, n_i is the sentence number of document i , and y is the sentence label, 1 for groundtruth evidence sentences while 0 for non-evidence sentences.

2.3 Full Pipeline

In each iteration, we have scored different sentences as candidate evidences. To obtain better performance, for each claim, we merge the results of different iterations, and rerank the sentences through their scores. Finally, according to the original setup of the task, we keep at most top 5 sentences as evidences, for further fact verification.

3 Multi-view Fact Verification

3.1 Multi-view Contextual Encoding

For each evidence sentence, we respectively obtain its representations through intra-document encoding and inter-document encoding.

- **Intra-document Encoding** aims to capture intra-document contextual information of each evidence sentence. It is similar to the sentence retrieval model in Section 2.2. Each evidence sentence is encoded within its intra-document context. Then its intra-document representation is also obtained through alignment.

- **Inter-document Encoding** is utilized to capture token-level information interaction among different evidence sentences to handle multi-hop problem. The claim, all evidence sentences and their document titles are concatenated as another input sequence:

$$[CLS] \text{ claim } [SEP] \text{ title1 } [SEP] \text{ evi1 } [SEP] \\ \text{ title2 } [SEP] \text{ evi2 } [SEP] \dots [SEP] \quad (6)$$

The concatenation is also encoded by BERT encoder. Then, similarly, we obtain claim, title or evidence representation from the hidden state of the first token. Finally, for each evidence, we obtain its inter-document representation through alignment with the claim representation and its corresponding title representation.

3.2 Dual Evidence Fusion Graph

Through multi-view contextual encoding, for each evidence, we can obtain two alignment representations from different contextual views. To further integrate multi-view evidence information to handle multi-hop problem, inspired by multi-relational graph convolutional network (Cao et al., 2019; Tu

et al., 2019, 2020), we propose dual evidence fusion graph network. As shown in Figure 3, *one evidence sentence* corresponds to *two different nodes* in this graph, whose initial representations respectively come from intra-document encoding and inter-document encoding. For each evidence sentence, the noun phrases and named entities are extracted as keywords through spaCy¹ tool. Then for a pair of nodes, the links between them are decided according to following rules:

- **Common Document** Two nodes are linked if they come from the same document.
- **Common Keyword** Two nodes are linked if they share overlapped keywords.
- **Claim Jump** Two nodes are linked if they respectively share overlapped keywords with the claim.

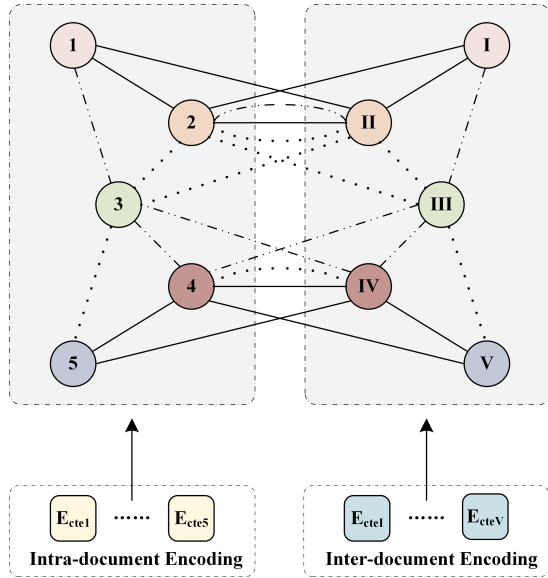


Figure 3: Dual Evidence Fusion Graph Network. Node 1 and node I denote different representations of the same evidence sentence from different encoding methods, similarly for node 2 and node II, etc. We define three kinds of edges in total.

For each claim, N selected evidence sentences introduce $2N$ evidence nodes. Let $H_i \in \mathcal{R}^{2N \times d}$ denotes the node representations at i -th graph layer, where d refers to the hidden dimension. The initial representation H_0 is the claim-title-evidence alignment representation through multi-view contextual encoding. Updated information $U_i \in \mathcal{R}^{2N \times d}$ after a single graph layer is defined as :

$$U_i = H_i W_0 + \sum_{j=1}^3 \tilde{A}_j H_i W_j \quad (7)$$

¹<https://spacy.io/>

where $\tilde{A}_j \in \mathcal{R}^{2N \times 2N}$ denotes corresponding row normalized adjacent matrix for different kinds of edges. Then the forget ratio $G_i \in \mathcal{R}^{2N \times d}$ between the updated and old information is:

$$G_i = \text{Sigmoid}(W_g[U_i, H_i]) \quad (8)$$

And the updated evidence representation through the graph layer is:

$$H_{i+1} = \text{Activation}(U_i) \odot G_i + H_i \odot (1 - G_i) \quad (9)$$

In this way, with several stacked layers, evidence representations are updated and multi-view evidence information is fused.

3.3 Confidence Aggregation

Aggregation aims to combines the evidence representations for final inference representation to verify the claim. Among the selected evidence sentences of a claim, some are groundtruth ones while others are not. To utilize evidence label information to enhance fact verification, like Tu et al. (2020), we adopt confidence aggregation.

Formally, let $H_k \in \mathcal{R}^{2N \times d}$ denotes evidence representations at the last graph layer. The confidence score of j -th evidence node \hat{y}_j is obtained from its representation H_k^j :

$$\hat{y}_j = \text{Sigmoid}(\text{MLP}(H_k^j)) \quad (10)$$

The final inference representation for fact verification R_v is the weighted sum of the evidence representations, where the weights are corresponding confidence scores:

$$R_v = \sum_{j=1}^{2N} \hat{y}_j H_k^j \quad (11)$$

and the fact verification result is obtained through a 3-way classification network:

$$\hat{v} = \text{Softmax}(W R_v + b) \quad (12)$$

The total loss consists of the binary cross entropy loss of evidence confidence, and the cross entropy loss of 3-way fact verification:

$$\mathcal{L}_{II} = \text{BCE}(y, \hat{y}) + \text{CE}(v, \hat{v}) \quad (13)$$

Here y is the evidence sentence label, 1 for groundtruth evidence sentences and 0 for non-evidence sentences. Besides, v is the fact verification label.

4 Experiment

4.1 Dataset

We conduct our experiments on FEVER 1.0 shared task (Thorne et al., 2018), which consists of 185,455 annotated claims with 5,416,537 Wikipedia documents from the June 2017 dumps. We adopt the original dataset split of the task, which includes a training set, a development set and an online blind test set. The detailed information is shown in Table 1.

Split	SUPPORTS	REFUTES	NEI	Total
train	80035	29775	35639	145449
dev	6666	6666	6666	19998
test	6666	6666	6666	19998

Table 1: Statistics information of FEVER 1.0 Shared Task.

Moreover, for a claim, there exist several groups of evidences, and each group itself is enough to independently verify the claim. To further study the impact of multi-view contextual information, we conduct a refined split on the development set. Specifically, samples of the development set can be divided into 5 parts and the ratio of different parts are displayed in Table 2:

- **Single.** All evidence groups contain exactly one sentence.
- **Single+.** At least one evidence group contains only one sentence, and at least one group contains multi sentences.
- **Multi.** All evidence groups contain exactly two sentences.
- **Multi+.** All evidence groups contain multi sentences, and at least one group contains more than two sentences.
- **NEI.** The sample is labeled as NEI with no evidence groups annotated.

Single	Single+	Multi	Multi+	NEI
56.87	3.78	5.03	0.99	33.33

Table 2: Ratio of different parts on the development set.

4.2 Experiment Setup

Our IMCI is implemented through Pytorch 1.2.0 and our experiments are conducted on a computation node with 4 NVIDIA Titan V GPU. Pre-trained BERT (Devlin et al., 2019) encoder is employed for all experiments. We also try RoBERTA

encoder (Liu et al., 2019) for fact verification. For the claims, we set max length as 64, and claims longer than this will be truncated. For the encoders, we set max input sequence length as 512, and sequence longer than this will be split with stride window size of 128. We utilize BERTAdam optimizer with initial learning rate of 1e-5 and warmup ratio of 0.1. For sentence retrieval, we adopt mini batch size of 4 and gradient accumulation step of 8. In each iteration, we train 2 epochs and select top 5 sentences as candidate evidences. For fact verification, we adopt mini batch size of 1 and gradient accumulation step of 32. For dual evidence graph, we stack 3 graph layers, where the hidden dimension is the same as that of the encoder. In each condition, we randomly start 4 times, train 4 epochs, and choose model parameters with the best performance on the development set.

4.3 Evaluation Metric

We adopt FEVER score as the dominant evaluation metric, which is the officially chief metric. FEVER score requires that fact verification label is correctly predicted, and at least one complete group of evidence sentences is found for SUPPORTS and REFUTES samples. The second important metric is label accuracy. For document retrieval and sentence retrieval, we take precision, recall as well as F1 into account. Here, we attach more importance to recall according to the task setting.

5 Results

5.1 Main Results

Main results on the blind test set are shown in Table 3. With multi-view contextual information joined, our ICMI framework obtains FEVER score of 70.10% and label accuracy of 73.04% with BERT_{base} encoder. The performance is comparable and even slightly promoted compared with the state-of-the-art one among all baselines with BERT_{base} encoder. Moreover, our model with RoBERTA_{base} encoder obtains FEVER score of 72.97% and label accuracy of 75.84%, and shows even higher performance than several baselines with large encoder. Furthermore, our model with RoBERTA_{large} encoder obtains FEVER score of 73.96% and label accuracy of 77.25%, and significantly outperforms all baselines. These indicate that our framework has made great progress to conduct more accurate fact extraction and verification.

Model	LA	FEVER
UKP-Athene(2018)	65.46	61.58
QFE(2019)	69.30	61.80
NSMN(2019a)	68.16	64.23
GEAR-BERT _{base} (2019)	71.60	67.10
SR-MRS-BERT _{base} (2019b)	72.56	67.26
DeSePtion-BERT _{base} (2020)	72.47	68.80
Transformer-XH-BERT _{base} (2020)	72.39	69.07
KGAT-BERT _{base} (2020)	72.81	69.40
CorefBERT-BERT _{base} (2020)	72.88	69.82
HESM-ALBERT _{base} (2020)	73.25	70.06
HESM-BERT _{base} (2020)	73.18	70.07
<i>ours</i> IMCI-BERT _{base}	73.04	70.10
KGAT-BERT _{large} (2020)	73.61	70.24
CorefBERT-BERT _{large} (2020)	74.37	70.86
HESM-ALBERT _{large} (2020)	74.64	71.48
KGAT-RoBERTa _{large} (2020)	74.07	70.38
CorefBERT-RoBERTa _{large} (2020)	75.96	72.30
<i>ours</i> IMCI-RoBERTa _{base}	75.84	72.97
<i>ours</i> IMCI-RoBERTa _{large}	77.25	73.96

Table 3: Overall performance on the online blind test det. FEVER is the officially chief score. LA denotes label accuracy.

5.2 Document Retrieval

Document retrieval results of different iterations on the development set are displayed in Table 4. With search engine power adopted through Wikipedia search, and statistical word frequency information joined through TF-IDF retrieval, coarse document retrieval obtains the highest recall of 92.77%. Besides, with dense semantic retrieval model guided and top one evidence hyperlinks joined, refined document retrieval shows even higher recall of 95.69%. Furthermore, refined document retrieval has much higher precision of 29.90% and F1 of 45.56%, respectively obtains 21.22% and 29.69% absolute increase than coarse retrieval. Therefore, our iterative fact extraction model has made great improvement on document retrieval.

Moreover, document retrieval results on different parts of the development set are displayed in Figure 4. Coarse document retrieval can handle Single and Single+ samples, where the recall are respectively as high as 97.56% and 98.02%. However, coarse document retrieval fails to handle multi-hop samples, where the recall of Multi and Multi+ samples are both pretty low, respectively 46.82% and 31.31%. Compared to coarse document retrieval, refined document retrieval shows compara-

Model	P	R	F1
UKP-Athene(2018)	-	90.32	-
NSMN(2019a)	51.04	89.23	64.94
GEAR-BERT _{base} (2019)	-	89.99	-
SR-MRS-BERT _{base} (2019b)	18.11	92.03	30.27
Coarse Retrieval	8.68	92.77	15.87
Refined Retrieval	29.90	95.69	45.56

Table 4: Document retrieval results on the development set. - denotes that the item is not available.

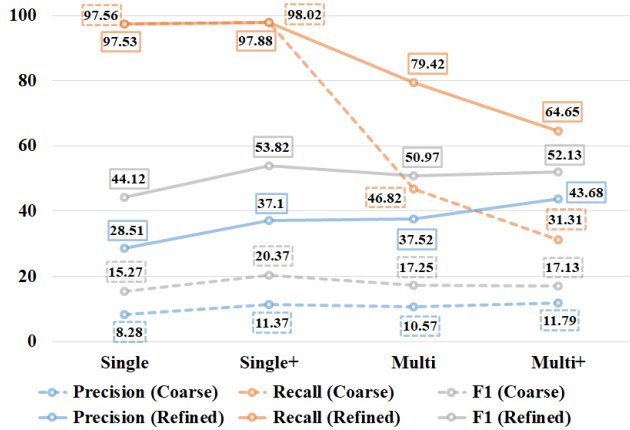


Figure 4: Document retrieval results on different parts of the development set. NEI samples are not taken into consideration since no evidence sentences are annotated for them.

ble recalls but much higher precision and F1 score on Single and Single+ samples. Furthermore, refined document retrieval makes great progress on multi-hop samples. For Multi and Multi+ samples, refined document retrieval respectively achieves 32.60% and 33.34% absolute increase on recall. Besides, the precision and F1 score also get significantly improved. These results indicate the high efficiency of our iterative multi-view fact extraction model. However, although refined document retrieval has achieved significant improvement, the recall of multi-hop samples is still far lower than single-hop ones.

5.3 Sentence Retrieval

Sentence retrieval results on the development set are summarized in Table 5. Our IMCI framework obtains the highest recall of 92.86%, and significantly outperforms all baselines.

For sentence retrieval of iteration 1, upstream coarse document retrieval obtains an extremely low precision of 8.68% (in Table 4). Thus, for each claim, on average our sentence retrieval model is re-

Model	P	R	F1
UKP-Athene(2018)	-	86.24	-
NSMN(2019a)	36.49	86.79	51.38
GEAR-BERT _{base} (2019)	24.08	86.72	37.69
SR-MRS-BERT _{base} (2019b)	44.47	86.60	58.77
HESM-BERT _{base} (2020) [#]	-	90.50	-
Iteration 1	25.31	90.30	39.54
w.o. Alignment	24.86	90.16	38.97
Iteration 1 (Top 1)	86.11	78.08	81.90
Iteration 2	25.90	91.98	40.42
IMCI [#]	25.74	92.86	40.30

Table 5: Sentence retrieval results on the development set. According to the original task setup, we keep top 5 sentences as evidence for each claim. # means the models adopt iterative sentence retrieval. *w.o.* means without the item.

quested to distinguish top 5 sentences as candidate evidences from more than 250 sentences. In this condition, with intra-document contextual information joined, the model obtains pretty high recall of 90.30%, and shows comparable performance with state-of-the-art iterative sentence retrieval model HESM (Subramanian and Lee, 2020). Besides, top one candidate evidences show a pretty high precision of 86.11%. This shows the importance of intra-document context, and is the base of refined document retrieval. Besides, the high precision also guarantees that top one candidate evidences can be considered as inter-document context in sentence retrieval of iteration 2. With multi-view contextual information joined, our sentence retrieval model of iteration 2 obtains even higher recall of 91.98%. Moreover, full pipeline reranking makes the recall get far more increase to 92.86%. These show the great power of multi-view contextual information on fact extraction.

Moreover, sentence retrieval results on different parts of the development set are displayed in detail in Table 6. For Single and Single+ samples, the recall are respectively high at 96.33% and 95.90%, while the precision and F1 score are pretty low. However, the recall of Multi samples is pretty low at 64.41%, while that of Multi+ samples is far lower at 26.26%. Therefore, taking these and the document retrieval results in Figure 4 into consideration, it seems that fact extraction for multi-hop samples is still a difficult problem, although our model has made several progress. A main concern is that bidirectional information interaction between multi-hop evidences may be not guaran-

teed during fact verification.

Part	P	R	F1
Single	23.00	96.33	37.14
Single+	52.06	95.90	67.48
Multi	32.98	64.41	43.63
Multi+	45.25	26.26	33.23

Table 6: Sentence retrieval results on different parts of the development set. NEI samples are not taken into consideration since no evidence sentences are annotated for them.

5.4 Fact Verification

Fact verification results on the development set are shown in Figure 5. With multi-view contextual information joined, our IMCI framework obtains the highest label accuracy of 75.83% and the highest FEVER score of 73.21%. When ignoring intra-document encoding, the label accuracy and FEVER score suffer severe decrease to 74.23% and 71.58%. This indicates the great importance of intra-document contextual information on fact verification. However, compared to intra-document encoding, inter-document encoding and dual evidence fusion graph have relatively weak influence on the performance. These two components mainly aim to handle multi-hop samples. However, multi-hop samples take pretty low ratio (about 6.02% in total in Table 2). Even worse, multi-hop samples have suffered serious performance damage on upstream fact verification task (in Table 6). Evidence confidence aggregation also makes some contribution, indicating the influence of evidence label information. Besides, we also study the influence of fact verification. It seems that progress on fact verification mainly contributes to FEVER score while shows weak influence on label accuracy.

Moreover, for our IMCI framework, the statistic information of prediction errors on fact verification is shown in Figure 6. The framework can correctly distinguish SUPPORTS and REFUTES examples, since SUPPORTS (REFUTES) and REFUTES (SUPPORTS) errors respectively take about 3.66% and 11.51%. This may indicate that the logical boundary between SUPPORTS and REFUTES is relatively clear. Besides, the framework hardly mistakes SUPPORTS examples for NEI examples. However, it may be difficult for the framework to distinguish REFUTES examples from NEI examples, as well as NEI examples from non-NEI examples, for REFUTES (NEI), NEI (SUPPORTS), and

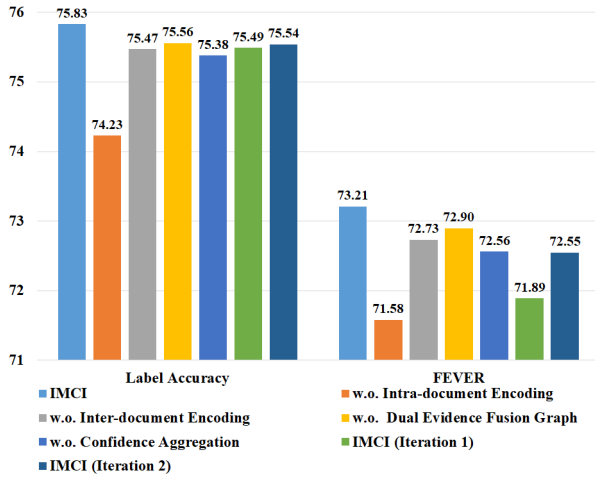


Figure 5: Fact verification results on the development set. These are averaged results on 4 random starts. *w.o.* means without the item.

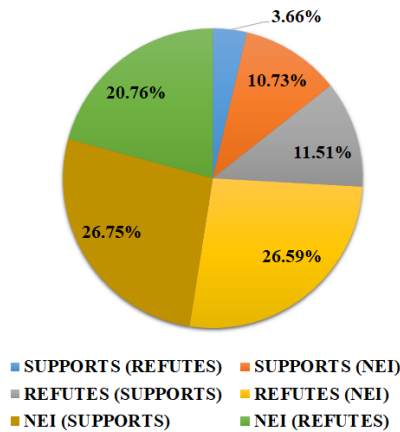


Figure 6: Statistic information of prediction errors on fact verification. (Label **out** of brackets denotes groundtruth, while label **in** brackets denotes wrong prediction.)

NEI (REFUTES) errors respectively take 26.59%, 26.75%, and 20.76%. The situation may be due to the pretty unbalanced label distribution of the training set (in Table 1). Besides, NEI may contain more complex logic semantic than the other two categories.

6 Related Work

• **Fake News Detection** Fake news detection has been attracting more attention. Ruffo et al. (2021) give a detailed survey about the development of this field. Textual-content based methods (Giachanou et al., 2019; Ghanem et al., 2020; Kaliyar et al., 2021) aim at understanding the linguistic and semantic information in the text to detect fake news.

User-role based methods (Vo and Lee, 2019; Giachanou et al., 2020) pay more attention to the role of users in the propagation of fake news. Multi-modal approaches (Zlatkova et al., 2019; Fung et al., 2021) involve multi-modal information, i.e. text, table, knowledge base, image, speech and video, to evaluate the credibility of news. Besides, bots and trolls aim at influencing users with commercial, political or ideological purposes by spreading disinformation deliberately. The detection of them (Stella et al., 2018; Sayyadharikandeh et al., 2020) is also an important direction. Moreover, Sheng et al. (2022) recently propose news environment perception for fake news detection, which focus on the background environment of fake news.

• **Fact Extraction** Fact extraction includes document retrieval and sentence retrieval. For document retrieval, Hanselowski et al. (2018) propose a constituency-based Wikipedia search model. Nie et al. (2019a) utilize a keyword matching model based on a quick string matching algorithm *Flash-Text* (Singh, 2017). Nie et al. (2019b) further adopt a combination model of keyword match and TF-IDF retrieval.

For sentence retrieval, Hanselowski et al. (2018), Nie et al. (2019a), and Zhou et al. (2019) respectively modify Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017). These models separately encode the claim and an evidence sentence, and adopt cross-attention mechanism to accomplish information interaction between the claim and the evidence sentence. Nie et al. (2019b) and Liu et al. (2020) adopt BERT-based model. Subramanian and Lee (2020) propose iterative fact verification models to retrieve evidence sentences and combine evidence sets.

• **Fact Verification** For fact verification, Nie et al. (2019b) concatenate the claim and the evidence sentences into a sequence as input to BERT encoder, and take the hidden state of the first special token [CLS], as final inference representation. Zhou et al. (2019) adopt graph neural network for evidence aggregating and reasoning. Zhong et al. (2020) introduce semantic role information to construct refined graph, and adopt graph convolutional network to handle the task. Liu et al. (2020) propose fine-grained kernel-based graph attention network for information interaction between the claim and the evidences. Subramanian and Lee (2020) propose to combine evidence sets during fact extraction, and conduct fact verification on evidence sets. Si et al.

(2021) introduce topic model and stance detection model, and study the influence of topic and stance information on fact verification.

7 Conclusion

In this paper, we propose to integrate multi-view contextual information for fact extraction and verification. Our experimental results show that our IMCI model can obtain state-of-the-art performance on the task. Moreover, the ablation study results indicate that multi-view contextual information is essential for both fact extraction and fact verification. In the future, we will explore much stronger model to utilize contextual information in a more efficient way.

References

- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. [Question answering by reasoning across documents with graph convolutional networks](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2306–2317. Association for Computational Linguistics.
- Qian Chen, Xiaodan Zhu, ZhenHua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. [Enhanced lstm for natural language inference](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1657–1668. Association for Computational Linguistics.
- Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.
- Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, ShihFu Chang, Kathleen R. McKeown, Mohit Bansal, and Avi Sil. 2021. [Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1683–1698. Association for Computational Linguistics.
- Bilal Ghanem, Paolo Rosso, and Francisco M. Rangel Pardo. 2020. [An emotional analysis of false information in social media and news articles](#). *ACM Transactions on Internet Technology*, 20(2):19:1–19:18.
- Anastasia Giachanou, Esteban ARissola, Bilal Ghanem, Fabio Crestani, and Paolo Rosso. 2020. [The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers](#). In *Proceedings of the 25th International Conference on Applications of Natural Language to Information Systems*, Lecture Notes in Computer Science, pages 181–192. Springer.
- Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. [Leveraging emotional signals for credibility detection](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 877–880. ACM.
- Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. [Ukp-athene: Multi-sentence textual entailment for claim verification](#). *CoRR*, abs/1809.01479.
- Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona T. Diab, and Smaranda Muresan. 2020. [Deseption: Dual sequence prediction and adversarial examples for improved fact-checking](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606. Association for Computational Linguistics.
- Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. [Fakebert: Fake news detection in social media with a bert-based deep learning approach](#). *Multimedia Tools and Applications*, 80(8):11765–11788.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. [Fine-grained fact verification with kernel graph attention network](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351. Association for Computational Linguistics.
- Giovanni Da San Martino, Stefano Cresci, Alberto Baron Cedenno, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. [A survey on computational propaganda detection](#). In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, pages 4826–4832. ijcai.org.
- Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. [Combining fact extraction and verification with neural semantic matching networks](#). In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 6859–6866. AAAI Press.
- Yixin Nie, Songhe Wang, and Mohit Bansal. 2019b. [Revealing the importance of semantic retrieval for](#)

- machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2553–2566. Association for Computational Linguistics.
- Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. [Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2335–2345. Association for Computational Linguistics.
- Giancarlo Ruffo, Alfonso Semeraro, Anastasia Giachanou, and Paolo Rosso. 2021. [Surveying the research on fake news in social media: a tale of networks and language](#). *CoRR*, abs/2109.07909.
- Mohsen Sayyadiharikandeh, Onur Varol, KaiCheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. [Detection of novel social bots by ensembles of specialized classifiers](#). In *Proceedings of 29th ACM International Conference on Information and Knowledge Management*, pages 2725–2732. ACM.
- Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. [Zoom out and observe: News environment perception for fake news detection](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4543–4556. Association for Computational Linguistics.
- Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. [Topic-aware evidence reasoning and stance-aware aggregation for fact verification](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 1612–1622. Association for Computational Linguistics.
- Vikash Singh. 2017. [Replace or retrieve keywords in documents at scale](#). *CoRR*, abs/1711.00046.
- Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. [Bots increase exposure to negative and inflammatory content in online social systems](#). *Proceedings of the National Academy of Sciences*, 115(49):12435–12440.
- Shyam Subramanian and Kyumin Lee. 2020. [Hierarchical evidence set modeling for automated fact extraction and verification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7798–7809. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [Fever: a large-scale dataset for fact extraction and verification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819. Association for Computational Linguistics.
- Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. [Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents](#). In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 9073–9080. AAAI Press.
- Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. [Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 2704–2713. Association for Computational Linguistics.
- Nguyen Vo and Kyumin Lee. 2019. [Learning from fact-checkers: Analysis and generation of fact-checking language](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344. ACM.
- Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. [Coreferential reasoning learning for language representation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7170–7186. Association for Computational Linguistics.
- Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. [Transformer-xh: Multi-evidence reasoning with extra hop attention](#). In *Proceedings of the 8th International Conference on Learning Representations*. OpenReview.net.
- Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. [Reasoning over semantic-level graph for fact checking](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180. Association for Computational Linguistics.
- Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. [Gear: Graph-based evidence aggregating and reasoning for fact verification](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 892–901. Association for Computational Linguistics.
- Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. [Fact-checking meets fauxtography: Verifying claims about images](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2099–2108. Association for Computational Linguistics.