Dual-Channel Evidence Fusion for Fact Verification over Texts and Tables

Nan Hu[♠], **Zirui Wu**[♠], **Yuxuan Lai**[♡♠], **Xiao Liu**[♠], **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,lxlisa,fengyansong}@pku.edu.cn
laiyx@ouchn.edu.cn

Abstract

Different from previous fact extraction and verification tasks that only consider evidence of a single format, FEVEROUS brings further challenges by extending the evidence format to both plain text and tables. Existing works convert all candidate evidence into either sentences or tables, thus often failing to fully capture the rich context in their original format from the converted evidence, let alone the context information lost during conversion. In this paper, we propose a Dual Channel Unified Format fact verification model (DCUF), which unifies various evidence into parallel streams, i.e., natural language sentences and a global evidence table, simultaneously. With carefully-designed evidence conversion and organization methods, DCUF makes the most of pre-trained table/language models to encourage each evidence piece to perform early and thorough interactions with other pieces in its original format. Experiments show that our model can make better use of existing pre-trained models to absorb evidence of two formats, thus outperforming previous works by a large margin. Our code and models are publicly available[1].

1 Introduction

The task of fact extraction and verification aims to extract evidence and verify a given claim. Previous efforts focus on dealing with text format evidence from unstructured documents (Nie et al., 2019; Zhong et al., 2020; Kruengkrai et al., 2021) or evidence from a single given table (Chen et al., 2020; Yang et al., 2020; Eisenschlos et al., 2020). Recently, Aly et al. (2021) propose a new realistic setting, FEVEROUS, i.e., fact extraction and verification over unstructured and structured information. In FEVEROUS, models should not only extract evidence sentences/table cells from millions of passages, but also combine the evidence in different formats to verify a given claim.

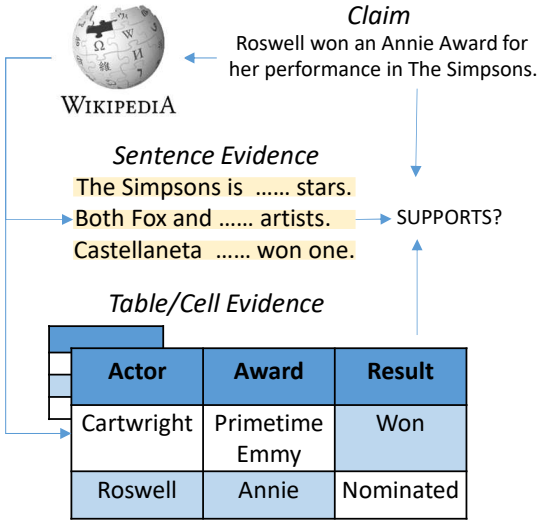

Figure 1: An excerpt example from FEVEROUS.

Previous works on FEVEROUS generally convert all evidence pieces into either plain text (Aly et al., 2021; Saeed et al., 2021; Malon, 2021) or several tables (Bouziane et al., 2021). However, format conversions inevitably lose rich context information for the converted evidence, thus may mislead the subsequent encoding and interaction steps. For example, in Figure 1, the entire top two rows are indispensable to understand the table cell *Won*. It is difficult to identify all related context cells and design a general conversion method to render them into sentences, but these connections can be easily caught by pre-trained table models (Herzig et al., 2020; Yin et al., 2020). On the other side, identifying/re-organizing crucial elements in a sentence to construct a table is also challenging. Simply inserting a whole sentence in a table cell (Bouziane et al., 2021) will make the new cells much larger (and unique) than normal ones, thus can not make the most of general pre-trained table models (Herzig et al., 2020; Yin et al., 2020) as we

*Corresponding Author.
[1]https://github.com/lanlanabcd/dual_channel_feverous

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5232 - 5242

expect.

Considering the inevitable expense in format conversion, we believe that each evidence in its original format can contribute necessary information to final verification, thus should be better encoded in its original format. This further indicates that we should design both sentence-to-table and table(cell)-to-sentence conversion methods to obtain all evidence in both formats, and maintain two parallel encoders to absorb the two formats, respectively. An advantage of doing so is to maximally encourage early interaction, which proves more effective than pair-wise encoding (Tymoshenko and Moschitti, 2021; Jiang et al., 2021)

When converting table evidence into sentences, previous works either convert table cells to a concatenation of key-value pairs (Aly et al., 2021; Malon, 2021), or construct sentences in a coordinate-description style (Kotonya et al., 2021a). They pay less attention to the conventional organization of tables structures. We observe that, in a table, the column headers usually represent the types/properties and the row headers often denote entities or scopes. We argue that one should consider these conventions to convert a table cell evidence into more natural sentences, and later pre-trained language models will be able to better capture the contextualized semantics of the table cells from generated sentences. On the other hand, existing pre-trained table models are trained to analyze one table at one time, while previous evidence conversion methods produce several small tables for one instance. It would be necessary to properly organize all evidence in one table so that pre-trained table models can allow the most interactions among all evidence pieces.

In this paper, we propose a dual channel unified format verification model (DCUF) to allow each evidence piece encoded in its original format and maximally maintain its original rich context, while encouraging further interactions with other evidence. DCUF converts each evidence into both textual and tabular formats in respective channels, and apply corresponding pre-trained models to learn the representations for the final verification. With the dual channel setting and carefully designed evidence conversion methods, DCUF makes better use of pre-trained language/table models to perform early and thorough interactions among all evidence and also between the claim and evidence.

In summary, we make the following contributions in this paper: (1) we propose DCUF, a novel model to maintain various evidence in two unified formats and allow each evidence piece to interact with other evidence in its best form. Experiments show that DCUF outperforms all previous works in literature. (2) we propose a context-aware evidence conversion method that can properly organize evidence of different formats, which fits current pre-trained language/table models hence take their most advantage to obtain accurate and focused representations.

2 Our Model

The FEVEROUS task can be formalized as, given a claim q and Wikipedia dump, a model is asked to find the evidence set consisting of sentences S and table cells C, and predict the veracity label of the claim accordingly. The veracity label set includes "SUPPORTS", "REFUTES" and "NOT ENOUGH INFORMATION".

2.1 Model Overview

Following the widely adopted fact verification pipeline (Thorne et al., 2018; Aly et al., 2021), we take three steps to solve the FEVEROUS task (i) retrieving pages from the Wikipedia dump; (ii) extracting evidence from the retrieved pages, and (iii) verifying the claim according to extracted evidence.

Specifically, for the document retrieval step, we narrow the search space with an information retrieval model DRQA (Chen et al., 2017) and then rerank the retrieved pages. For the evidence retrieval step, we design a multi-turn cell selector to extract sentence evidence and table evidence respectively, and select evidence cells from tables. Finally, we propose a Dual Channel Unified Format verification model (DCUF, shown in Figure 2) for the verification step. DCUF converts evidence to a unified table/sentence format with carefully-designed evidence conversion and re-organization methods in each channel, and combine dual-channel encodings to make the final prediction.

2.2 Document Retrieval

An efficient and effective document retriever is required since the Wikipedia dump containing millions of pages. We first narrow the search space to several hundred pages (m_0) with an efficient information retrieval method based on TF-IDF, namely, DRQA (Chen et al., 2017). A RoBERTa-based

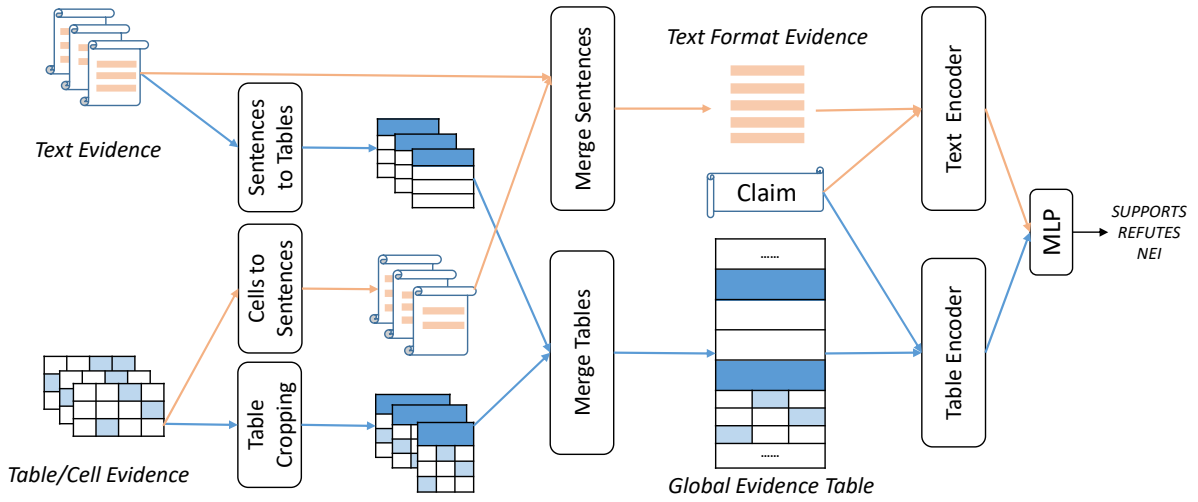

Figure 2: The architecture of our verification model. Orange lines show how we arrange the evidence into a text format, while blue ones show how to arrange into a table format.

re-ranker (Saeed et al., 2021) and a BM25-based re-ranker are then applied in parallel to re-rank the m_0 document candidates. We combine the results of two re-rankers and keep top m documents since BM25 focuses more on entity matching and RoBERTa-based re-ranker pays more attention to the overall sentence structure. The document scores are calculated as the sum of their rankings in the two re-rankers. Documents with lower scores have higher priority.

We further notice that the first several words of a claim always contain the page titles needed. We therefore derive a position-aware sub-sequence matching to strengthen the page retriever. We also remove pages with a long Wikipedia title starting with a specified year that is not contained in the claim.

2.3 Evidence Retrieval

We use DRQA (Chen et al., 2017) to extract k sentences $S = \{s_i\}_{i=1}^{k}$ and n tables $T = \{t_i\}_{i=1}^{n}$ from the retrieved pages, respectively. Then we select cells from the extracted tables. Many instances in the FEVEROUS dataset require evidence cells from more than one table, and each retrieved table has different relevance score to the claim. However, the widely-used cell extractor (Aly et al., 2021) reserves cells from only one table in their implementation.

We thus propose a Multi-turn Cell Selector (MCS), which retrieves cells from all evidence tables and consider the importance of the retrieved tables. A basic cell selector concatenates a given claim q and a flattened candidate evidence table t_i and feeds it into a sequence tagger to decide whether each cell in the table should be selected. MCS implements this procedure in a multi-turn manner, since each table has a different relevance score to the given claim. In the first turn, MCS selects the table most related to the given claim and feeds it to the basic cell selector. All cells with a selection score larger than the threshold g will be added to the cell evidence set C. In the second turn, all tables ranked second in T are the input to the basic selector. If the number of cells in C has not reached the upper limit, MCS adds the newly selected cells in the second turn to C. MCS repeats this procedure for n loops and we will get the cell evidence set $C = \{c_i\}_{i=1}^{o_j}$. o_j is the number of cells selected as evidence for the j^{th} instance.

2.4 Unified Format Encodings

Since the evidence can be of two formats, textual format and tabular (or cell) format, we convert each evidence of one format to another, so that we will get a unified representation of all evidence and could easily assemble them. We propose two conversion methods, i.e., cells-to-sentences and sentences-to-tables for the original tabular evidence and textual evidence, respectively. We carefully design the evidence conversion and re-organization methods to make converted evidence natural, thus take better advantage of the pre-trained language/table models.

2.4.1 Text Format Encoding

We consider the table conventions, i.e., row headers in general tables usually represent attribute types,

and convert table cells into natural sentences, thus make better use of pre-trained language models to perform early interaction over the claim and all evidence.

There are two types of tables on the Wikipedia pages, (i) general tables and (ii) Infoboxes. For general tables, we ignore header cells selected by the evidence retriever and convert content cells selected into text format. For each cell, we identify its row header cell and column header cell. We find that most general-typed tables are column tables, which means they only contain column headers. However, the first column of a table, in many cases, indicates the object name and others are attributes. Therefore, if a cell does not have an explicit header cell, we choose the first cell of the same row to be its row header cell. We observe that the row header cell for a general table always indicates the object name, the column header cell indicates the attribute type, and the cell itself is the attribute value for the object. We thus form the corresponding text for a cell in a general table as "<column header> for <row header> is <cell value>." For Infoboxes, it is a different story. The row header is always the attribute type, the column header is the field that the attribute belongs to, and the Wikipedia title is the object name. Therefore, the text representation of each cell in an Infobox is formulated as "<column header> : <row header> of <Wikipedia title> is <cell value>". As shown in Figure 3, "21-24 minutes" is a selected cell, its row header "Running Time" is the attribute type, and the header cell "Production" is the field which the attribute "Runing time" belongs to. Thus, the cell will be converted to "Production: Running time of The Simpsons is 21-24 minutes."

Claim verification on a set of evidence containing many cells often requires operations on cells in the same column, such as maximum and summary (Aly et al., 2021). Therefore, we pack the texts from cells in the same column together. These texts are joined by semicolons and form a piece of column text. Each column is converted into a sentence. As shown in Figure 3, the cells "December 17, 1989" and "October 11, 1990" are jointly converted to one sentence, "Season premiere for 1 is December 17, 1989; Season premiere for 2 is October 11, 1990." The given claim, the extracted text evidence, and the column texts are concatenated together to form the input to the text format evidence encoder, separated by the separator "</s>" in the RoBERTa model. Column texts from the same table are adjacent, and sentences from the same Wikipedia page are arranged together.

2.4.2 Table Format Encoding

As shown in the right part of Figure 3, we propose to construct a single evidence table containing all evidence candidates, both textual and tabular, since most existing table pre-training models are designed to and trained to analyze one single table at one time. Making better use of the pre-training models helps to perform early and thorough evidence interaction.

The method of converting text format evidence to tabular format is straightforward. We group the evidence sentences from the same Wikipedia page. If there are n evidence sentences from the same Wikipedia page, we construct a table of $n+1$ rows and 1 column. The only cell in the first row is a header cell containing the Wikipedia title. And for other rows, each row has one cell, and in that cell is a piece of sentence evidence from that page. Assuming there are evidence sentences from m pages, we will get m table units after this step.

As the pre-training table model has a capacity limit, we crop irrelevant cells first to compress the extracted tables. To be precise, rows and columns containing no selected cells are removed. After that, we add one cell row to the top of each cropped table, this cell contains the title of the Wikipedia page from which the table is extracted. If there are n extracted tables, we will get n cropped table from this step.

We get a global evidence table by stacking the m tables from sentences and n tables from tabular evidence, as illustrated in Figure 3. Then, we feed the claim and the global evidence table to a pre-trained table model, TAPAS (Herzig et al., 2020), and get the tabular format evidence representation.

2.4.3 Dual-Channel Verdict Prediction

The final verdict prediction is based on the dual channel encoding. Therefore, each evidence can be encoded in its original format while interacting with all evidence pieces and the claim.

We concatenate the text format evidence encoding h_{text} and the tabular format evidence encoding h_{tab} to obtain a joint format encoding for prediction. With a feed-forward network and a softmax layer, we obtain the veracity probability distribution of the claim and the predicted label is the one

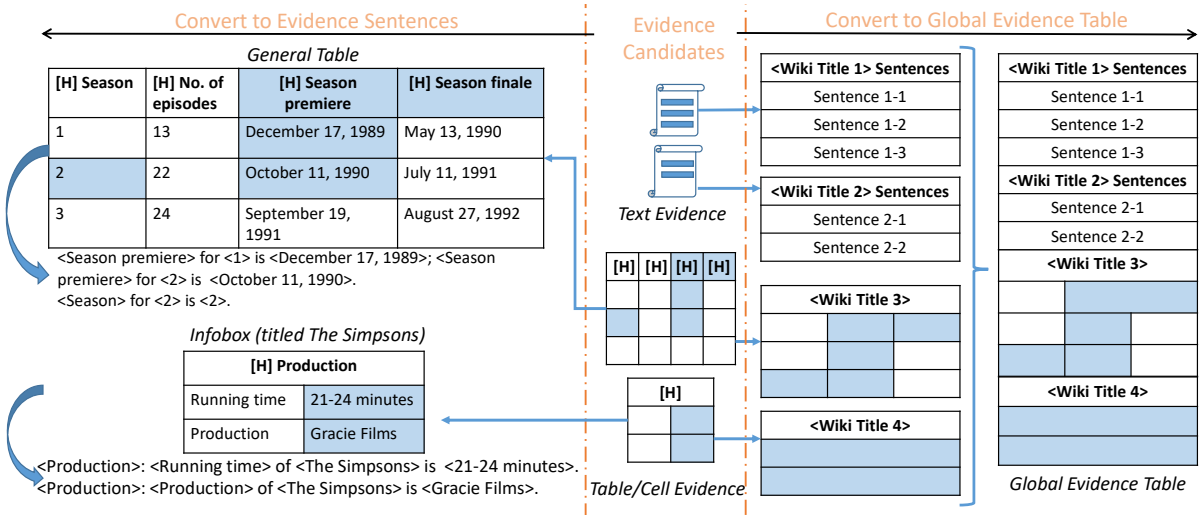

Figure 3: An example of our evidence conversion methods. Cells with [H] are explicitly specified header cells. Light blue cells are the selected cells. The left part shows how we convert evidence to a unified text format, while the right shows how to convert all evidence candidates into a global evidence table.

with the largest probability:

$$h = [h_{\text{text}}; h_{\text{tab}}] \quad (1)$$

$$\mathrm{p}\left(y|q, S, T, C\right) = \text{Softmax}\left(\text{FNN}\left(h\right)\right) \quad (2)$$

$$\hat{y} = argmax_y \mathrm{p}\left(y|q, S, T, C\right) \quad (3)$$

where $\mathrm{p}\left(y|S, T, C\right)$ represents the probability of each alternative label y given the claim q, evidence sentences S and evidence tables T.

To strengthen the model's ability to predict the veracity label with the evidence set containing irrelevant pieces, we construct two instances for each claim in the training set. One is the claim with the gold evidence provided by the FEVEROUS dataset, and the other is the claim with extracted evidence pieces from previous evidence extraction steps. We use the cross-entropy loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\mathrm{p}\left(\hat{y} = y_i|q, S, T, C\right)\right) \quad (4)$$

where y_i is the true veracity label of the i^{th} instance. N is the size of the training set.

3 Experiments

We evaluate our models on the FEVEROUS dataset, where each claim is annotated with a gold veracity label and several gold evidence sets. Any one of the evidence sets is sufficient to verify the claim. More details about the FEVEROUS dataset are in Appendix.

The FEVEROUS dataset provides two official metrics, namely label accuracy (Acc.) and FEVEROUS score. Label accuracy calculates the ratio of the instances whose veracity label is correctly predicted. FEVEROUS score is the ratio of the instances whose veracity label is correct and the extracted evidence set is sufficient. Here, sufficient evidence sets are defined as the evidence sets covering one of the gold evidence sets provided in the FEVEROUS dataset. Note that there are at most 25 table cells and 5 sentences to calculate the scores.

3.1 Implementation Details

In the document retrieval step, the number of pages retrieved by the BM25-based retriever m_0 is 150. We keep the top 5 pages for evidence extraction after page re-ranking. For evidence retriever, the top k =5 sentences and top n =2 tables are extracted from the retrieved pages. The gate of cell selection g is 0.25 and a maximum of 25 cells are selected in total for each claim.

We use an Adam optimizer (Kingma and Ba, 2015) with a linear learning rate scheduler. The rate of warm-up steps is 20%. The peaking learning rate for parameters in pre-trained language models is 10^{-5} and 10^{-3} for other parameters. The batch size is 24, implemented with gradient accumulation techniques. DCUF takes about 5 hours to run on a single NVIDIA A100 Tensor Core GPU (40GB).

The RoBERTa-based re-ranker is initialized from the hugging face checkpoint[2] without further fine-tuning. The TAPAS checkpoint is fine-tuned with a table fact verification task, Tabfact (Chen

[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2

Models	Development					Test				
	Feverous	Acc.	E-P	E-R	E-F1	Feverous	Acc.	E-P	E-R	E-F1
Official Baseline	19	53	12	30	17	17.73	48.48	10.17	28.78	15.03
EURECOM	19	53	12	29	17	20.01	47.79	13.73	33.73	19.52
Z team	–	–	–	–	–	22.51	49.01	7.76	42.64	13.12
CARE	26	63	7	37	12	23	53	7	37	11
NCU	29	60	10	42	17	25.14	52.29	9.91	39.07	15.81
Papelo	28	66	–	–	–	25.92	57.57	7.16	34.60	11.87
FaBULOUS	30	65	8	43	14	27.01	56.07	7.73	42.58	13.08
DCUF	**35.77**	**72.91**	**15.06**	**43.22**	**22.34**	**33.97**	**63.21**	**14.79**	**44.10**	**22.15**

Table 1: Model performance on the development set and test set. Acc. is Accuracy. E-P, E-R and E-F1 is Evidence Precision, Evidence Recall and Evidence F1, respectively.

et al., 2020) [3]. Same as the baselines, the sentence evidence encoder is RoBERTa-large tuned with several NLI and verification tasks[4].

Document Retriever Details For BM25 re-reranker, all page candidates of every 200 adjacent instances are merged to build the BM25 index for these instances. Each document in the Wikipedia dump is represented by the concatenation of its title and the first 64 words of its content for the BM25 re-ranker. We use NLTK[5] to remove stop words and lemmatize the remains. For position-aware entity matching, if a sub-sequence, with more than two words, in the first ten words of a given claim is a page title in the Wikipedia dump and it is not in the m documents we replace the page of the lowest priority with it.

3.2 Main Results

The overall performance of our model on the development set and the test set are shown in Table 1. We get an increase of 5.77% on the FEVEROUS score and 7.91% on the accuracy over the previous best model FaBULOUS (Bouziane et al., 2021) on the development set. For the test set, the increase is 6.96% and 7.14% in Feverous score and label accuracy, respectively. These results suggest the effectiveness of our proposed DCUF model. The evidence format of a global evidence table is consistent to the input of pre-trained table models. Thus, DCUF can make better use of the internal ability of pre-trained models than previous works which concatenate linearized tables or max-pool lots of claim-table pair encoding (Bouziane et al., 2021). Moreover, DCUF also performs better than another well-performing model, CARE (Kotonya et al., 2021b). DCUF converts cells to meaningful sentences that are similar to the inputs of PLMs pre-training stage, which makes better use of the PLMs ability.

We also conduct experiments with the gold evidence to investigate the effectiveness of our verification model. The results are shown in Table 2. DCUF obtains an increase of 2.88% on accuracy over the previous best result and 3.68% over RoBERTa-based models. With all evidence candidates in the same format and preserving context information, our system can make better use of the pre-trained language/table models and perform early and thorough interactions among evidence and between the claim and evidence.

Models	Accuracy
Fabulous-Max	80
Papelo-RoBERTa	82.9
NCU	84
Fabulous-Joint	84
Papelo-T5	84.8
DCUF	**87.68**

Table 2: Model accuracy on the development set with gold evidence as model input.

Evidence Extraction Results The document retrieval results are shown in Table 3. In Table 3, experiments show that both BM25 re-ranker and RoBERTa re-ranker can improve the document retrieval quality to a great extent compared to vanilla DrQA page retriever, with whole evidence set recall improvement of 8.63% and 12.56% respectively. The combination of them can further enlarge this gap to 16.22%. We find that the average number of pages in the merged set of top-5 BM25 re-ranked

[3] https://huggingface.co/google/tapas-large-finetuned-tabfact

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

[5] https://www.nltk.org

Models	MAP	Rec-single	Rec-set
DrQA@150	91.13	90.00	89.29
DrQA	69.32	66.91	65.50
BM25 RR	77.98	75.32	74.13
RoBERTa RR	82.56	78.65	78.06
Combined RR	85.13	82.29	81.72
+ Rule Enhancement	88.11	85.20	84.82

Table 3: Document retrieval results on the development set. RR means re-ranker. Rec-single is the recall of single evidence. Rec-set considers all evidence for a instance as a whole. It is the recall with top-5 pages retrieved without special statement.

Models	Table	Sent	Cell	All
Official Baseline	56	53	28.70	30
FaBULOUS	-	56.6	34.2	40.4
MCS	75.59	62.54	58.41	43.22

Table 4: Categoried evidence recall on the development set. The recall is calculated with at most 3 tables, 5 sentences and 25 table cells.

results and top-5 RoBERTa re-ranked results is 7.75 in the development set. It proves that these two re-rankers tend to focus on different aspects when evaluating the correlation of the given claim and a page candidate. The rule-based enhancement methods, namely matching position-aware entities and removing pages with unmatched year, bring a a further improvement of 3.10%. It indicates that document retrievers should take the word positions in the given claim into consideration. Without any training procedure or Dense Retriever (which is time-consuming), we get a whole set recall of 84.82% when retrieving 5 pages for each claim.

Table 4 shows the evidence extraction results. With MCS, we achieve an increase of 29.71% on cell recall and 13.22% on the overall evidence recall over FaBULOUS. The result indicates that considering only one table is not enough and we should pay attention to the relevance scores of the input tables especially when the cell selector is somewhat weak.

3.3 Ablation Study

We evaluate the effect of each part of DCUF with a collection of ablation experiments. The Experiment settings are as follows. (1) **w/o Table Format Encoding** We only use the unified text format evidence encoding for verdict prediction. (2) **w/o Table Format Encoding** We only use the unified table format evidence encoding for verdict prediction. (3) **w/o Dual-Channel Predictor** We use the verdict predictor in the baseline to predict the veracity label.

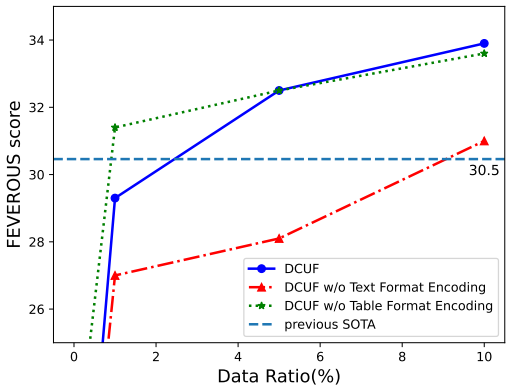

Figure 4: Few shot results on the development set.

Models	Feverous	Accuracy
DCUF	**35.77**	**72.91**
w/o Table Format Encoding	35.16	72.05
w/o Text Format Encoding	32.41	66.80
w/o Dual Channel Predictor	27.91	58.40

Table 5: Ablation results on the development set.

The results are shown in Table 5. The FEVEROUS score and accuracy drop consistently when we remove the unified table format encoding or the unified text format encoding. Especially, the FEVEROUS score drops by 3.34% when only using the unified tabular encoding for prediction. With only unified text format encoding, the FEVEROUS score drops by 0.61%, which may contribute to pre-trained table models, such as TAPAS, being still weaker compared to pre-trained language models. However, with our carefully-designed context-aware unified format conversion, verdict prediction upon one format encoding outperforms all previous results. To relieve the effect of different evidence extraction methods, we train the verdict prediction model in the baseline with our evidence extraction results. We see a great drop, with accuracy dropping of 14.51% and FEVEROUS score by 7.86%, which proves that our combined unified format verdict prediction model can keep the context needed when converting the evidence format and help the extracted evidence perform early interaction in a unified format.

4 Analysis

Few-shot Results Figure 4 shows the FEVEROUS score on the development set when we train unified DCUF, DCUF without text format encoding and DCUF without table format encoding verdict predictors on 1%, 5% and 10% instances of the training set respectively. We find that, training on only 10% instances, all of the three settings outperform previous SOTA results, with the dual channel predictor achieving the best FEVEROUS score of 33.9%. Meanwhile, with only 1% of the training set, namely, 713 instances, the unified text format predictor outperforms previous SOTA result by 0.9%. These improvements may contribute to our carefully-designed format conversion methods. The text format evidence converted from cells is similar to sentences in natural language when keeping much context information in the conversion procedure. And the concatenated claim-evidence string is the same as previous fact verification tasks on which the RoBERTa checkpoint is fine-tuned. Meanwhile, a given claim, and a single global evidence table are consistent to the input requirements in the pre-training step of TAPAS. With few new parameters introduced and an input form strictly complying with the requirements of pre-trained models and previous single format fact-checking tasks, DCUF could make the most of the pre-training stage and, thus, learn well in the few-shot setting.

We also find that when training with only 1% instances, the unified text format predictor outperform other two settings. As the number of training instances increase, dual-channel predictor learns how to combine information from the two channel, thus achieving better results.

Error Analysis We find that when converting cells to text with rule-based methods, there are inevitably noise or not fluent sentences introduced.

One problem is caused by the latent header cells. As shown in Table 6(a), the cell "Team" is selected but not explicitly marked as a header cell, so a meaningless sentence "Senior career*: Years for Aramais Yepiskoposyan is Team." is derived from this cell. Meanwhile, although the header cells indicating the attribute type are usually nouns, there are some exceptions. Infobox Table 6(b) is an example. The selected cell "Sir Roger Manwood" is converted to a fluent and meaningful evidence sentence "Information: Founder for Sir Roger Manwood's School is Sir Roger Manwood.". However, "Information: Established for Sir Roger Manwood's School is 1563; 457 years ago." is not a sentence

(a) Case A

Wikipedia page: Aramais Yepiskoposyan

Senior career*		
Years	Team	Apps
1986-1991	FC Ararat Yerevan	10
1997	FC Kuban Krasnodar	8

(b) Case B

Wikipedia page: Sir Roger Manwood's School

Information	
Founder	Sir Roger Manwood
Established	1563; 457 years ago

(c) Case C

Wikipedia page: List of The Simpsons cast members

Episodes	Actor	Character(s)
179	Marcia Wallace	Edna Krabappel
52	Phil Hartman	Troy McClure

Table 6: Example tables in the FEVEROUS dataset.

that conforms to English grammar. And sometimes, the latent row header which indicates the object name is confusing. For example, the cell "Marcia Wallace" in Table 6(c) is converted to a sentence "Actor for 179 is Marcia Wallace." Without the header "Episodes", what "179" refers to is confusing. Assessing and polishing the converted sentence may help to solve the problems presented above.

5 Related Works

Fact Verification over Unstructured Evidence Thorne et al. (2018) proposed FEVER, a large-scale dataset of claims based on Wikipedia articles. Language models have better performance compared to other methods e.g. ESIM-based models(Hanselowski et al., 2018; Nie et al., 2019). BERT-based models make the prediction based on collected evidence in a direct aggregating rule(Soleimani et al., 2020) or a graph-based approach(Zhou et al., 2019; Zhong et al., 2020).

Fact Verification over Structured Evidence Benchmarks for fact verification on structured evidence are built on tables collected from Wikipedia(Chen et al., 2020) or scientific articles(Wang et al., 2021). Many previous works search latent programs as an intermediary to reason over the given table. They directly encode programs (Chen et al., 2020)or construct heterogeneous graphs (Shi et al., 2020; Yang et al., 2020) with the claim, the table and the programs. Another

way is to linearize the input table and perform table pre-training (Chen et al., 2020) and add additional table-aware embeddings (Herzig et al., 2020; Eisenschlos et al., 2020) to enhance the table encoding. However, in these datasets, the evidence is only one given table, and models are not requested to find out the evidence cells explicitly.

Fact Verification over Structured and Unstructured Evidence FEVEROUS (Aly et al., 2021) is the first dataset of fact verification on structured and unstructured evidence. Many previous works follow the baseline settings and convert all evidence to text format to perform evidence interaction. They transform each cell to header-value pairs (Aly et al., 2021; Malon, 2021) or in a cell location indication type (Kotonya et al., 2021a). They pay less attention to making the converted text more consistent with natural language expressions or identifying what the context cells represent. Bouziane et al. (2021) propose to convert all evidence to tables. They simply convert each sentence to a 2-cell table with the Wikipedia title and itself instead of packing closely-tied evidence and building a global evidence table. There are also works focusing on the first two steps two improve the final results. Saeed et al. (2021) propose to add a document re-ranker to strengthen the document retrieval. Multi-hop Dense Retriever (Bouziane et al., 2021) and T5 generator (Malon, 2021) are introduced to better extract multi-hop evidence.

6 Conclusion

In this paper, we propose DCUF, a dual channel unified format model for fact verification over structured and unstructured data. With context-aware evidence format conversion, DCUF gets a unified text format representation of all evidence and a global evidence table of them at the same time. The dual channel design helps us make the most of existing pre-trained language/table models to encourage all evidence pieces to interact with each other in their best forms as early as possible. Experiments show that, with dual-channel unified format encoding, our proposed DCUF achieves state-of-the-art performance and also comparable results in few-shot settings.

Acknowledgements

This work is supported in part by NSFC (62161160339). We would like to thank the anonymous reviewers and action editors for their helpful comments and suggestions. For any correspondence, please contact Yansong Feng.

References

Rami Aly, Zhijiang Guo, M. Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *ArXiv*, abs/2106.05707.

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.

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.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

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.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with T5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021,*

(Volume 2: Short Papers), Virtual Event, August 1-6, 2021, pages 402–410. 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. 2021a. Graph reasoning with context-aware linearization for interpretable fact extraction and verification. *ArXiv*, abs/2109.12349.

Neema Kotonya, Thomas Spooner, Daniele Magazzeni, and Francesca Toni. 2021b. Graph reasoning with context-aware linearization for interpretable fact extraction and verification. *arXiv preprint arXiv:2109.12349*.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2447–2460. Association for Computational Linguistics.

Christopher Malon. 2021. Team papelo at feverous: Multi-hop evidence pursuit. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

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)*.

Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. 2020. Learn to combine linguistic and symbolic information for table-based fact verification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5335–5346, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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.

Kateryna Tymoshenko and Alessandro Moschitti. 2021. Strong and light baseline models for fact-checking joint inference. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4824–4830. Association for Computational Linguistics.

Nancy XR Wang, Diwakar Mahajan, Marina Danilevsk Rosenthal, et al. 2021. Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts). *arXiv preprint arXiv:2105.13995*.

Xiaoyu Yang, Feng Nie, Yufei Feng, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020. Program enhanced fact verification with verbalization and graph attention network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7810–7825, Online. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

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, Online. 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

The FEVEROUS is an open-domain English dataset. It contains 87,026 claims, and the claim length is 225.3 on average. Each claim averagely needs 1.4 sentences and 3.3 cells (0.8 tables) to be verified. 34,963 instances need only text format evidence, 28,760 only table format and 24,667 need a combination of the two formats. There are 49,115 instances labeled SUPPORTS, 33,669 labeled Refutes and the rest 4,242 instances are labeled NEI. Detailed label and evidence distributions are shown in Table 7.

	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%)	3589 (42%)
Cells	25,020 (32%)	2,738(32%)	2816 (33%)
Sentence+Cells	20,865 (27%)	2,468 (25%)	2062 (24%)

Table 7: FEVEROUS Distribution.