# UNIFEE: Unified Evidence Extraction for Fact Verification

**Nan Hu**[♠*], **Zirui Wu**[♠*], **Yuxuan Lai**[♡♠], **Chen Zhang**[♠], **Yansong Feng**[♠◇†]
[♠]Wangxuan Institute of Computer Technology, Peking University, China
[◇]The MOE Key Laboratory of Computational Linguistics, Peking University, China
[♡]Department of Computer Science, The Open University of China
{hunan,ziruiwu,erutan,zhangch,fengyansong}@pku.edu.cn
laiyx@ouchn.edu.cn

## Abstract

FEVEROUS is a fact extraction and verification task that requires systems to extract evidence of both sentences and table cells from a Wikipedia dump, then predict the veracity of the given claim accordingly. Existing works extract evidence in the two formats separately, ignoring potential connections between them. In this paper, we propose a Unified Evidence Extraction model (UNIFEE), which uses a mixed evidence graph to extract the evidence in both formats. With the carefully-designed unified evidence graph, UNIFEE allows evidence interactions among all candidates in both formats at similar granularity. Experiments show that, with information aggregated from related evidence candidates in the fusion graph, UNIFEE can make better decisions about which evidence should be kept, especially for claims requiring multi-hop reasoning or a combination of tables and texts. Thus it outperforms all previous evidence extraction methods and brings significant improvement in the subsequent claim verification step.

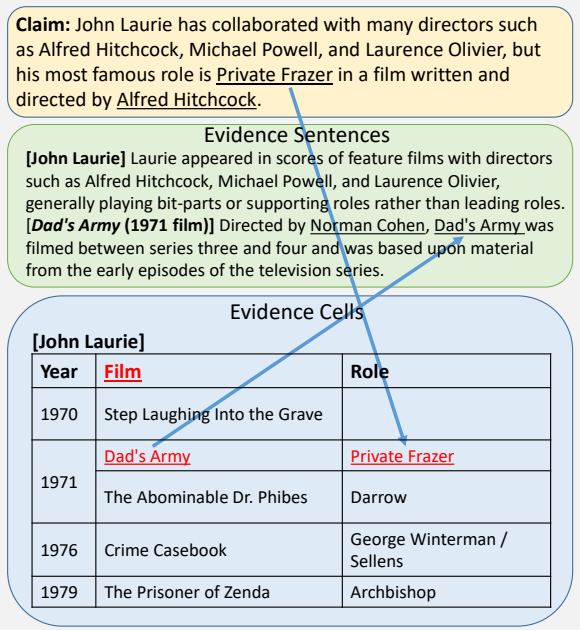


Figure 1: An example in FEVEROUS. The green rectangle shows the gold evidence sentences, while the blue one is the gold evidence table, where the gold evidence cells are underlined.

## 1 Introduction

FEVEROUS (Aly et al., 2021) is a fact extraction and verification task over structured and unstructured information. Models should extract fine-grained evidence in two formats, namely, sentences and table cells, from a Wikipedia dump and verify the given claim accordingly.

Many previous works focus on the claim verification procedure. They propose various methods to fuse evidence in different formats while leaving the evidence extraction within each format separately (Kotonya et al., 2021; Malon, 2021; Bouziane et al., 2021; Hu et al., 2022). For those efforts on evidence extraction, Saeed et al. (2021) use a neural re-ranker to refine page retrieval. Bouziane et al. (2021) propose a reinforced adaptive retrieval embedding paradigm, but they ignore the possible connections among evidence either in the same format or across different formats. However, there are many gold evidence pieces with fewer lexical or semantic overlaps with the claim, while their necessity is mainly determined by other evidence. Without considering evidence connections during extraction, models may be prone to miss these evidence candidates, thus, propagating errors to subsequent claim verification.

Figure 1 shows an example from FEVEROUS. To refute the given claim, models should find that *Private Frazer* is a role in the film *Dad's Army*, and recognize that its director is *Norman Cohen* instead of *Alfred Hitchcock*. However, the overlap between the claim and the required evidence sentence *Directed by Norman Cohen ...* is only two words:

[*] The first author and the second author contribute equally to this work.
[†] Corresponding Author.



*Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1150–1160
May 2-6, 2023 

*directed by*, without any key entities, which is generally not strong enough to recognize its usefulness. With the evidence cells in the table, one can find that *Private Frazer* is a role in the film *Dad's Army*, which links the evidence sentence and the given claim. Therefore, for the evidence extraction step it is essential to encourage interactions between evidence pieces of different formats.

Meanwhile, evidence in the same format also provides context information to each other, especially for table cells. In Figure 1, the cell *Dad's Army* is not mentioned in the claim, but the column header, *Film*, and the cell in the same row, *Private Frazer*, are strong clues appearing in the given claim, which implies that the cell *Dad's Army* is highly likely to be useful evidence. We argue that models should allow the evidence to accumulate context information from all closely-associated evidence in both formats to extract comprehensive evidence sets for claim verification.

In this paper, we propose a Unified Evidence Extraction model (UNIFEE), which exploits graph attention networks (GAT) to encourage interactions among connected candidates during evidence extraction. We design a novel graph structure, which accommodates all evidence candidates and introduces column nodes to allow evidence candidates of different formats to interact with each other in a similar granularity. This graph also facilitates the interactions among cells in several nested tables with better representation of the table layout. Compared to previous flatten-based methods, our method exploits the structural relationship among cells in complicated tables and enhances the interactions between evidence in both formats.

Experiments on FEVEROUS show that our UNIFEE improves the evidence extraction performances and further boosts the final fact verification scores significantly. The ablation experiments and case study demonstrate the effectiveness of the proposed evidence extraction method and our graph designs. Our code is released to the public for further exploration. [1].

Our contributions can be summarized as follows:

• We propose a novel evidence extraction method, featuring a mixed graph structure with carefully designed column nodes and layout-aware table representations. Our method enables early and thorough evidence interactions within one format or across formats at a similar granularity.

• Experiments show that our method outperforms previous works by a large margin on both the evidence extraction and the final verification results. Thorough analysis shows that our model can recall more multi-hop evidence and also evidence with fewer lexical overlaps with the given claim.

## 2 Our Model

We take the widely-adopted three-step pipeline for FEVEROUS (Aly et al., 2021), which consists of retrieving pages from Wikipedia dump, extracting sentences and table cells as evidence from the pages, and predicting the veracity label of the given claim based on the evidence set. In this work, we focus on the second step, i.e., evidence extraction. We use the DRQA (Chen et al., 2017) based document retrieval method and the dual-channel verification method proposed by Hu et al. (2022) for the remaining two steps.

Formally, during evidence extraction, $n'_s$ sentences and $n_t$ tables are extracted from $n_p$ retrieved passages by a pair-wise classification model firstly. This pre-processing step narrows the search space of sentences and table cells from thousands to a few dozens, which makes the mixed graph lighter. Then, $n_c$ cells and $n_s$ sentences are extracted by our proposed Unified Evidence Extraction model (UNIFEE). These fine-grained evidence pieces are used for fact verification later.

In UNIFEE, we design a mixed evidence graph (§2.1) to organize evidence of two formats. Then, a graph neural network is applied to the graph to accumulate contextualized information from all evidence candidates with similar granularity, and identify useful evidence from a unified perspective (§2.2).

### 2.1 Mixed Evidence Graph

There are three kinds of nodes in our mixed evidence graph. Sentence nodes and cell nodes stands for sentences and cells, e.g., *FC Ararat Yerevan* in table 1, evidence candidates. Besides, we further introduce a novel node type representing table columns , e.g., *Senior career* Apps 10 164 8 26* in table 1. The intuition is that textual evidence is mainly at the sentence level but table cells are generally words or phrases. The unbalanced granularities will hinder their interactions. Therefore, we introduce this column node, to act as a mediator for information transmission between the sentence nodes and the cell nodes.

[1] https://github.com/WilliamZR/unifee

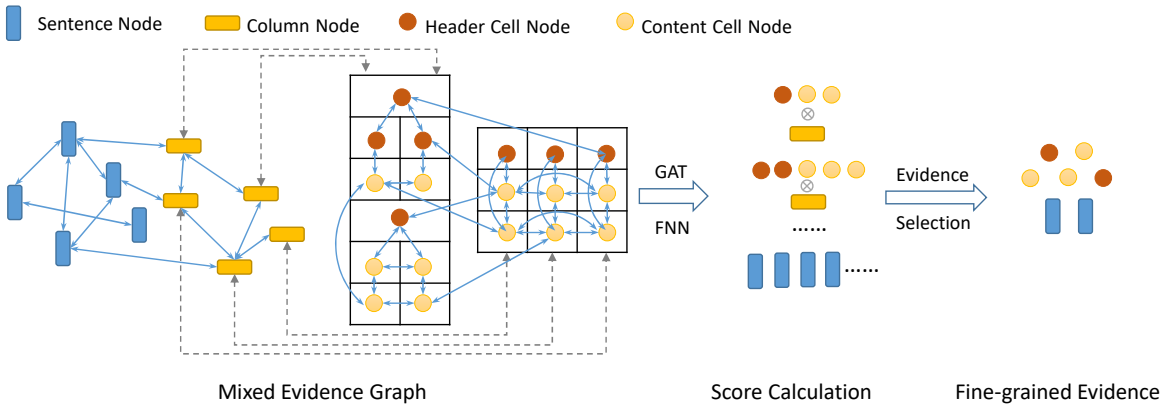


Figure 2: The overview of UNIFEE. The blue lines are edges in the mixed evidence graph, and the gray dash lines indicate which column each column node comes from .

Specifically, the mixed evidence graph $G$ has three types of nodes: sentence nodes, cell nodes (including header cells and content cells), and column nodes, as shown in Figure 2.

The edge connections are elaborated as follows:

**Edges among Sentence Nodes and Column Nodes** We add bi-directed edges between two nodes $v_i$ and $v_j$ for $v_i, v_j \in V^S \cup V^{Col}$ if and only if the two following conditions:

- $v_i$ and $v_j$ represent evidence in the same Wikipedia page.
- $v_i$ and $v_j$ contain common entities or hyperlinks.

Here the text representation of a column node simply is the concatenation of all cells in that column.

**Edges among Cell Nodes** Although we expect direct messaging between cells, simply connecting cells in the same rows or the same columns is not enough. The tables in Wikipedia pages are often organized as complicated nested tables. Previous works simply flattening a table or converting cells to strings with human-designed templates will lose delicate structural information layout and cannot reflect the cross-table connections.

Meanwhile, we notice that the tables often have large headers in the middle of themselves in reality, which separates the influenced columns into several parts, and, in this case, cells in the same column could be irrelevant. Disconnecting these cells in the graph will help us remove confusing edges and make the information transmission more reasonable.

Table 1 is a table from the Wikipedia page *Aramais Yepiskoposyan*. We can see that three long headers *Personal Information*, *Senior career* and *National team* cut the table into three parts. The cell organization and contents within each part are consistent while differing a lot across different parts. For example, although the header cell *Personal Information* and the content cell *FC Ararat Yerevan* are in the same column, they do not have a subordinate relationship.

| Personal information | | | |
|---|---|---|---|
| Date of birth | 27 September 1968 (age 53) | | |
| Height | 1.75 m (5 ft 9 in) | | |
| Position(s) | Midfielder | | |
| Senior career* | | | |
| Years | Team | Apps | (Gls) |
| 1986–1991 | FC Ararat Yerevan | 10 | (0) |
| 1992–1997 | FC Chernomorets Novorossiysk | 164 | (8) |
| 1997 | FC Kuban Krasnodar | 8 | (0) |
| 1999–2000 | FC Irtysh Pavlodar | 26 | (2) |
| 2000 | FC Spartak Anapa (amateur) | | |
| National team | | | |
| 1997 | Armenia | 1 | (0) |

Table 1: A table instance from the Wikipedia.

We believe that a graph deliberately connecting cells in all table candidates can better model the table structures and connections. According to the table layout and lexical features, we connect two cell nodes $v_i^C$ and $v_j^C$ by bi-directional edges:

- If $v_i^C$ and $v_j^C$ are two content cells in the same column and there is no table-header-like cell between them. For example, the header cells *Senior Career** and *Team* separate the content cells *Midfielder* and *FC Ararat Yerevan*.
- If $v_i^C$ and $v_j^C$ are two cells in the same column, $v_i^C$ is a header cell above $v_j^C$, and no header cell exists between them.
- If $v_i^C$ and $v_j^C$ are two cells in the same row.
- If $v_i^C$ and $v_j^C$ are two cells sharing at least one entity or hyperlink in their contents.

## 2.2 Evidence Extraction with Mixed Graphs

In this section, we will show how we extract the evidence set with the mixed evidence graph, which encourages evidence to accumulate context information from closely-associated candidates of both formats and make a better decision to form a comprehensive and accurate evidence set for the verification step.

We formulate the task as: given a claim $Q$ and $n$ retrieved Wikipedia pages $P = \{p_i\}_{i=1}^{n}$, the models should extract a sentence evidence set $E = \langle S, C \rangle$, where $S$ are sentences in $P$ and $C$ are cells in $P$. The evidence set $E$ can be used to support or refute the claim $Q$.

**Node Initialization** The column nodes and cell nodes share a common embedding module, which is more efficient and can build an implicit connection between cells and columns apart from the explicit edges in the graph. Specifically, we feed the claim-table pairs into TAPAS (Herzig et al., 2020), a pre-trained table representation model, to get the embedding of cell nodes and column nodes. Each cell node is initialized by averaging the last hidden states of its tokens, $n(v_i^C)$. $\#\text{TOKENS}\,(cell_i)$ means the number of tokens in $cell_i$. $T$ means the table $cell_i$ belongs to.

$$\mathbf{n}(\mathbf{v_i^C}) = \frac{\sum_{t \in \text{TOKENS}(cell_i)} (\text{Tapas}_t(Q, T))}{\#\text{TOKENS}\,(cell_i)}$$

And each column node is initialized as the mean-pooling of the embeddings of all cells in this column $n(v_i^{Col})$. $\#\text{CELLS}\,(col_i)$ is the number of cells in the column $i$.

$$\mathbf{n}(\mathbf{v_i^{Col}}) = \frac{\sum_{t \in \text{CELLS}(col_i)} (\text{Tapas}_t(Q, T))}{\#\text{CELLS}\,(col_i)}$$

A one-layer GAT is applied to the graph, which updates the node representations by aggregating representations from their neighbors. $h_i^{(l)}$ is the current representation of node $i$. $W^f$, $\vec{a}$ and $W$ are training parameters.

$$h_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W^f h_j^{(l)}$$

$$\alpha_{ij}^{(l)} = \text{softmax}_\text{i}\left(e_{ij}^{(l)}\right)$$

$$e_{ij}^{(l)} = \text{LeakyReLU}\left(\vec{a}^T \left[W h_i^{(l)} \| W h_j^{(l)}\right]\right)$$

With a two-layer feed-forward network and a softmax layer, we obtain the probability of each fine-grained evidence candidate being retrieved. The evidence with a score above a predefined threshold is kept.

The loss is calculated as the weighted sum of the sentence loss $L^S$, column loss $L^{Co}$, and the cell loss $L^{Ce}$. We use cross entropy as the loss function for each part separately.

$$L = L^S + \alpha \cdot L^{Co} + \beta \cdot L^{Ce}$$

$$L^S = -\frac{1}{N^S} \sum_{i=1}^{N^S} \log\left(\text{p}\left(\hat{y}^S = y_i^S | Q, S, C\right)\right)$$

$$L^{Co} = -\frac{1}{N^{Co}} \sum_{i=1}^{N^{Co}} \log\left(\text{p}\left(\hat{y}^{Co} = y_i^{Co} | Q, S, C\right)\right)$$

$$L^{Ce} = -\frac{1}{N^{Ce}} \sum_{i=1}^{N^{Ce}} \log\left(\text{p}\left(\hat{y}^{Ce} = y_i^{Ce} | Q, S, C\right)\right)$$

where $N^S$, $N^{Co}$, and $N^{Ce}$ are the total numbers of sentence, column, and cell candidates in the training set, respectively. $\alpha$ and $\beta$ are the coefficients of $L^{Co}$ and $L^{Ce}$. If a cell or a sentence is in any of the gold evidence sets, we give it the label as 1, otherwise 0. For the labels of the column nodes, if any of the cells in that column is labeled as 1, this column is labeled as 1, otherwise 0.

The score of a sentence candidate is simply calculated as the softmax result of the label 1. In the training procedure, the cell score is calculated similarly. While in the evidence prediction step, the probability of a cell being retrieved is calculated as the product of the score of the cell node itself and the score of the column node it belongs to.

## 2.3 Document Retrieval and Claim Verification

For the two remaining steps in the whole pipeline, i.e., document retrieval and claim verification, we follow (Hu et al., 2022). For the document retrieval step, we retrieve evidence pages from the Wikipedia dump with a DrQA retriever (Chen et al., 2017) and re-rank the pages with a combination of BM25 and RoBERTa-base (Liu et al., 2019) re-rankers. For the verification step, we perform bi-directional evidence conversions and dual-channel evidence fusion to predict the final veracity label of the given claim.

## 3 Experiments

### 3.1 Evaluation Metrics

There are two main official metrics in the FEVEROUS task, i.e., accuracy and FEVEROUS score. The accuracy calculates the proportion of instances for which the model predicts a correct veracity label. The FEVEROUS score considers not only the correctness of the final veracity label but also the extracted evidence set. It calculates the proportion when the extracted evidence set covers one of the gold evidence sets and the predicted veracity label is consistent with the gold label.

Three additional official metrics are used to evaluate the quality of the extracted evidence sets, i.e., Evidence Precision, Evidence Recall, and Evidence F1. Multiple gold evidence sets are provided in FEVEROUS for each claim. A piece of extracted evidence is correct if and only if it is contained by any gold evidence set. For each instance, the Evidence Precision is the proportion of the correct predicted evidence. The overall Evidence Precision is the average score over all instances. The Evidence Recall is the proportion of instances with a correct extracted evidence set. An extracted evidence set is correct if and only if it covers any of the gold evidence sets. The Evidence F1 is the harmonic mean of the Evidence Precision and Evidence Recall.

### 3.2 Implementation Details

We retrieve $n_p = 5$ pages from the Wikipedia dump for each claim, and further narrow the evidence searching space to $n'_s = 5$ sentences and $n_t = 3$ tables with RoBERTa-base and DrQA model (same as the official baseline), separately. When we choose the final evidence set according to UNIFEE, the threshold for selecting cells and sentences is both 0.01. At most 25 cells are selected for each claim. As for the results, 12.1 cells and 4.7 sentences are extracted for each instance on average.

We use an Adam optimizer (Kingma and Ba, 2015) with a batch size of 4. The learning rates for pre-trained parameters and the others are $10^{-6}$ and $10^{-4}$, respectively. We train UNIFEE for 3 epochs, which takes 21 hours on a NVIDIA A40 GPU. We use Stanza toolkit[2] to identify entities in the claim and evidence for graph construction. We use base-size RoBERTa[3] and TAPAS[4] for sentence and table embedding in evidence extraction. The sentence encoder in claim verification is RoBERTa-large fine-tuned with NLI and verification tasks[5], same as the official baseline.

Details of the FEVEROUS dataset are shown in Appendix § A.

### 3.3 Baseline Models

The official baseline (Aly et al., 2021) uses DrQA (Chen et al., 2017) to extract sentences and tables. It uses pre-trained models to retrieve cells from flattened tables. EURECOM (Saeed et al., 2021) proposes a neural re-ranker to improve document retrieval. Both NCU (Gi et al., 2021) and Z team (Kotonya et al., 2021) linearize cells to unify evidence element format with sentences. NCU concatenates the claim and the evidence elements as input into RoBERTa for binary classification. Z team constructs a fully-connected graph of evidence element to extract evidence. A next-hop predictor is introduced by Papelo (Malon, 2021) to extract the required evidence. Sentences and cells are retrieved in multi-hops simply based on word overlap with imagined evidence. FaBULOUS (Bouziane et al., 2021) trains a Multi-Hop Dense Retriever (Xiong et al., 2020) to retrieve sentences and cells separately. DCUF(Hu et al., 2022) strengthens the official evidence extraction method with a multi-turn cell retriever.

### 3.4 Main Results

Our main results are summarized in Table 2. We obtain remarkable improvements over the best previous model, DCUF, in evidence extraction. In the development/test set, the improvements are 3.98%/3.56%, 11.86%/9.77%, and 5.96%/5.22% in evidence precision, recall, and f1 score, respectively. These results suggest that, with the context information accumulated from other candidates in the mixed evidence graph, our model can extract evidence more accurately. We also find that UNIFEE can easily drop irrelevant evidence and keep crucial evidence even though it has less word-overlapping with the claim. Specifically, compared with the official baseline, UNIFEE extracts 25% more gold evidence, with an average word overlap of 15.8%,

[2] https://github.com/stanfordnlp/stanza/

[3] https://huggingface.co/roberta-base

[4] https://huggingface.co/google/tapas-base

[5] https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

| Models | Development/Test | | | | |
|---|---|---|---|---|---|
| | FEVEROUS Score | Accuracy | Evidence Precision | Evidence Recall | Evidence F1 |
| Official Baseline | 19 / 17.73 | 53 / 48.48 | 12 / 10.17 | 30 / 28.78 | 17 / 15.03 |
| EURECOM | 19 / 20.01 | 53 / 47.79 | 12 / 13.73 | 29 / 33.73 | 17 / 19.52 |
| Z team | – / 22.51 | – / 49.01 | – / 7.76 | – / 42.64 | – / 13.12 |
| NCU | 29 / 25.14 | 60 / 52.29 | 10 / 9.91 | 42 / 39.07 | 17 / 15.81 |
| Papelo | 28 / 25.92 | 66 / 57.57 | – / 7.16 | – / 34.60 | – / 11.87 |
| FaBULOUS | 30 / 27.01 | 65 / 56.07 | 8 / 7.73 | 43 / 42.58 | 14 / 13.08 |
| UNIFEE* | **43.48 / 39.36** | **72.35 / 62.41** | **19.04 / 18.35** | **55.08 / 53.87** | **28.30 / 27.37** |
| DCUF | 35.77 / 33.97 | 72.91 / 63.21 | 15.06 / 14.79 | 43.22 / 44.10 | 22.34 / 22.15 |
| UNIFEE | **44.86 / 41.50** | **73.67 / 65.04** | **19.04 / 18.35** | **55.08 / 53.87** | **28.30 / 27.37** |

Table 2: Model performance on the development set and the test set. UNIFEE* means UNIFEE with the verification model of the official baseline instead of DCUF's.

far less than the 28.9% word overlap averagely in the gold evidence extracted by the official baseline.

Our model also achieves better performance on the fact verification metrics of FEVEROUS. Compared to previous state-of-the-art model DCUF, UNIFEE obtains an improvement of 9.09%/7.53% on the FEVEROUS score and 0.76%/1.83% on accuracy at the development/test set. Moreover, we experiment with a simpler verification model taken from the official baseline instead of that from DCUF to see the effectiveness of our evidence extraction step (UNIFEE*). Compared to Bouziane et al. (2021), Kotonya et al. (2021) and Hu et al. (2022) , who use complex verification models with template-based format conversion and(or) dual-channel setting, UNIFEE* still achieves better results. These results suggest that , with a high-quality evidence set, even a simple verification model can achieve good performance.

We also measure the improvement of our model on instances of different challenges in the dev set. Compared to the official baseline, UNIFEE shows an improvement of 65.93% (from 21% to 35%) on instances requiring multi-hop reasoning and 45.32% (from 41% to 59%) on instances requiring the combination of tables and texts, far above the average increase ratio 35.84%. It shows that with information accumulated from connected evidence candidates, UNIFEE can recall more required evidence within a limited evidence size.

Meanwhile, compared to the previous models using multi-turn retriever (Malon, 2021; Bouziane et al., 2021; Hu et al., 2022) for evidence extraction, our method does not require iterative running to retrieve multi-hop evidence and thus is less time-consuming.

### 3.5 Ablation Study

We evaluate the effectiveness of our Unified Evidence Extraction model with ablation experiments. (1)w/o Fusion Graph: We use the baseline proposed in Aly et al. (2021) to retrieve evidence based on our document retrieval results. (2) w/o Column Nodes: we connect sentences and cells directly if they have common hyperlinks or entities without column nodes as intermediaries. (3) w/o Cross-format Interactions: We deprecate sentence nodes and column nodes, only using cell nodes to retrieve cell evidence. (4)w/o Threshold: We retrieve top 5 sentences and top 25 cells instead of using a threshold.

The results are listed in Table 3. When we apply the evidence extraction model proposed in the official baseline (Aly et al., 2021) on our page retrieval results, the evidence precision, recall, and F1 drop by 3.15%, 14.54%, and 5.47%, respectively. It proves that our UNIFEE extracts the required evidence and removes confusing evidence candidates more confidently with context information accumulated from all evidence candidates of both formats. Meanwhile, the metrics drop a lot when we directly connect sentence candidates and cell candidates, with a decrease of 2.78% on evidence F1. It suggests that with column nodes as intermediaries, UNIFEE allows cross-format evidence interactions at a similar granularity, thus making the information flow over the graph more reasonable and efficient. In the w/o Cross-format Interactions setting, when we remove column nodes and do not add any edges between sentence nodes and cell nodes, the evidence F1 decreases to 25.89%. It shows that context information from evidence of another format is also crucial for the evidence extraction step.

| Models | P | R | F1 |
|---|---|---|---|
| UNIFEE | **19.04** | 55.08 | **28.30** |
| w/o Fusion Graph | 15.89 | 40.54 | 22.83 |
| w/o Column Nodes | 16.74 | 53.66 | 25.52 |
| w/o Cross-format Interactions | 16.98 | 54.51 | 25.89 |
| w/o Threshold | 12.48 | **56.22** | 20.42 |

Table 3: Ablation study regarding the retrieved evidence on the development set.

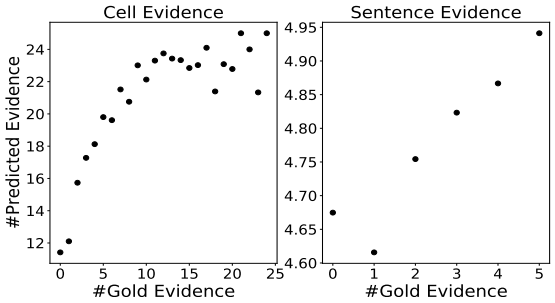


Figure 3: The relationship between the number of gold evidence and the average number of predicted cell/sentence evidence the evidence extraction step.

We notice that , without cross-format evidence interactions, UNIFEE still outperforms the official baseline by 2.94% and improves the evidence recall by 13.97%, which means that our carefully-designed edges between cells can better represent the layout of complex nested tables, thus improving the cell selection results.

From the w/o Threshold setting, we can see that if we, as many previous works do, directly select the top 25 cells and 5 sentences even whose scores are very low, there will be a slight improvement in the evidence recall, from 55.08% to 56.22%. However, it is at the expense of a 6.56% drop in evidence precision and a 7.88% drop in evidence F1. More irrelevant evidence is selected, which easily confuse the verification model in the final step. With the rough threshold in evidence selection, UNIFEE can dynamically select useful evidence according to their context-aware relevance regardless of the numbers and formats , which helps even simple verification models achieve better performance.

## 4 Analysis

### 4.1 Adaptive Evidence Selection

The number of sentences and cells required to verify claims in the FEVEROUS varies a lot. Some claim needs only sentence evidence, some only cell evidence, and many others a combination of both. We find that our UNIFEE can make an adaptive selection of evidence according to the given claim, without being bound to certain formats or a fixed evidence number. With joint evidence extraction in a mixed evidence graph and evidence selection with a threshold, UNIFEE keeps more helpful evidence in the fine-grained evidence retrieval step.

We analyze the relationship between evidence in the gold evidence set and that in the extracted evidence set. The results are shown in Figure 3.

We can see that UNIFEE extracts more cells for claims requiring more cell evidence. For claims that only need sentence evidence, UNIFEE extracts 11 cells on average, and for claims requiring 25 cells, it extracts cells to the maximum limit. We can observe a similar trend for sentence evidence. These results prove that UNIFEE introduces fewer noisy evidence pieces of unwanted formats than the previous works that set a fixed number of each format of evidence, while keeping the required evidence.

### 4.2 Positions of Cell Candidates

We find that the baseline cell extraction model is prone to select cells in the first few lines. If the cell evidence is at the bottom of a table, it is likely to be left out. A possible explanation is that when flattening the table into lines and applying a sequence tagging model to extract evidence cells, these cells are too far from their header cells, making it hard to capture the crucial context information. The average row position of the predicted cells in the baseline is 4.37 on average, much smaller than 6.71, the average of the gold ones. By contrast, it is easier for UNIFEE to exclude the interference of the cell positions thanks to its graph structure, where the cells in different lines are almost equivalent. Therefore, it can extract evidence cells at the bottom of the tables, and the average row position of its predicted cells increases to 6.39.

### 4.3 Case Study

A case is shown in Figure 4. Models are expected to find the cells *The Irish Times* and *4/5 stars* in the same row to refute the given claim *The Fine Art of Surfacing received a rating of 5/10 from The Irish Times*. The baseline model leaves out the cell *4/5 stars* since it has little word coverage with the given claim. Instead, it selects the distracting cell *3/10*. With the carefully-designed graph structure in UNIFEE, our model can obtain context information from the cell *The Irish Times* in the same row and the cell *Rating* in the same column. Meanwhile,

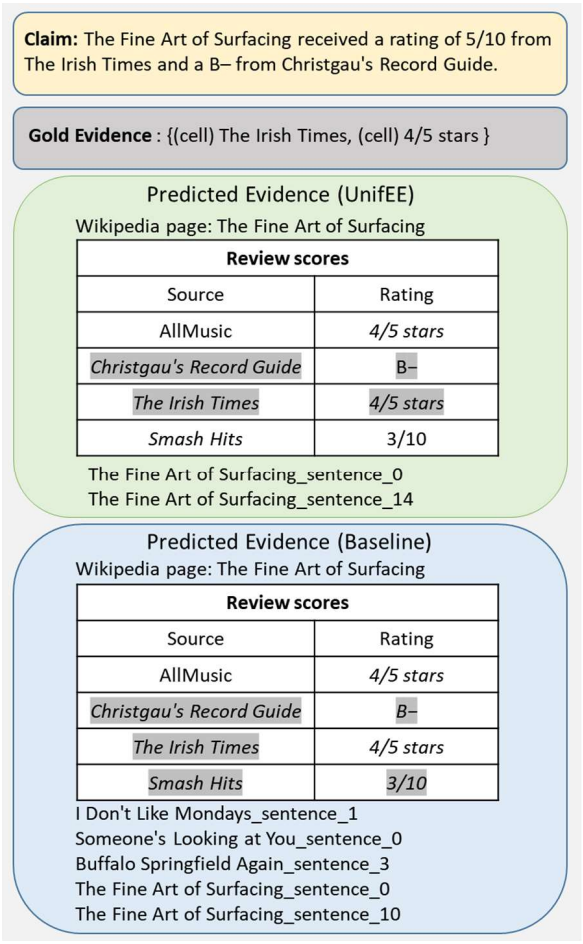


Figure 4: The extracted evidence set by UNIFEE and the official baseline. Cells with a gray background are the predicted ones.

the column where *4/5* is situated will be given a high score with the information from the key cell *Rating*, *B-* and even the distracting cell *3/10*. As a result, the cell *4/5* can be easily retrieved.

Figure 4 also shows that UNIFEE introduces less irrelevant evidence, especially when it comes to the unwanted evidence format. The cells *Christgauś Record Guide* and *B-* can support one aspect of the claim, so our model only introduces two more sentences in need. By contrast, the baseline model retrieves two irrelevant cells, and 5 sentences from many different pages, which may confuse the verification model in the next step.

## 5 Related Works

In this work, we focus on the evidence extraction step of the fact verification based on table and text, i.e. the FEVEROUS dataset (Aly et al., 2021). Other fact verification datasets only concentrate on unstructured data (Thorne et al., 2018) or a single given table (Wang et al., 2021; Chen et al., 2020; Kwiatkowski et al., 2019). These datasets do not require fine-grained evidence extraction or only extract evidence of a single format.

For evidence sentence extraction, most previous works rank claim-sentence pairs with ESIM (Hanselowski et al., 2018; Zhou et al., 2019) or PLM (Liu et al., 2020; Soleimani et al., 2020). There are also works using the multi-turn retrieving method for multi-hop evidence (Stammbach and Ash, 2020) or training the evidence extraction and claim verification model jointly to relieve the error propagation (Yin and Roth, 2018). For cell evidence extraction, SEM-TAB-FACT (Wang et al., 2021) and FEVEROUS (Aly et al., 2021) are the only two verification datasets requiring fine-grained cell selection to our best known. Acharya (2021) propose to parse the claim into logical forms and identify cells with string matching and dependency parsing. Jindal et al. (2021) use an ensemble of BERT fine-tuned on a single-cell NLI task and a cell-wise similarity algorithm to capture the additional relationship. Many previous works on FEVEROUS flatten each table to a sequence and perform binary sequence labeling to select cells (Aly et al., 2021; Hu et al., 2022). However, these models extract evidence according to every single sentence/cell/table, they neglect the connections of evidence candidates. Meanwhile, they are not tailored for evidence extraction of two formats and cross-format evidence interactions, which are the main challenges of our task.

Many efforts aim at fusing evidence of two formats in FEVEROUS. Kotonya et al. (2021) and Gi et al. (2021) convert the extracted table cells to strings with human-designed templates to get a unified evidence set of sentences. Bouziane et al. (2021) convert each evidence sentence to a small table and verify the claim according to a set of evidence tables. Hu et al. (2022) perform bi-directional format conversion and apply dual-channel encoding to the evidence set. However, their format fusion is only performed in the verification step after evidence extraction. Without evidence fusion in the extraction step, much evidence has been left, especially cell evidence and evidence with fewer overlaps with the given claim.

For works fusing evidence in the extraction step, Kotonya et al. (2021) and Malon (2021) convert cells to strings and treat all evidence candidates as sentences for evidence selection. They neglect the

delicate structure of nested tables and lose much context information after the format conversion.

Our proposed UNIFEE introduces column nodes and layer-aware table representations, which can gather complex table information and facilitate cross-format interactions in a similar granularity in the evidence extraction step.

## 6 Conclusion

In this paper, we propose a Unified Evidence Extraction model (UNIFEE) for fact extraction over structured and unstructured data. UNIFEE adopts a mixed evidence graph to encourage evidence interactions among evidence candidates in both formats without manually designed conversion rules. Experiments on the FEVEROUS benchmark demonstrate that, with UNIFEE, a simple claim verification model outperforms previous SOTA results by a large margin. Further analysis shows that UNIFEE enhances the contextualized modeling of cells in complicated nested tables, thus largely improving the evidence extraction performance.

## 7 Acknowledgements

This work is supported in part by the NSFC Grants (No.62161160339, 62206070). For any correspondence, please contact Yansong Feng.

## 8 Limitations

Although we consider evidence interactions in the evidence extraction step and find out required evidence with less overlapping with the claim, it is hard for our method to recall multi-hop evidence in different pages since these pages are left out in the document retrieval step.

Apart from cell evidence and sentence evidence, there is a small proportion of evidence in the FEVEROUS dataset whose type is *table caption*, *list* or so. We simply ignore evidence of these types in the evidence retrieval step. To further improve the quality of evidence extraction step, we should also take these evidence types into consideration.

Another limitation is that the instances of the three veracity labels is unbalanced. From the details of each split shown in Appendix A, only 3% of the training split is labelled NEI, which makes it hard for models to learn predicting this class accurately. We have not tried solving this problem yet.

## References

Kaushik Acharya. 2021. KaushikAcharya at SemEval-2021 task 9: Candidate generation for fact verification over tables. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1271–1275, Online. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Mostafa Bouziane, Hugo Perrin, Amine Sadeq, Thanh Nguyen, Aurélien Cluzeau, and Julien Mardas. 2021. FaBULOUS: Fact-checking based on understanding of language over unstructured and structured information. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 31–39, Dominican Republic. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

In-Zu Gi, Ting-Yu Fang, and Richard Tzong-Han Tsai. 2021. Verdict inference with claim and retrieved elements using roberta. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65.

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. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Nan Hu, Zirui Wu, Yuxuan Lai, Xiao Liu, and Yansong Feng. 2022. Dual-channel evidence fusion for fact verification over texts and tables. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5232–5242, Seattle, United States. Association for Computational Linguistics.

Aditya Jindal, Ankur Gupta, Jaya Srivastava, Preeti Menghwani, Vijit Malik, Vishesh Kaushik, and Ashutosh Modi. 2021. BreakingBERT@IITK at SemEval-2021 task 9: Statement verification and evidence finding with tables. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 327–337, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Neema Kotonya, Thomas Spooner, Daniele Magazzeni, and Francesca Toni. 2021. Graph reasoning with context-aware linearization for interpretable fact extraction and verification. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 21–30, Dominican Republic. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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. *ArXiv preprint*, 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, Online. Association for Computational Linguistics.

Christopher Malon. 2021. Team papelo at FEVEROUS: Multi-hop evidence pursuit. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 40–49, Dominican Republic. Association for Computational Linguistics.

Mohammed Saeed, Giulio Alfarano, Khai Nguyen, Duc-Hong Pham, Raphael Troncy, and Paolo Papotti. 2021. Neural re-rankers for evidence retrieval in the feverous task. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. *Advances in Information Retrieval*, 12036:359.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, et al. 2020. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. 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 Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

## A Statistics of the FEVEROUS dataset

FEVEROUS[6] is an open-domain dataset based on English Wikipedia which contains 95.6 million sentences and 11.8 million tables. The dataset has 87,026 claims, with an average length of 25.3. An average of 1.4 sentences and 3.3 cells (0.8 tables) are required to verify each claim. Only text evidence is required in 34,963 instances, only tables in 28,760 instances, and both formats are required in 24,667 instances. 49,115 instances are marked as SUPPORTS, 33,669 as REFUTES, and the remaining 4,242 are marked as NEI. Table 4 displays specific label and evidence distributions in each split.

| | **Train** | **Dev** | **Test** |
|---|---|---|---|
| Supported | 41,835(59%) | 3,908(50%) | 3,372 (43%) |
| Refuted | 27,215(38%) | 3,481(44%) | 2,973 (38%) |
| NEI | 2,241 (3%) | 501 (6%) | 1,500 (19%) |
| Total | 71,291 | 7,890 | 7,845 |
| Sentences | 31,607(41%) | 3,745(43%) | 3,589 (42%) |
| Cells | 25,020 (32%) | 2,738(32%) | 2,816 (33%) |
| Sentence+Cells | 20,865 (27%) | 2,468 (25%) | 2,062 (24%) |

Table 4: Details of each split in FEVEROUS

[6] https://fever.ai/dataset/feverous.html