Table-based Fact Verification with Salience-aware Learning

Fei Wang, Kexuan Sun, Jay Pujara, Pedro Szekely and **Muhao Chen**
Department of Computer Science & Information Sciences Institute
University of Southern California
{fwang598,kexuansu,jpujara,szekely,muhaoche}@usc.edu

Abstract

Tables provide valuable knowledge that can be used to verify textual statements. While a number of works have considered table-based fact verification, direct alignments of tabular data with tokens in textual statements are rarely available. Moreover, training a generalized fact verification model requires abundant labeled training data. In this paper, we propose a novel system to address these problems. Inspired by counterfactual causality, our system identifies token-level salience in the statement with *probing-based salience estimation*. Salience estimation allows enhanced learning of fact verification from two perspectives. From one perspective, our system conducts *masked salient token prediction* to enhance the model for alignment and reasoning between the table and the statement. From the other perspective, our system applies *salience-aware data augmentation* to generate a more diverse set of training instances by replacing non-salient terms. Experimental results on TabFact show the effective improvement by the proposed salience-aware learning techniques, leading to the new SOTA performance on the benchmark. [1]

1 Introduction

Fact verification, the problem of determining whether a statement is entailed or refuted by evidence, has quickly become a critical problem in NLP to combat information pollution (Rashkin et al., 2017; Thorne et al., 2018; Zhang et al., 2019; Zellers et al., 2019; Wadden et al., 2020). Successful fact verification enables downstream tasks such as misinformation detection, fake news identification, factual error correction, and deceptive opinion detection (Ott et al., 2011; Shu et al., 2017; Yoon et al., 2019; Cao et al., 2020).

Recently, table-based fact verification (Chen et al., 2020a; Zhong et al., 2020; Yang et al., 2020) has garnered attention. As a ubiquitous and clean format of semi-structured knowledge, tables are regarded as reliable sources of evidence to verify the textual statements (Chen et al., 2020a). Leveraging tabular data for fact verification requires identifying relevant evidence in tables, and conducting logical reasoning according to the selected evidence. Prior studies have attempted to generate logical programs to capture logical operations and relations between the statement and the table (Zhong et al., 2020; Yang et al., 2020; Shi et al., 2020). More recent work shows that Transformer-based language models with general and task-specific pre-training over textual and tabular data can achieve SOTA performance without counting on explicit logical programs (Eisenschlos et al., 2020; Dong and Smith, 2021).

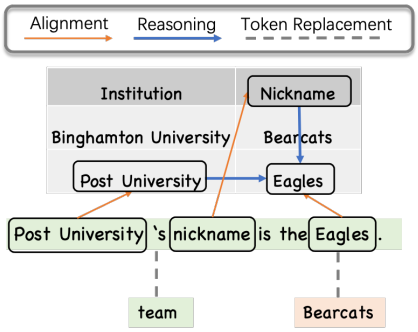

Figure 1: An example of table-based fact verification, with green for entailed statements and red for refuted statements. Alignment and reasoning are essential for both table-based fact verification and masked salient token prediction (e.g. "Eagles"). Token replacement may lead to similar (e.g. "'s" to "team") or different (e.g. "Eagles" to "Bearcats") statements.

However, to provide a reliable solution to the table-based fact verification task, several critical challenges are still overlooked by prior studies. One challenge is to effectively provide connections among components of the statement and substructures of the table, and accordingly conduct the in-

[1] Our code is publicly available at https://github.com/luka-group/Salience-aware-Learning

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4025–4036
November 7–11, 2021.

ference. Being unaware of such fine-grained connections and logical relations could raise the risk of misalignment, incorrect reasoning and ignoring salient components of a statement, and therefore leads to incorrect verification results. For example, to verify the statement in Fig. 1, the model should implicitly or explicitly infer all the five arrows accurately. Although some works have tried to perform token-level interactions and generate logical programs to connect statements and tables and conduct logical reasoning (Zhong et al., 2020; Yang et al., 2020), the supervision signals to guide the learning process are typically sparse. Another challenge is that training a well-generalized fact verification model non-trivially requires abundant labeled training data. Limited training data can only cover limited statement patterns and hinder robustness and generalizability of model inference. Previous works either trained on limited data (Zhong et al., 2020; Yang et al., 2020) or augment training data with specific statement generation templates (Eisenschlos et al., 2020). Yet, in real-world scenarios, statements and evidences can be presented in very diverse ways, and such diversity is difficult to be comprehensively captured by specific templates.

To this end, we propose a novel salience-aware learning system for table-based fact verification. Starting from a TAPAS (Herzig et al., 2020) language model fine-tuned on the TabFact dataset, our system identifies salient and non-salient tokens in statements with a *probing-based salience estimation* method inspired by counterfactual causality (Pearl, 2009) (§3.2). Then, the system leverages the estimated salience information from two perspectives. From one perspective, to enhance the model for capturing fine-grained connections and supporting the reasoning between statements and tables, the system conducts *masked salient token prediction* as an auxiliary task (§3.3). More specifically, this task is to predict the masked salient token in an entailed statement given the corresponding table by reusing the embedding layer of TAPAS as a language model head. The fact verification task can receive indirect supervision from the auxiliary task, as both of them requires table-text alignment and logical reasoning. From the other perspective, to improve the model robustness, instead of using templates for statement augmentation like prior work (Eisenschlos et al., 2020), we develop a *salience-aware data augmentation* technique (§3.4). Intuitively, replacing non-salient tokens provides unseen statements while preserving the meaning and correctness of the original statement. This strategy enhances the size and comprehensiveness of the training data and further complements training with more supervision signals.

The main contributions of this paper are threefold. First, we propose a probing-based salience estimation method to evaluate the importance of each token in a statement according to the counterfactual causality theory. Second, we propose a novel salience-aware learning system that helps the fact verification model to find the connections between the table and the statement, and enhance the inference ability of the model with the auxiliary task of masked salient token prediction. Third, to complement with insufficient training signals and improve the model robustness on heterogeneous statements, we incorporate a probabilistic data augmentation method driven by non-salient tokens. We evaluate our system based on the TabFact benchmark, which shows promising performance on this task and drastically outperforms prior methods. Detailed analysis demonstrates the effectiveness and essentially of both masked salient token prediction and salience-aware data augmentation techniques for the improved performance.

2 Related Work

In this section, we provide a selected summary for two related research topics.

2.1 Fact Verification

Fact verification have become an essential research topic in recent years with the rising concerns of misinformation (Vlachos and Riedel, 2014; Wang, 2017; Thorne et al., 2018; Khattar et al., 2019; Zellers et al., 2019; Chen et al., 2020a). Early works on fact verification are mainly based on unstructured textual evidence (Yin and Roth, 2018; Nie et al., 2019; Zhou et al., 2019).

Recently, much attention has been paid to table-based fact verification (Chen et al., 2020a; Zhong et al., 2020; Yang et al., 2020; Eisenschlos et al., 2020; Shi et al., 2020; Dong and Smith, 2021). Chen et al. (2020a) released the TabFact benchmark, and motivated two lines of research. Considering the importance of logical operations in this task, some works introduce such inductive bias by explicitly generating and capturing logical programs. Latent Program Algorithm (LPA) (Chen et al., 2020a) collected potential program candi-

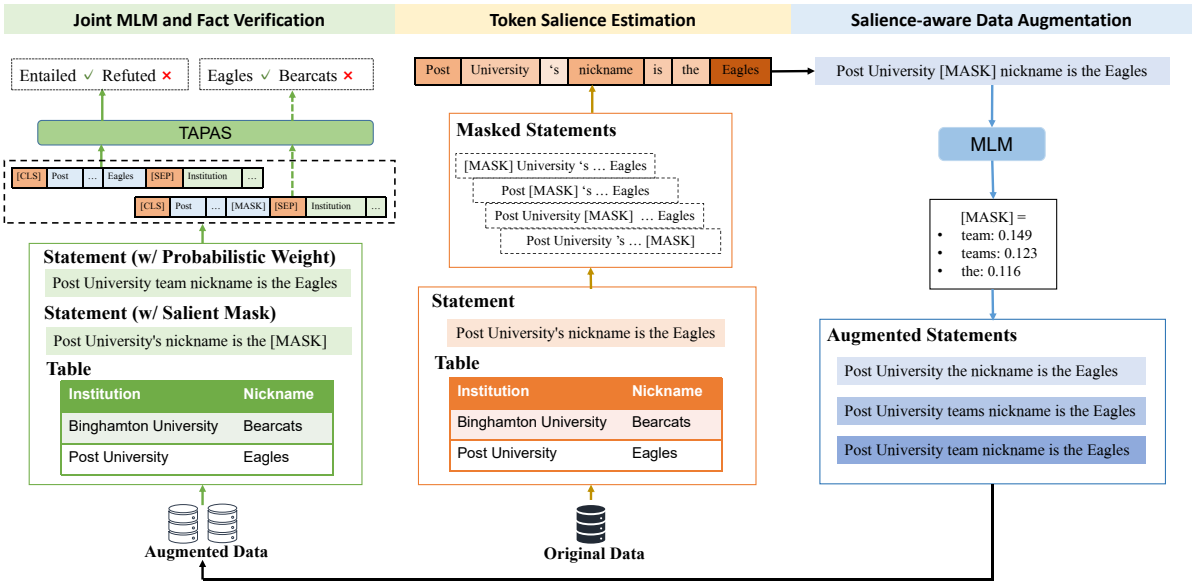

Figure 2: Workflow of the proposed system. The system is composed of three parts. The arrows illustrate how information is transferred. For tokens, a lighter background color indicates a lower salience score. For augmented statements, a lighter background color indicates a smaller probability.

dates and execution results according to a search algorithm, and then trained a Transformer-based (Vaswani et al., 2017) model to assign a confidence score to each program based on matching to the statement. Through this line, later works have explored improved ways to generate and capture logical programs (Zhong et al., 2020; Yang et al., 2020). LogicalFactChecker (Zhong et al., 2020) generated logical programs using a sequence-to-action generation approach, where it applied neural module networks (Andreas et al., 2016) to capture the logical structure of programs. HeterTFV (Shi et al., 2020) learned to combine linguistic information and symbolic information with a heterogeneous graph attention network. ProgVGAT (Yang et al., 2020) verbalized the execution processes of the generated programs, and applied graph attention networks (Veličković et al., 2017) to capture each execution tree. Beside logical programs, other studies applied pre-trained language models to linearized tables and perform fact verification as natural language inference (NLI) (Chen et al., 2020a; Eisenschlos et al., 2020; Dong and Smith, 2021). Table-BERT (Chen et al., 2020a) applied BERT (Devlin et al., 2019) as an NLI model. Eisenschlos et al. (2020) and Dong and Smith (2021) improve this strategy by conducting task-specific pre-training to TAPAS (Herzig et al., 2020), a Transformer-based language model pre-trained on both textual and tabular data.

Our work takes advantages of both lines of research on table-based fact verification, introducing cross-structural alignment bias and logical reasoning bias to pre-trained language models. Besides, previous works focus on significant words in statements, while we apply data augmentation to improve model robustness to insignificant words.

2.2 Counterfactual Causality in NLP

Counterfactual thinking and causal inference have inspired several studies in natural language processing, including counterfactual story rewriting (Qin et al., 2019), paraphrasing diversification (Park et al., 2019), measuring fairness in text classification (Garg et al., 2019), debiasing in machine translation (Saunders and Byrne, 2020) and visual question answering (Niu et al., 2021). This direction has also developed data augmentation strategies in various NLP tasks (Zmigrod et al., 2019; Kaushik et al., 2019; Fu et al., 2020; Zeng et al., 2020). Especially, counterfactual causality has been used to measure the causal effects of specific inputs in visual question answering (Niu et al., 2021).

Inspired by these applications, we apply the thought of counterfactual causality on table-based fact verification, and detect token-level salience in statements in a probing manner.

3 Method

In this section, we describe the technical details of the proposed system. Our system extends the NLI

formulation of table-based fact verification (Eisenschlos et al., 2020) with the pretrained language model TAPAS as the backbone (§3.1). As a preliminary step, our system estimates token-level salience in a probing manner for each statement (§3.2). The proposed salience-aware learning leverages the estimated salience information from two perspectives. From one perspective, it enhances the main task learning with an auxiliary task of masked salient token prediction (§3.3). In this auxiliary task, our system masks salient tokens in entailed statements and requires the model to jointly solve the cloze task along with the main task of fact verification. From the other perspective, our system incorporates a probablistic data augmentation technique (§3.4) by replacing non-salient tokens in statements according to a pretrained masked language model (MLM). This is followed by the technical details of training and inference processes (§3.5). The overall architecture of our system is shown in Fig. 2.

3.1 Base Model for Fact Verification

Our system adopts the TAPAS (Herzig et al., 2020) model from the previous SOTA method as the base model. In this way, we also formulate the main task of table-based fact verification as an NLI task following Eisenschlos et al. (2020).

For a brief description of TAPAS, it extends BERT's architecture (Devlin et al., 2019) with additional positional embeddings to represent tabular structure. Specifically, in addition to the embeddings used by BERT, the model applies column and row embeddings to represent the column index and row index of the cell enclosing the token, and rank embeddings to represent the numeric rank of the cell referring to the token if the column is sortable. It flattens the table into a sequence of words and concatenates them with textual sequence if any as input. The model is pre-trained using an MLM objective. Eisenschlos et al. (2020) designed task-specific intermediate pretraining tasks to improve the model performance on table-based fact verification. We use the model released by them as our basic model. Following their setting, we add a `[CLS]` token at the beginning of the input sequence, and separate the statement and the linearized table with a `[SEP]` token. Then, our system adopts the TAPAS model to encode the input sequence and model the probability of entailment with a task-specific prediction head taking the final representation of the `[CLS]` token as input. Specifically, the task-specific prediction head is implemented as an MLP with the sigmoid activation fuction for binary classification, which is consistent with Eisenschlos et al. (2020).

3.2 Probing-based Salience Estimation

Lexical tokens usually have different levels of importance with regard to the overall content or purpose of a description (Chiarcos et al., 2011; Liu et al., 2018; Xiong et al., 2018). For example, in the sentence "`Post University has used the Eagles as its nickname`", the tokens like "`Eagles`" and "`nickname`" are more important than others such as "`has used`" and "`as`" for determining if the sentence is refuted or entailed. We refer to such highly important tokens as *salient* tokens, and less important ones as *non-salient* tokens. To make use of token-level salience in the table-based fact verification task, the immediate challenge is to *estimate the salience of each token in a statement*.

Inspired by the counterfactual theories of causation (Pearl, 2009; Lewis, 2013), we address the challenge with a probing-based salience estimation method. Counterfactual causality has been widely used in social science for measuring the causal effects of specific factors (Tetlock and Belkin, 1997; Brady, 2008; Morgan and Winship, 2015), and has also been introduced to deep learning (Tang et al., 2020; Niu et al., 2021). In our context of fact verification, the intuition of counterfactual causation is to testify that: *If the model has not seen the token, will it still make the same prediction?* The counterfactual lies between the fact that the token is seen and the imagination that the token is masked. The comparison between them naturally reflects the effect of the token, because the token is the only thing changed between the two situations.

Technically, to estimate the salience of a token in a statement, we compare the confidence score to the gold fact verification label between the statements with that token unmasked and masked. Formally, given the table T, original statement S and its counterfactual version S'_t with the target token t masked, the salience score of t in this statement is

$$salience(t) = \left| P(y|S,T) - P(y|S'_t,T) \right|$$

where y indicates the gold label for fact verification and P is given by the TAPAS model finetuned on TabFact. Larger difference between the predictions for S and S'_t indicates the token is more salient.

3.3 Masked Salient Token Prediction

Salient tokens in statements, such as lexemes that appear in table cells, and those referring to aggregations and their results, directly contribute to table-text alignment and reasoning. Hence, they are critical to table-based fact verification as shown in Fig. 1. Considering the supervision signals for the verification task are sparse and not necessarily sufficient to capture fine-grained table-text alignment and the logical relation, we introduce masked salient token prediction as an auxiliary task.

This task is to predict a masked salient token in an entailed statement given the masked statement and the respective table. We mask the most salient token in each statement according to the salience score estimated in §3.2. The reason to do so is that it is hard to find a general threshold to split tokens in different statements into salient and non-salient groups. The effectiveness of the salience-aware masking will be further evaluated in §4.2.

Both of table-based fact verification and masked salient token prediction share the same TAPAS encoder and the latter reuses the embedding layer as the language modeling head (i.e. linear layer with weights tied to the input embeddings). In this way, all parameters that are updated for the auxiliary task are shared with those in the main task. Both tasks are jointly learned, so that the auxiliary task seeks to provide indirect supervision signals to improve the main task. The objective function and training details are described in §3.5.

3.4 Salience-aware Data Augmentation

To effectively learn a robust and generalized NLI model to verify statements based on tables, one requirement is sufficient training data. Previous work has explored augmenting data by filling in specific statement generation templates with entities or values from the table (Eisenschlos et al., 2020). These selected tokens are always detected as salient tokens by the method described in §3.2 as they are important to fact verification. However, previous works ignored the fact that the statement can be presented in heterogeneous ways, and a reliable table-based fact verification model should also be adaptive and robust to heterogeneous statements. In this context, it is intuitive to consider that the non-salient tokens should not interfere the meaning and evidential support of a statement. Accordingly, we introduce an efficient probabilistic data augmentation technique that leverages the salience of tokens from the other perspective.

We augment training data by replacing the least salient token in each statement with reasonable alternatives. Since we expect non-salient token substitution to cause inconsequential meaning change to the original statement, such automatically generated instances will be augmented into the training data along with the original labels. Similar to §3.3, we select the least salient token to augment, because it is hard to find a fixed threshold that works for all statements to justify whether each of their tokens is important enough or not.

In detail, for each human-annotated statement, we mask the least salient token and request a BERT model to provide the top k tokens to fill in the blank. Each a filled token gives an augmented instance of statement. BERT is pretrained on large textual corpora with the MLM objective, so its predictions can reflect the real-world language expressions[2]. Considering the top k token substitutions are not equally confident according to the BERT predictions and potential noise in data augmentation, we down-weight each augmented data instance in training according to the token prediction probabilities (denoted by w_{ij} for the j-th augmented instance derived from the i-th original instance). Related details are presented shortly in §3.5.

3.5 Training and Inference

We train the model to jointly conduct the main table-based fact verification task (§3.1) using augmented data described in §3.4 along with the auxiliary task of masked salient token prediction (§3.3).

In detail, there are two learning objectives: the binary classification objective L_v for the main task and the MLM objective L_m for the auxiliary task. For fact verification, we denote the gold label of the i-th instance in the original dataset as y_i (1 for entailed and 0 for refuted). With salience-aware data augmentation, each original instance in the dataset is augmented to $k+1$ instances (including itself). The training instances are also assigned with the probability-based training weight w_{ij} as described in §3.4 ($w_{i0} = 1$ for the original instance). Then, given the model prediction $p_{ij} \in [0, 1]$ on each instance, the loss function is defined as the following weighted cross-entropy, where N_v is the number of

[2]We do not use TAPAS for data augmentation because the table is not used as input for masked sentence completion.

instances in the original dataset:

$$L_v = -\sum_{i=1}^{N_v}\sum_{j=0}^{k} w_{ij}(y_i \log(p_{ij}) + (1-y_i)\log(1-p_{ij})).$$

For the auxiliary task, given the gold label y_i^j (1 for the target token, 0 for other tokens) and model outputs p_i^j of each candidate token $c_j \in V$ for the i-th instance, the loss function is defined as below, where N_m is the number of all entailed statements in the dataset:

$$L_m = -\sum_{i=1}^{N_m}\sum_{j=1}^{|V|} y_i^j \log(p_i^j).$$

The overall learning objective is to optimize the following joint loss, where α is a coefficient to balance between the two task objectives:

$$L = \frac{\alpha}{N_v k} L_v + \frac{(1-\alpha)}{N_m} L_m.$$

In inference, given a statement and a table, we use the prediction head of fact verification independently and perform the verification without augmenting the test data, following the details in §3.1.

4 Experiment

In this section, we conduct experiments on the TabFact dataset. We first introduce the dataset, a series of recent baselines and details of our method (§4.1). Then we show the overall performance and ablation results (§4.2). We also provide case studies for in-depth analysis (§4.3).

4.1 Experimental Settings

Dataset and Evaluation. We evaluate our model on the TabFact benchmark (Chen et al., 2020a) that is widely used by studies on this task[3]. The dataset contains 118,275 statements and 16,573 tables. Each table thereof comes along with 2 to 20 statements, and consists of 14 rows and 5 columns in average. Each statement is paired with a table and is labeled as entailed or refuted by information in the table. We use the originally released train, validation and test splits for evaluation, for which the statistics are listed in Tab. 1. Tables in these splits do not have overlaps. Specifically, statements in the test split are further labeled into simple or complex categories according to their verification difficulty. Additionally, a small subset of the test split is used to compare machine performance and human performance. Being consistent with previous studies (Chen et al., 2020a; Zhong et al., 2020; Yang et al., 2020; Eisenschlos et al., 2020), we report the model performance on the validation and test splits, two of the difficulty-specific subsets, as well as the small subset with human performance, and use accuracy as the evaluation metric.

Split	#Statement	# Table
Train	92,283	13,182
Validation	12,792	1,696
Test	12,799	1,695
Simple	50,244	9,189
Complex	68,031	7,392

Table 1: Statistics of the TabFact dataset.

Baselines. We compare our system with the following competitive baselines:

- Latent Program Algorithm (LPA) (Chen et al., 2020a) synthesizes logical programs based on the given statement and table, executes programs to return bool labels, and aggregates the results according to the confidence score of each program assigned by a Transformer-based model.
- LogicalFactChecker (Zhong et al., 2020) captures token-level semantic interaction between a statement, a table and a derived program using BERT with graph-based masking. Logical semantics of each program is captured with neural module networks (Andreas et al., 2016).
- HeterTFV (Shi et al., 2020) constructs a heterogeneous graph to incorporate the statement, the table and the program, and applies a heterogeneous graph attention network to capture both linguistic and symbolic information.
- ProgVGAT (Yang et al., 2020) generates a program and verbalize the execution progress as evidences. The system applies a graph attention network (Veličković et al., 2017) to capture the execution graph, the table and statement.
- Table-BERT (Chen et al., 2020a) applies BERT for NLI taking a statement as the hypothesis and a linearized table as the premise.
- TAPAS (Herzig et al., 2020) is a Transformer-based model pre-trained on textual and tabular data. Dong and Smith (2021) and Eisenschlos et al. (2020) have formulated table-based

[3] https://tabfact.github.io/

Model	Val	Test	Test (simple)	Test (complex)	Small Test Set
Human Performance	-	-	-	-	92.1
LPA	65.2	65.0	78.4	58.5	68.6
LogicalFactChecker	71.8	71.7	85.4	65.1	74.3
HeterTFV	72.5	72.3	85.9	65.7	74.2
ProgVGAT	74.9	74.4	88.3	67.6	76.2
Table-BERT	66.1	65.1	79.1	58.2	68.1
TAPAS (Dong and Smith, 2021)	-	76.0	89.0	69.8	-
TAPAS (Eisenschlos et al., 2020)	81.0	81.0	92.3	75.6	83.9
ours	**82.7**	**82.1**	93.3	**76.7**	84.3
– w/o augmented data	82.4	82.1	93.4	76.6	**84.4**
– w/o auxiliary task	81.8	81.9	**93.6**	76.3	84.1

Table 2: Performance on the official splits of TabFact in terms of verification accuracy (%). Baselines are organized into *logical program-driven* (i.e. LPA, LogicalFactChecker, HeterTFV and ProgVGAT) and *non-logical program-driven* (i.e. Table-BERT and TAPAS). Human performance is reported by Chen et al. (2020a).

fact verification as an NLI task, and applied TAPAS with task-specific intermediate pretraining. The latter one achieves the current SOTA performance on TabFact.

Model Configurations. Our system also adopts the officially released TAPAS-Large model, which applies intermediate pre-training and is fine-tuned on TabFact, as our basic model[4]. Following Eisenschlos et al. (2020), we set the max input length to 512. We use $10,000$ training steps, and optimize the learning objective with an AdamW optimizer (Loshchilov and Hutter, 2019) which sets the learning rate to $5e^{-5}$, a batch size of 32 and a warmup ratio of 0.1. All hyper-parameters are decided according to the validation performance. For multi-task learning, we set the coefficient between two losses α to 0.5. For data augmentation, we use the uncased BERT-Large model as the MLM. For computational efficiency, we select the top $k = 3$ predictions for probabilistic data augmentation.

4.2 Results

Overall Performance. Tab. 2 presents the results of different verification models. Among the baseline methods, TAPAS with task-specific intermediate pretraining demonstrates the best performance. It implies that explicit logical programs is not a necessity for reasoning between the table and the statement. We observe that our system outperforms the best baseline with 2.1% relative improvement on the validation set and 1.4% relative improvement on the test set in terms of accuracy. It is noteworthy that, our system applies the same backbone model and pretraining process as the previous best method, so that all the improvements are attributed to the salience-aware learning strategies. Besides, our system reduces the gap between machine performance and human performance on the small test set to 7.8%. These experimental results verify our hypothesis that masked salient token prediction and salience-aware data augmentation are conducive to table-based fact verification.

	Strategy	Val	Test
Masking	Random	82.1	81.9
	Salient	**82.4**	**82.1**
Augmentation	Uniform	81.5	81.3
	Probabilistic	**81.8**	**81.9**

Table 3: Ablation results for masking strategy and augmentation strategy. To avoid co-effects, we conduct experiments on masking (or augmentation) strategy without using augmented data (or auxiliary task).

Effect of Masked Salient Token Prediction. The performance of the base model with masked salient token prediction is marked as "w/o augmented data" in Tab. 2. The auxiliary task solely brings along 1.7% relative improvement on the validation set and 1.4% relative improvement on the test set. This demonstrates that the indirect supervision brought by the auxiliary task can directly benefit the main task training. Tab. 3 compares salient masking and random masking for the auxiliary task. For fair comparison, we mask one token in each entailed statement for both strategies. The results show that salient masking reduces error rate on the validation set by relative 1.7% (and by relative 1.1% on the test set) in comparison with random masking. This is not surprising since random masking may mask

[4]https://github.com/google-research/tapas

Token Salience Estimation	Augmentation
The file format mobipocket comes with all three supports .	works 0.615 worked 0.051 compatible 0.029
The player from the Santiago province lives in the city navarrete .	population 0.352 people 0.184 one 0.035
Canton , Ohio was the location for the event , fightfest 2 , which lasted only 3 rounds .	of 0.709 to 0.001 in 0.001

Table 4: Examples of salience estimation and data augmentation. Darker background indicates more salience. Blue rectangles mark the targeted most salient tokens in masked salient token prediction. Red rectangles mark the least salient tokens that are to be substituted by the augmentation tokens, for which weights are listed.

non-salient tokens which are not decisive for table-text alignment and logical inference.

Effect of Salience-aware Data Augmentation. The performance of the base model with salient-aware data augmentation is marked as "w/o auxiliary task" in Tab. 2. The data augmentation independently brings 1.0% relative improvement on the validation set and 1.1% relative improvement on the test set. The results demonstrate that table-based fact verification requires abundant training data and verify the effectiveness of the proposed data augmentation strategy. Tab. 3 compares probabilistic weights and uniform weights. The results show that probabilistic data augmentation reduces error rate on the validation set by 1.6% relatively (and by 3.2% relatively on the test set) in comparison with uniform data augmentation. This observation is reasonable because the augmented data are not equally confident according to the MLM predictions. Moreover, the predicted probabilities from the pretrained language model correlate with real-world distribution of English language.

Performance on Simple and Complex Instances. We further compare the performance of baselines and variants of our system on two groups of test instances labeled with different verification difficulties. Our system outperforms all the baselines on both simple and complex instances with at least 1.0% absolute improvement. Ablation results in Tab. 2 also show that the auxiliary task improves the base model more on complex instances while data augmentation improves the base model more on simple instances. These results are consistent with the features of the two salience-aware learning strategies. Masked salient token prediction seeks to enhance the model to capture table-text alignment and the underlying logical relations, so that complex instances requiring more complicated reasoning gain more benefits. Salience-aware data augmentation seeks to augment statements by simply replacing non-salient tokens. This strategy increases the training data but does not augment the implicit logical form covered by the dataset so that the improvement on complex instances is not as significant as that on simple instances.

4.3 Case Study

We present a case study with three representative examples to illustrate salience estimation and data augmentation in Tab. 4. The detected salient tokens can be entities and numeric values from the table, tokens indicating relations, and the results of logical operations. Non-salient tokens can be common nouns, verbs, prepositions and so on. These tokens are detected as non-salient because they are not closely associated with facts in the given table. For example, the table in the second example is about the residence of different athletes, so "`player`" in the statement may be substituted to related terms without interfering the verification result. It is noteworthy that entities consisting of multiple words tend to have relatively small salience scores for some parts. It may be due to that verification models can identify the corresponding cell by part of the entity. But it also raises the risk of incorrect verification or polluted data augmentation when modifying a part of a multi-word entity.

5 Conclusion

In this paper, we proposed a novel system for table-based fact verification. Our system employs salience-aware learning and introduce complementary supervision signals by leveraging both salience

and non-salient tokens from different perspectives. The system consists of three key techniques, including probing-based salience estimation, masked salient token prediction and salience-aware data augmentation. Experiments on the TabFact benchmark show that our system leads to significant improvements over the current SOTA systems. For future work, we plan to extend salience-aware learning to other NLU tasks, including NLI (Bowman et al., 2015; Williams et al., 2018) and Tabular QA (Sun et al., 2016; Chen et al., 2020b). Applying the idea of salience estimation to NLG tasks, such as controlled table-to-text generation (Parikh et al., 2020) and paraphrasing (Iyyer et al., 2018; Huang and Chang, 2021), is another meaningful direction.

Ethical Consideration

This work does not present any direct societal consequence. The proposed work seeks to develop a salience-aware learning framework for fact verification using tabular data as evidence. We believe this leads to intellectual merits that benefit claim and statement verification for Web corpora, as well as detection of misinformation. It potentially also has broad impacts for NLU and NLG tasks where tables serve as a medium of knowledge sources. The experiments are conducted on a widely-used open benchmark.

The goal of this research topic is to help identify misinformation, which seeks to benefit societal fairness. While we treat tables as reliable sources of evidences like relevant studies do, we do not hypothesize that the populated information by Web users in tables is not completely free of societal bias. We believe this is a meaningful research direction for further exploration. While not being explicitly studied in this work, the incorporation of salience-aware inference could be a way to control or mitigate societal biases.

Acknowledgement

We appreciate the anonymous reviewers for their insightful comments.

This research is supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under contract number FA8650-17-C-7715, by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research, and by the National Science Foundation of United States Grant IIS 2105329. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Henry E Brady. 2008. Causation and explanation in social science. In *The Oxford Handbook of Political Science*.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

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

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1026–1036.

Christian Chiarcos, Berry Claus, and Michael Grabski. 2011. Introduction: Salience in linguistics and beyond. In *Salience*, pages 1–28. De Gruyter Mouton.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Rui Dong and David A Smith. 2021. Structural encoding and pre-training matter: Adapting bert for table-based fact verification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2366–2375.

Julian Eisenschlos, Syrine Krichene, and Thomas Mueller. 2020. Understanding tables with intermediate pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 281–296.

Tsu-Jui Fu, Xin Eric Wang, Matthew F Peterson, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. 2020. Counterfactual vision-and-language navigation via adversarial path sampler. In *European Conference on Computer Vision*, pages 71–86. Springer.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, 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.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1022–1033.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921.

David Lewis. 2013. *Counterfactuals*. John Wiley & Sons.

Zhengzhong Liu, Chenyan Xiong, Teruko Mitamura, and Eduard Hovy. 2018. Automatic event salience identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1226–1236.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Sunghyun Park, Seung-won Hwang, Fuxiang Chen, Jaegul Choo, Jung-Woo Ha, Sunghun Kim, and Jinyeong Yim. 2019. Paraphrase diversification using counterfactual debiasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6883–6891.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5046–5056.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736.

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.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web*, pages 771–782.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725.

Philip E Tetlock and Aaron Belkin. 1997. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122.

Chenyan Xiong, Zhengzhong Liu, Jamie Callan, and Tie-Yan Liu. 2018. Towards better text understanding and retrieval through kernel entity salience modeling. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 575–584.

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.

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.

Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. 2019. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 791–800.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280.

Yi Zhang, Zachary Ives, and Dan Roth. 2019. Evidence-based trustworthiness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 413–423.

Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin. 2020. Logicalfactchecker: Leveraging logical operations for fact checking with graph module network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6065.

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.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.