Distilling the Evidence to Augment Fact Verification Models

Beatrice Portelli[1] **Jason Zhao**[2] **Tal Schuster**[2] **Giuseppe Serra**[1] **Enrico Santus**[2,3]

[1] University of Udine [2] CSAIL, MIT
[3] Decision Science and Advanced Analytics for for MAPV & RA, Bayer
portelli.beatrice@spes.uniud.it, jzhao7@mit.edu,
tals@csail.mit.edu, giuseppe.serra@uniud.it, esantus@mit.edu

Abstract

The alarming spread of fake news in social media, together with the impossibility of scaling manual fact verification, motivated the development of natural language processing techniques to automatically verify the veracity of claims. Most approaches perform a claim-evidence classification without providing any insights about why the claim is trustworthy or not. We propose, instead, a model-agnostic framework that consists of two modules: (1) a span extractor, which identifies the crucial information connecting claim and evidence; and (2) a classifier that combines claim, evidence, and the extracted spans to predict the veracity of the claim. We show that the spans are informative for the classifier, improving performance and robustness. Tested on several state-of-the-art models over the FEVER dataset, the enhanced classifiers consistently achieve higher accuracy while also showing reduced sensitivity to artifacts in the claims.

1 Introduction

The increased quantity of information that circulates in social media and on the Web every day, together with the high cost of assessing its veracity, has demanded the application of natural language processing (NLP) techniques to the task of fact verification. In the last years, the NLP community has proposed a large number of datasets and approaches for addressing this task, facing complicated challenges that are still far from being solved.

The task of fact verification can be split into (i) retrieving one or more candidate pieces of evidence; (ii) assessing whether they are either *supporting* or *refuting* a claim, or whether they contains *insufficient* information to state either of the above. In this paper, we mostly focus on the reasoning between the claim and the evidence.

To generate models that work on real world data, fact verification solutions are expected to: (i) perform well not only on synthetic datasets but also in realistic scenarios, where both text form and text content are highly unpredictable; (ii) produce transparent decisions, providing an explanation for their verdict, so that the readers may consider whether trusting them or not.

Claim	Susan Sarandon was nominated for five Emmy Awards.
Evidence	**[wiki/Susan_Sarandon]** On television, she is a five-time Emmy Award nominee, including for her guest roles on the sitcoms Friends 2001 and Malcolm in the Middle (2002), and the TV films Bernard and Doris (2007) and You Don't Know Jack (2010).
Label	SUPPORT
Claim	Fantastic Beasts and Where to Find Them was released only in North America on November 18, 2016.
Evidence	**[wiki/Fantastic_Beasts_and_Where_to_Find_Them_(film)]** Fantastic Beasts and Where to Find Them premiered in New York City on 10 November 2016 and was released worldwide on 18 November 2016 in 3D, IMAX 4K Laser and other large format cinemas.
Label	REFUTE
Claim	Ian Brennan is a film screenwriter.
Evidence	**[wiki/Ian_Brennan_(writer)]** Ian Brennan (born April 23, 1978) is a television writer, actor, producer and director.
Label	NOT ENOUGH INFORMATION

Figure 1: Examples of claim-evidence pairs from the FEVER dataset. The evidence spans extracted by our system are underlined and presented in color.

To address these two requirements, we propose a model-agnostic framework that includes two modules: (i) a span extractor that aims to identify in the evidence the pieces of relevant information that are informative with respect to the claim; (ii) a classifier that uses the claim, evidence and extracted spans to predict whether the evidence is *supporting*, *refuting* or containing *insufficient* information. The spans extracted by the first module are useful to enhance the classifier and inform the user. Humans can in fact exploit the spans to effectively understand why a claim is true or false.

We evaluate our pipeline with three highly performing neural models on the FEVER dataset (Thorne et al., 2018), comparing the uninformed to the informed setting. While this dataset includes

Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), pages 47–51

ground truth for both evidence retrieval and evidence classification, in this paper we only exploit the latter annotations. Our experiments show that the models informed with the extracted spans consistently achieve higher performance than their uninformed counterparts, demonstrating the usefulness of spans. We also evaluate our models on the challenging SYMMETRIC FEVER dataset (Schuster et al., 2019), which tests system's robustness in absence of FEVER's artifacts. We find the models trained with our pipeline to achieve higher accuracy.

Finally, we assess the quality of the extracted spans as decision rationales to be shown to end-user. Manually examining a subset of outputs shows that 67% of the *support* and 88% of the *refute* spans are well explanatory with respect to the decision, leading to an aggregated score of 75%.

2 Related Work

Fake news detection has recently gained interest in the NLP community. Most of the initial works have focused on style (Feng et al., 2012) and linguistic approaches (Pérez-Rosas and Mihalcea, 2015). Despite the good performance in synthetic datasets, these methods failed when applied to real-world data. New approaches based on fact verification over retrieved evidence have therefore taken the stage in the literature.

Datasets. Several fact verification datasets were developed over the last decade. Vlachos and Riedel (2014) created a dataset which consisted of 221 statements and hyperlinks to pieces of evidence of various formats. Many datasets were created in the following years, with collections of claims of increasing size and various kinds of additional information. Among them Ferreira and Vlachos (2016)'s debunking dataset (300 rumoured claims and 2,595 associated news articles) and Wang (2017)'s LIAR dataset (12,836 short statements labeled for veracity, topic and various metadata on the speaker). In the last years, most systems have been developed over FEVER (Thorne et al., 2018), a large-scale dataset for Fact Extraction and VERification that consists of 185,445 claims and their related evidence, labeled as either supporting, refuting or not containing enough information.

Approaches. There has been a large development since the first approaches for fact verification (Ferreira and Vlachos, 2016; Wang, 2017; Long et al., 2017). To provide a strong baseline for FEVER, Thorne et al. (2018) proposed a pipeline consisting of document and sentence retrieval and a multi-layer perceptron as textual entailment recognizer. More sophisticated models followed. Among them, the Bi-Directional Attention Flow (BiDAF) network (Seo et al., 2016a), originally introduced for machine comprehension, has been recently adapted to the task of fact verification (Tokala et al., 2019). BiDAF combines LSTMs with both a context-to-query and query-to-context attention, to produce a query-aware context representation at multiple hierarchical levels. Nie et al. (2019) introduced the Neural Semantic Matching Networks (NSMNs), which aligns two encoded texts and computes the semantic matching between the aligned representations with LSTMs and used it to earn the first place in the first competitions organized on the FEVER dataset. Soleimani et al. (2019) exploits the contextualized representations of a pre-trained BERT (Devlin et al., 2019) model for both sentence selection and fact verification.

3 Method

Given a claim $C = \{c_1, \ldots, c_n\}$ and a piece of evidence $E = \{e_1, \ldots, e_m\}$, two word sequences of length n and m respectively, the fact verification problem requires to predict the relation $rel = \{(S)\text{upports}, (R)\text{efutes}, (I)\text{nsufficient}\}$ between E and C.

Framework. We propose a pipeline of two modules: a span extractor M_{span} and a classifier $M_{\text{classifier}}$. The goal of $M_{\text{span}}(C, E)$ is to identify polarizing pieces of information $\{e_{i_1}, \ldots, e_{i_N}\}$ in E without which $rel(E, C)$ would be neutral (i.e. C would neither be entailed nor contradicted by E). The identified pieces of information are passed to $M_{\text{classifier}}$, together with C and E, to perform a three-label classification aimed at predicting $rel(E, C)$: $M_{\text{classifier}}(C, E, \{e_{i_1}, \ldots, e_{i_N}\}) = l \in \{S, R, I\}$.

3.1 Span Extractor

We utilize the TokenMasker architecture from Shah et al. (2020) for M_{span}. This masker was developed to identify the minimal group of tokens without which E would be neutral with respect to C. M_{span} is trained by getting feedback from a pre-trained neutrality classifier. Shah et al. (2020) use the ESIM model with GloVe embeddings trained on FEVER as a neutrality classifier. We choose to use the RoBERTa model (Liu et al., 2019) in-

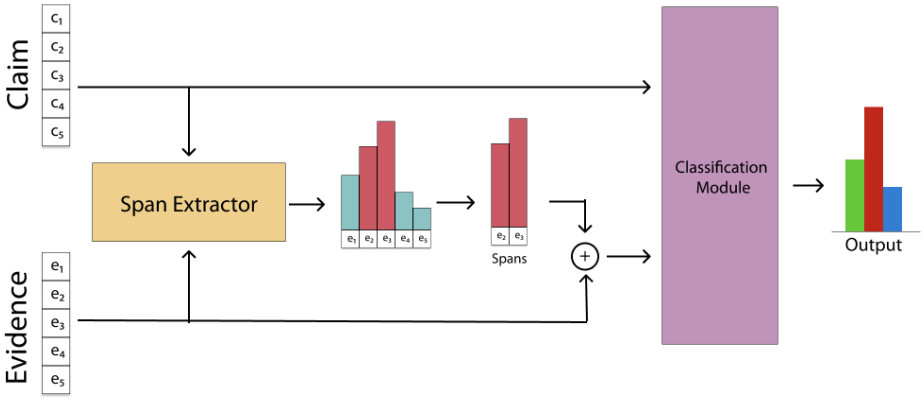

Figure 2: Framework outline: (i) the claim and the evidence pass through the span extractor, which quantifies the relative importance of their words; (ii) claim, evidence and spans are then passed to the classification module, which decides whether the evidence is supporting, refuting or insufficient to judge the claim.

stead, pretrained on an entailment task over a multi-genre corpus (i.e. three-label classification: entailment/neutral/contradiction on the MULTINLI dataset (Williams et al., 2018)).

The choice of using a rationale-style extractor (Shah et al., 2020) is due to its ability to provide informative spans that can be used as explanations to the relation of the evidence with the claim. This approach was shown to perform better than simply relying on the internal attention weights of a classifier (Lei et al., 2016; Jain and Wallace, 2019).

3.2 Classifiers

To test our assumption, we consider three neural network architectures that have achieved the best performance on the first FEVER shared Task recently: BiDAF (Seo et al., 2016b), NSMN (Nie et al., 2019) and BERT (Devlin et al., 2019). Note that the architecture of $M_{\text{classifier}}$ is independent of M_{span}. The spans extracted by $M_{\text{classifier}}$ are forwarded to the classifier by concatenating them to the original evidence, followed by a separator token.

BiDAF consists of four layers: (i) the embedding layer, which encodes two raw text sequences (i.e. C and E) into two vector sequences $\hat{C}$ and $\hat{E}$; (ii) the attention layer, which computes the attention scores between the two sequences and returns two attended sequences C_A and E_A; (iii) the modeling layer, which takes C_A and E_A as input and outputs two fixed size vectors, $\hat{C}_A$ and $\hat{E}_A$, that capture the semantic similarity between the original sequences; and (iv) the output layer, which takes $\hat{C}_A$ and $\hat{E}_A$ and returns the output labels.

NSMN encodes C and E into vector sequences $\hat{C}$ and $\hat{E}$, similarly to BiDAF. It then applies an alignment layer, which computes the alignment matrix, $\mathbf{A} = \hat{C}^T\hat{E}$, and the aligned representations, C_A and E_A, using $\hat{C}$, $\hat{E}$, $\mathbf{A}$. It follows a matching layer, which performs semantic matching using LSTM between C_A and $\hat{C}$, as well as E_A and $\hat{E}$, to output matching matrices $\mathbf{M_C}$ and $\mathbf{M_E}$, which are finally pooled by the output layer and mapped to output labels.

BERT (we use the base-uncased version) consists of 12 encoder layers with self-attention $(enc_1, \ldots, enc_{12})$ and one classification layer. Each encoder enc_i takes an input sequence I_{i-1} and outputs I_i, a sequence of the same length where each token is replaced with an embedding capturing its relationship with the other words in I_{i-1}. The output of enc_i becomes the input of enc_{i+1}. I_0 is set as the concatenation of C and E, preceded by the special `[CLS]` token. The output of the last encoder enc_{12} is therefore an highly embedded representation of C and E. It is passed to the classification layer which maps the representation of the `[CLS]` token to the output labels.

4 Experiments

We evaluate the three classifiers described in section 3 in two conditions: uninformed (W/O) and informed (With), where the latter refers to the utilization of the information extracted by M_{span}.

4.1 Data

We use the FEVER dataset to train all of our classifiers. We evaluate the classifiers both on FEVER and on SYMMETRIC FEVER.

FEVER dataset (Thorne et al., 2018): the current

largest available Wikipedia-based dataset, consisting of 185,445 claims. Each claim is matched with supporting or refuting evidence from Wikipedia or with a "not enough information" label.

We use the development set from FEVER's shared-task as our test set (containing 19,998 samples). We randomly split FEVER's training set into our training and validation sets. Following this process, we have 125,451 samples in our training set (73,369 support, 23,109 refute, and 28,973 insufficient information).

While evidence sentences for supporting and refuting examples are provided in the ground truth, those for the "insufficient information" were obtained by us. We use the document retrieval module of the best performing system on the first FEVER Shared Task (Nie et al., 2019). Given a claim and the Wikipedia dump provided with the FEVER dataset, this document retrieval module returns a list of Wikipedia articles which are possibly related to the claim, ranked with a score calculated by comparing the claim, the title of the article and its first sentence. We keep the highest scoring document. Thereafter, we pick the sentence with the highest TF-IDF similarity with the claim. Also, to disambiguate pronouns, we extend all evidence sentences by appending the title of their Wikipedia page.

SYMMETRIC FEVER (Schuster et al., 2019): a smaller unbiased extension of FEVER, consisting of 712 claim-evidence pairs which were synthetically generated from FEVER to remove strong cues in the claims which could allow predicting the label without looking at the evidence (give-away phrases).

4.2 Hyperparameters

TokenMasker is trained on the same dataset and configuration as Shah et al. (2020). However, we replace their neutrailty classifier with a RoBERTa classifier, pretrained on MNLI. This model is trained once and used in inference mode for all subsequent experiments.

BiDAF is trained for 12 epochs using cross entropy loss and Adam optimizer with initial learning rate 1e-3. We use a dropout probability of 0.2 and a batch size of 8.

NSMN is trained for 12 epochs using cross entropy loss and Adam optimizer with initial learning rate 1e-4. We use a dropout probability of 0.5 and a batch size of 8.

BERT is fine-tuned for 8 epochs using cross entropy loss and Adam optimizer with initial learning rate 2e-5. We use a dropout probability of 0.1 and a batch size of 16.

These hyperparemeters were found to achieve the highest accuracy on our validation set. For our final classifiers, we fix these settings and retrain them using the full FEVER training set.

Model	W/O	With	Test set
BiDAF	73.90%	*75.12%	FEVER
NSMN	72.88%	**74.56%	
BERT	84.16%	84.33%	
BiDAF	49.16%	**52.24%	SYMMETRIC
NSMN	53.35%	54.56%	
BERT	71.12%	71.49%	

Table 1: Accuracy of the models on the FEVER and the SYMMETRIC datasets. Results for BERT are the average over 5 runs with the same hyperparameters. Significance: * if $p < 0.1$, ** if $p < 0.05$.

4.3 Results

Table 1 shows the results obtained in our experiments on both FEVER and the SYMMETRIC dataset. Scores are much higher in the first dataset as the systems can rely on give-away phrases, some words in the claims which have a high correlation with the correct output label regardless of the evidence. This situation does not exist in the SYMMETRIC dataset, where the give-away phrases have been eliminated. As expected, all systems perform worse on this dataset, but the drop in performance is more significant for the uninformed models (W/O) than for the informed (With) ones. In fact, the informed models consistently perform better than the uninformed ones (W/O), often obtaining statistical significance. While the difference in performance between W/O and With is particularly relevant for BiDAF and NSMN, it thins for BERT, which is already a strong classifier leveraging on a robust pretraining.

Output Explainability. We also manually evaluated the spans for 100 randomly extracted claim-output pairs, to assess whether they represented an understandable explanation for the verdict. The spans were deemed explanatory in 88% of the cases for *refute* claims and 67% of the *support* claims, which leads to an aggregated score of 75%. The extracted spans are therefore not only informative to the classifier, but can also be used to produce human-readable justifications for a positive or negative relation.

5 Conclusions

This paper has introduced a classifier-agnostic framework that allows fact verification models to improve their performance and robustness, utilizing concise spans of the available evidence sentences. The experiments have shown that the extracted spans are indeed informative for the final classifier, supporting the usefulness of the framework. Furthermore, this work opens the possibility of providing to the human users a justification for the model's predictions.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *In NAACL-HLT*.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *ACL*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *HLT-NAACL*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL-HLT*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *IJCNLP*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *In AAAI*.

Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *EMNLP*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *In EMNLP-IJCNLP*.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016a. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016b. Bidirectional attention flow for machine comprehension. *ArXiv*, abs/1611.01603.

Darsh J Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *In AAAI*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. Bert for evidence retrieval and claim verification.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *In NAACL-HLT*.

Santosh Tokala, Vishal G, Avirup Saha, and Niloy Ganguly. 2019. AttentiveChecker: A bi-directional attention flow mechanism for fact verification. In *In NAACL-HLT*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *LTCSS@ACL*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *In NAACL-HLT*.