Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks

Jing Ma[1], **Wei Gao**[2], **Shafiq Joty**[3,4], **Kam-Fai Wong**[1,5]
[1]The Chinese University of Hong Kong, Hong Kong SAR
[2]Victoria University of Wellington, New Zealand
[3]Nanyang Technological University, Singapore
[4]Salesforce Research Asia, Singapore
[5]MoE Key Laboratory of High Confidence Software Technologies, China
[1]{majing,kfwong}@se.cuhk.edu.hk
[2]wei.gao@vuw.ac.nz, [3]srjoty@ntu.edu.sg

Abstract

Claim verification is generally a task of verifying the veracity of a given claim, which is critical to many downstream applications. It is cumbersome and inefficient for human fact-checkers to find consistent pieces of evidence, from which solid verdict could be inferred against the claim. In this paper, we propose a novel end-to-end hierarchical attention network focusing on learning to represent coherent evidence as well as their semantic relatedness with the claim. Our model consists of three main components: 1) A coherence-based attention layer embeds coherent evidence considering the claim and sentences from relevant articles; 2) An entailment-based attention layer attends on sentences that can semantically infer the claim on top of the first attention; and 3) An output layer predicts the verdict based on the embedded evidence. Experimental results on three public benchmark datasets show that our proposed model outperforms a set of state-of-the-art baselines.

1 Introduction

The increasing popularity of social media has drastically changed how our daily news are produced, disseminated and consumed.[1] Without systematic moderation, a large volume of information based on false or unverified claims (e.g., fake news, rumours, propagandas, etc.) can proliferate online. Such misinformation poses unprecedented challenges to information credibility, which traditionally relies on fact-checkers to manually assess whether specific claims are true or not.

Despite the increased demand, the effectiveness and efficiency of human fact-checking is handicapped by the volume and fast pace the noteworthy claims being produced on daily basis. Therefore, it is an urgent need to automate the process and ease the human burden in assessing the veracity of claims (Thorne and Vlachos, 2018).

Not surprisingly, various methods for automatic claim verification have been proposed using machine learning. Typically, given the claims, models are learned from auxiliary relevant sources such as news articles or social media responses for capturing words and linguistic units that might indicate viewpoint or language style towards the claim (Jin et al., 2016; Rashkin et al., 2017; Popat et al., 2017; Volkova et al., 2017; Dungs et al., 2018). However, the factuality of a claim is independent of people's belief and subjective language use, and human perception is unconsciously prone to misinformation due to the common cognitive biases such as naive realism (Reed et al., 2013) and confirmation bias (Nickerson, 1998).

A recent trend is that researchers are trying to establish more objective tasks and evidence-based verification solutions, which focus on the use of evidence obtained from more reliable sources, e.g., encyclopedia articles, verified news, etc., as an important distinguishing factor (Thorne and Vlachos, 2018). Ferreira and Vlachos (2016) use news headlines as evidence to predict whether it is for, against or observing a claim. In the Fake News Challenge[2], the body text of an article is used as evidence to detect the stances relative to the claim made in the headline. Thorne et al. (2018a) formulate the Fact Extraction and VERification (FEVER) task which requires extracting evidence from Wikipedia and synthesizing information from multiple documents to verify the claim. Popat et al. (2018) propose DeClarE, an evidence-aware neural attention model to aggregate salient words from source news articles as the

[1] The latest Pew Research statistics show that 68% American adults at least occasionally get news on social media. http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

[2] http://www.fakenewschallenge.org/

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2561–2571
Florence, Italy, July 28 - August 2, 2019.

c:	The test of a 5G cellular network is the cause of unexplained bird deaths occurring in a park in The Hague, Netherlands. *Verdict*: ***False***
s_1:	**[*Contradict*]**: Lots of tests going on with it in the Netherlands, but there haven't been test done in The Haque during the time that the mysterious starling deaths occurred.
s_2:	**[*Contradict*]**: One such test did occur in an area generally near Huijgenspark, but it took place on 28 June 2018.
s_3:	**[*Entail*]**: It's not clear whether tests with 5G have been carried out again, but so far everything points in the direction of 5G as the most probable cause.
s_4:	**[*Neutral*]**: Between Friday, 19 Oct and Saturday, 3 Nov 2018, 337 dead starlings and 2 dead common wood pigeons were found.
s_5:	**[*Entail*]**: The radiation created on the attempt of 5G cellular networks are not harmful only for birds but also for humans too.
s_6:	**[*Neutral*]**: 5G network developers promise faster data rates in addition to reduce energy and financial cost.
s_7:	**[*Neutral*]**: Parts of the park are blocked and dogs are no longer allowed to be let out, the dead birds are always cleaned up as quickly as possible.

Table 1: Sentences topically coherent (s_1–s_4) and not coherent (s_5–s_7) with each other relative to the claim c, where their semantic entailment relations (i.e., entail, contradict, neural) with c are shown.

main evidence to obtain claim-specific representation based on the attention score of each token.

Inspired by the FEVER task (Thorne et al., 2018a) and DeClarE (Popat et al., 2018), we propose our approach to claim verification by using representation learning to embed sentence-level evidences based on coherence modeling and natural language inference (NLI). The example in Table 1 illustrates our general idea: given a claim "*The test of a 5G cellular network is the cause of unexplained bird deaths occurring in a park in The Hague, Netherlands*" and its relevant articles, we try to embed into the claim-specific representation those evidential sentences (e.g., s_1–s_4) that are not only topically coherent among themselves considering the claim, but could also semantically infer the claim based on textual entailment relations such as entail, contradict, and neutral. It is hypothesized that sentence-level evidence can convey more complete and deeper semantics, thus providing stronger NLI capacity between claim and evidence, which would result in better claim-specific representation for the more accurate fact-checking decision.

In this work, we propose an end-to-end hierarchical attention network for sentence-level evidence embedding that aims to attend on important sentences (i.e., evidence) by considering their topical coherence and semantic inference strength. Different from DeclarE (Popat et al., 2018), our model can determine the verdict of a claim more reasonably with evidential sentences embedded into the learned claim representation. Meanwhile, with the help of attention, crucial evidence can be highlighted and referred for better interpretability of the verdict. Our model is also advantageous over pipeline methods such as Neural Semantic Matching Network (NSMN) (Nie et al., 2019) which topped the FEVER shared task (Thorne et al., 2018b), because our model can be trained to address evidence representation learning directly rather than rank and select sentences semantically similar to the claim. Our contributions are summarized as follows:

- We propose a novel claim verification framework based on hierarchical attention neural networks to learn sentence-level evidence embeddings to obtain claim-specific representation.
- We use a co-attention mechanism to model sentence coherence and integrate the coherence- and entailment-based attentions into our proposed hierarchical attention framework for better evidence embedding.
- We experimentally confirm that our method is much more effective than several state-of-the-art claim verification models using three public benchmark datasets collected from snopes.com, politifact.com and Wikipedia.

2 Related Work

The literature on fact-checking and credibility assessment has been reviewed by several comprehensive surveys (Shu et al., 2017; Zubiaga et al., 2018; Kumar and Shah, 2018; Sharma et al., 2019). We only briefly review prior works closely related to ours.

Many studies on claim verification extracted veracity-indicative features that can reflect stances and writing styles from relevant texts such as news articles, microblog posts, etc. and used the traditional supervised models to learn the parameters (Castillo et al., 2011; Qazvinian et al., 2011; Rubin et al., 2016; Ferreira and Vlachos, 2016; Rashkin et al., 2017). Deep learning models such as recurrent neural networks (RNN) (Ma et al., 2016), convolutional neural networks (CNN) (Wang, 2017) and recursive neural networks (Ma et al., 2018) were also exploited to learn the feature representations.

More recently, semantic matching methods were proposed to retrieve evidence from relatively trustworthy sources such as checked news and Wikipedia articles. Popat et al. (2018) attempted to debunk false claims by learning claim representations from relevant articles using an attention mechanism to focus on *words* that are closely related to the claim. Following NLI (Bowman et al., 2015), which is a task of classifying the relationship between a pair of sentences, composed by a premise and a hypothesis, as Entails, Contradicts or Neutral, Thorne et al. (2018a) formulated claim verification as a task that aims to classify claims into Supported, Refuted or Not Enough Info (NEI). They released a large dataset containing mutated claims based on relevant Wikipedia articles and developed a basic pipeline with document retrieval, sentence selection, and NLI modules. Similar pipelines were developed by most of the participating teams (Nie et al., 2019; Padia et al., 2018; Alhindi et al., 2018; Hanselowski et al., 2018) in FEVER shared task (Thorne et al., 2018b). Apart from the document retrieval function, our model is end-to-end and aims to learn sentence-level evidence with a hierarchical attention framework.

Attention is in general used to attend on the most important part of texts, and has been successfully applied in machine translation (Luong et al.), question answering (Xiong et al., 2016) and parsing (Dozat and Manning, 2016), and is adopted in our model for attending on important sentences as evidence. Our work is also related to coherence modeling. Different from traditional coherence studies focusing on discourse coherence among sentences that are widely applied in text generation (Park and Kim, 2015; Kiddon et al., 2016) and summarization (Logeswaran et al., 2018), we try to capture evidential sentences topically coherent not only among themselves but also with respect to the target claim.

3 Problem Statement

We define a claim verification dataset as $\{\mathcal{C}\}$, where each instance $\mathcal{C} = (y, c, S)$ is a tuple representing a given claim c which is associated with a ground-truth label y and a set of n sentences $S = \{s_i\}_{i=1}^{n}$ from the relevant documents of the claim. We assume the relevant documents are retrieved from text collections containing variable number of sentences, and we disregard the order of sentences and which documents they are from. Our task is to classify an instance into a class defined by the specific dataset, such as veracity class labels, e.g., *True/False*, or NLI-style class labels, e.g., *Supported/Refuted/NEI*.

Our approach exploits and integrates two core semantic relations: 1) coherence of the sentences given the claim; 2) entailment relation between the claim and each sentence, which are described more specifically below.

Coherence Evaluation: According to the coherence theory of truth, the truth of any (true) proposition consists in its coherence with some specified set of propositions (Young, 2018). In order to focus on the useful evidence in a set of relevant sentences S, we propose a coherence-based attention component by cross-checking if any sentence $s_i \in S$ coheres well with the claim and with other sentences in S in terms of topical consistency.

Textual Entailment: Entailment is used to measure whether a piece of evidence semantically infers a given claim. We propose an entailment-based attention component that can be pre-trained to capture entailment relations (Dagan et al., 2010; Bowman et al., 2015) based on sentence pairs labeled with NLI-specific classes: *entails*, *contradicts* and *neutral*. This pre-trained component together with the entire claim verification framework then will be trained end-to-end to attend on the salient sentences for inferring the claim.

4 End-to-End Claim Verification Model

In this section, we introduce our end-to-end hierarchical attention network for claim verification, which consist of two attention layers, i.e., coherence-based attention and entailment-based attention, for learning evidence embeddings. Figure 1 gives an overview of our framework, which will be depicted in detail in the subsections.

4.1 Sentence Representation

Given a word sequence $T = (w_1 \dots w_t \dots w_{|T|})$ which could be either a claim or a sentence, each $w_t \in \mathbb{R}^d$ is d-dimensional vector which can be initialized with pre-trained word embeddings. We map each w_t into a fixed-sized hidden vector using standard GRU (Cho et al., 2014). We then obtain the sentence-level representation for a claim c and each sentence $s_i \in S$ using two GRU-based RNN

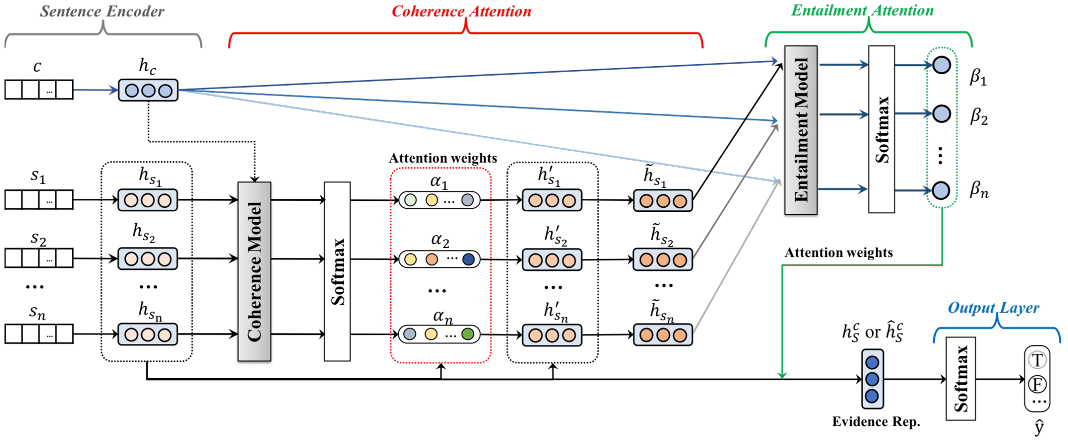


Figure 1: Our end-to-end hierarchical attention networks for claim verification.

encoders (one for c and the other for s_i):

$$\begin{aligned} h_c &= h_{|c|} = \mathrm{GRU}(w_{|c|}, h_{|c|-1}, \theta_c) \\ h_{s_i} &= h_{|s_i|} = \mathrm{GRU}(w_{|s_i|}, h_{|s_i|-1}, \theta_S) \end{aligned} \tag{1}$$

where $|.|$ denotes the number of words, $w_{|c|}$ is the last word of c, $w_{|s_i|}$ is the last word of s_i, θ_c contains the claim encoder parameters, θ_S contains the sentence encoder parameters, and h_c, $h_{s_i} \in \mathbb{R}^{1\times l}$ are l-dimensional vectors.

4.2 Coherence-based Evidence Attention

Our assumption is that sentences used as evidence should be topically coherent given a claim. For example, for the claim in Table 1, which is about the connection between 5G test and birds' death in a park in Hague, the sentences s_1-s_4 are topically coherent by specifically addressing the event's detail while s_5-s_7 are marginal as s_6 and s_7 diverge from the focus and s_5 is a too general statement even though it might imply a possibility.

Our model cross-checks all the sentences to capture the coherence among them using an attention mechanism. We consider the relation from two perspectives: 1) *global coherence* measures the consistency of each sentence regarding the entire set as a whole; and 2) *local coherence* measures the consistency of each sentence considering its relation with another sentence. For each s_i, we use a biaffine attention (Dozat and Manning, 2016), which naturally fits our problem, to get the attention weights:

$$\begin{aligned} \tilde{a}_i &= (H_S \cdot W_c) \cdot h_{s_i}^\top + H_S \cdot u^\top \\ \tilde{\alpha}_i &= \mathrm{softmax}(\tilde{a}_i) \end{aligned} \tag{2}$$

where $H_S = [h_{s_1}; \dots; h_{s_n}] \in \mathbb{R}^{n\times l}$ is the matrix representing all sentences, and $W_c \in \mathbb{R}^{l\times l}$ and $u \in \mathbb{R}^{1\times l}$ contain the weights of the biaffine transformation. The term $H_S \cdot u^\top \in \mathbb{R}^{n\times 1}$ denotes the global coherence where each element is a prior probability of a sentence s_j being coherent with any sentences in S; the term $(H_S \cdot W_c) \cdot h_{s_i}^\top \in \mathbb{R}^{n\times 1}$ is the local coherence where each element $h_{s_j} \cdot W_c \cdot h_{s_i}^\top$ represents the relative likelihood of s_j being coherent with s_i. Therefore, $\tilde{\alpha}_i \in \mathbb{R}^{n\times 1}$ is a n-dimensional weight vector for s_i where each element $\tilde{\alpha}_{ij}$ for $j \in [1, \dots, n]$ denotes the coherence attention weight between s_i and s_j.

Extension of Coherence Attention

The coherence attention in Eq. 2 ignores the claim information. To prevent off-topic coherence which deviates from claim's focus, we propose to assess each sentence's coherence by jointly considering the claim and all sentences, which shares a similar intuition with the co-attention method in question-answering (Lu et al., 2016; Xiong et al., 2016).

Unlike the question-answer co-attention focusing on mutual selection of salient words in question and documents, we focus on sentence-level attention, for which we have multiple sentences but only one claim. So, we only need a claim-guided sentence attention. We use a gating unit to endow the model with the capacity of deciding how much information it should accept from the claim. The new attention weight of s_i is computed by:

$$\begin{aligned} \bar{h}_{s_i} &= g^{c\to s_i} \odot h_{s_i} + (1 - g^{c\to s_i}) \odot h_c \\ \bar{a}_i &= (\bar{H}_S \cdot W_c) \cdot \bar{h}_{s_i}^\top + \bar{H}_S \cdot u^\top \\ \bar{\alpha}_i &= \mathrm{softmax}(\bar{a}_i) \end{aligned} \tag{3}$$

where $g^{c\to s_i} = \sigma(W_g \cdot h_{s_i} + U_g \cdot h_c)$ is the gate function with trainable parameters W_g and U_g, $\bar{H}_S = [\bar{h}_{s_1}; \dots; \bar{h}_{s_n}]$ denotes the stacked output of the gating unit, and other settings are same as the biaffine coherence attention (see Eq. 2).

Based on the attention weights, each sentence can be represented as the weighted sum of all sentences, capturing its overall coherence:

$$h'_{s_i} = \sum_j \alpha_{ij} \cdot h_{s_j} \tag{4}$$

where α_{ij} is the attention weight between s_i and s_j obtained from Eq. 2 ($\tilde{\alpha}_i$) or Eq. 3 ($\bar{\alpha}_i$).

Finally, we concatenate the coherence-based sentence embedding h'_{s_i} with the original embedding h_{s_i} to obtain a richer sentence representation:

$$\tilde{h}_{s_i} = \tanh(W_{co} \cdot [h_{s_i}, h'_{s_i}] + b_{co}) \tag{5}$$

where W_{co} and b_{co} are parameters for transforming the concatenation into a l-dimensional vector.

4.3 Entailment-based Evidence Attention

We further enhance the sentence representation by capturing the entailment relations between the sentences and the claim based on the NLI method (Bowman et al., 2015) for strengthening the semantic inference capacity of our model.

Given c and s_i, we represent each such pair by integrating three matching functions between h_c and $\tilde{h}_{s_i}$: 1) concatenation $[h_c, \tilde{h}_{s_i}]$; 2) element-wise product $h_c \odot \tilde{h}_{s_i}$; and 3) absolute element-wise difference $|h_c - \tilde{h}_{s_i}|$. The similar matching scheme was commonly used to train NLI models (Conneau et al., 2017; Mou et al., 2016; Liu et al., 2016; Chen et al., 2016). We then perform a transformation to obtain the joint representation $h^c_{s_i}$ as follow:

$$h^c_{s_i} = \tanh\left(W_e \cdot \left[h_c, \tilde{h}_{s_i}, h_c \odot \tilde{h}_{s_i}, |h_c - \tilde{h}_{s_i}|\right]\right) \tag{6}$$

where W_e are trainable weights for transforming the long concatenation into an l-dimensional vector. We omit the bias to avoid notational clutter.

To capture entailment-based evidence, we again apply attention over the original sentences guided by the joint representation $h^c_{s_i}$ which is obtained *on top of* the coherence attention. This yields:

$$\begin{aligned} b_i &= \tanh(V_e \cdot h^c_{s_i} + b_e) \\ \beta_i &= \frac{\exp(b_i)}{\sum_i \exp(b_i)} \\ h^c_S &= \sum_i \beta_i \cdot h_{s_i} \end{aligned} \tag{7}$$

where V_e and b_e are parameters turning $h^c_{s_i}$ to an entailment score b_i, β_i is the entailment-based attention weight of s_i which is used to produce the final representation h^c_S of an entire instance.

Note that the hierarchy of our attention structure is conveyed by the query part $h^c_{s_i}$, and we apply the weight β_i on the original representation h_{s_i} rather than h'_{s_i} (Eq. 4) or $\tilde{h}_{s_i}$ (Eq. 5), which is empirically better based on our trials since the latter two may contain more redundant information due to the sum over an entire set when computing h'_{s_i}.

4.4 The Overall Model

The attention vector h^c_S is the high-level representation of the claim with the embedded evidence based on the hierarchical attention method. We use a fully connected output layer to output the probability distribution over the *veracity* classes:

$$\hat{y} = \mathrm{softmax}(V_o \cdot h^c_S + b_o) \tag{8}$$

where V_o and b_o are the weights and bias in output layer. Note that Eq. 8 assumes that using h^c_S alone can determine the *veracity* as true or false without direct reference to the claim again. This may be suitable for *news* data as the salient news sentences often straightforwardly comment on the claim's veracity. However, some claim verification tasks such as FEVER (Thorne et al., 2018a) are particularly defined to classify if the *factual* evidence from the source like Wikipedia, which rarely remark on the veracity of the *mutated* claim, can infer the claim as being supported, refuted or NEI. In such case, we replace h^c_S in Eq. 8 with the richer representation $\hat{h}^c_S = [h_c, h^c_S, h_c \odot h^c_S, |h_c - h^c_S|]$ to facilitate the *inference* from the evidence to the claim in accordance with such NLI style of the task definition. Interestingly, such treatment does not work for *veracity classification* of *news* claim (see Section 5.2), which may be because the veracity features of news claim have been already embedded into h^c_S and the richer representation $\hat{h}^c_S$ involving the claim could introduce unnecessary noise to a non-NLI type of task unlike FEVER.

To fine-tune our model, we also pre-train the coherence- and entailment-related parameters for avoiding the sole reliance on the potentially limited supervision from the task-specific labels.

Pre-training Coherence Model

Without ground truth for learning the coherence model, we use a pair-wise training strategy to optimize a large margin objective. For each claim

c, we randomly choose another "negative" claim c'. Then we construct a tuple (s, X^+, X^-), where $X^+ = (c, S)$ and $X^- = (c', S')$ are tuples consisting of different claims and their relevant article sentences, and $s \in S$ is a sentence selected randomly. Generally, (s, X^+) should exhibit higher topical coherence than (s, X^-) since the former reports the same claim c. We seek for parameters that assign a higher score to (s, X^+) than (s, X^-) by minimizing the following margin-based ranking loss:

$$\mathcal{L}_c = \max\left\{0, 1 + r(s, X^-) - r(s, X^+)\right\} \quad (9)$$

and $r(,)$ is the ranking function turning the coherence-based sentence embedding to a ranking score:

$$r(s, X) = \tanh\left(W_c' \cdot \mathrm{cohAtt}(s, X) + b_c'\right) \quad (10)$$

where $\mathrm{cohAtt}(,)$ is a shorthand of Eq. 4, and W_c' and b_c' are the weights and bias of an added ranking output layer which is not a part of our end-to-end model. The pre-trained model is used to initialize all the parameters needed for computing Eq. 4.

Pre-training Entailment Model

We use the Standford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) to pre-train the parameters of entailment-based attention model. Specifically, we train a model for Recognizing Textual Entailment (RTE) as follow:

$$\bar{y} = \mathrm{softmax}(V_e' \cdot h_{RTE} + b_e') \quad (11)$$

where $\bar{y}$ is the entailment class label, i.e., *entails*, *contradicts*, or *neutral*, h_{RTE} has the same form as Eq. 6 while the input claim-sentence pair is replaced by a pair of premise and hypothesis in the SNLI corpus (each element is encoded by a GRU sentence encoder), and V_e' and b_e' are the weights and bias of the RTE output layer which is not part of our end-to-end model. The pre-trained model is used to initialize the parameters W_e in Eq. 6.

For pre-training, we minimize the square loss between the distributions of the predicted and the ground-truth entailment classes.

Overall Training

After pre-training, all the model parameters are trained end-to-end by minimizing the squared error between the class probability distribution of the prediction and that of the ground truth over the claims. Parameters are updated through back-propagation (Collobert et al., 2011) with AdaGrad (Duchi et al., 2011) for speeding up convergence. The training process ends when the model converges or the maximum epoch number is met. We represent input words using pre-trained GloVe Wikipedia 6B word embeddings (Pennington et al., 2014). We set d to 300 for word vectors and l to 100 for hidden units, and *no parameter depends on n which varies with different claims*.

5 Experiments and Results

5.1 Datasets and Evaluation Metrics

We use three public fact-checking datasets for evaluation: 1) Snopes and 2) PolitiFact, released by Popat et al. (2018), containing 4,341 and 3,568 *news* claims, respectively, along with relevant articles collected from various web sources; 3) FEVER, released by Thorne et al. (2018a), which consists of 185,445 claims accompanied by human-annotated relevant Wikipedia articles and evidence-bearing sentences, and many claims in FEVER are human altered by mutating the original claims from Wikipedia.

Each Snopes claim was labeled as *true* or *false*, while each PolitiFact claim was originally assigned one of six veracity labels: *true*, *mostly true*, *half true*, *mostly false*, *false*, and *pants on fire*. Unlike Popat et al. (2018) converting all the classes into true or false, we merge mostly true, half true and mostly false into *mixed*, and treat false and pants on fire as false. Thus, we have a more practical classification on PolitiFact, i.e., *true*, *false* and *mixed*. We use micro-/macro-averaged F1, class-specific precision, recall and F-measure as evaluation metrics. We hold out 10% of the claims for tuning the hyper parameters, and conduct 5-fold cross-validation on the rest of the claims.

On FEVER dataset, each claim, which is classified as *Supported*, *Refuted* or *NEI*, can be verified with its ground-truth label and a set of human-annotated evidential sentences extracted from its relevant Wikipedia pages. This task is similar as predicting the entailment relation by aggregating the sentences to infer the NLI-style label of the target claim, instead of directly predicting the claim's veracity as true or false. FEVER shared task used label accuracy, F_1 score of evidential sentence selection, and FEVER score as evaluation metrics (Thorne et al., 2018b).

Method	Snopes								PolitiFact				
			True			False					True	False	Mixed
	micF1	macF1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	micF1	macF1	F_1	F_1	F_1
CNN	0.721	0.636	0.477	0.440	0.460	0.802	0.822	0.812	0.453	0.402	0.368	0.566	0.270
LSTM	0.689	0.642	0.441	0.512	0.517	0.834	0.716	0.771	0.463	0.413	0.452	0.561	0.228
SVM	0.704	0.649	0.459	0.584	0.511	0.832	0.747	0.786	0.450	0.421	0.440	0.547	0.277
DeClarE	0.762	0.695	0.559	0.556	0.553	0.839	0.837	0.837	0.475	0.443	0.447	0.576	0.307
HAN-na	0.750	0.674	0.535	0.500	0.517	0.821	0.841	0.831	0.470	0.431	0.456	0.594	0.242
HAN-ba	0.771	0.738	0.556	**0.765**	0.644	**0.899**	0.774	0.832	0.520	0.471	0.475	**0.629**	0.308
HAN	**0.807**	**0.759**	**0.637**	0.665	**0.651**	0.874	**0.860**	**0.867**	**0.523**	**0.487**	**0.495**	0.627	**0.340**
HAN-nli	0.747	0.670	0.534	0.491	0.512	0.817	0.841	0.830	0.485	0.432	0.467	0.599	0.230

Table 2: Results of comparison among different models on Snopes (left) and PolitiFact (right) datasets

Method	Snopes		PolitiFact	
	micF1	macF1	micF1	macF1
HAN-na	0.750	0.674	0.470	0.431
+ ba	0.776	0.727	0.495	0.455
+ ca	0.788	0.741	0.516	0.473
+ ea	0.779	0.728	0.508	0.463
+ ba + ea	0.771	0.738	0.520	0.471
+ ca + ea	0.807	0.759	0.523	0.487

Table 3: Results of ablation test across different attentions on Snopes (left) and PoliFact (right) datasets.

5.2 Experiments on Veracity-based Datasets

We compare our model and several state-of-the-art baseline methods described below. 1) **SVM**: A linear SVM model for fake news detection using a set of linguistic features (e.g., bag-of-words, ngrams, etc.) handcrafted from relevant sentences (Thorne and Vlachos, 2018); 2) **CNN** and **LSTM**: The CNN-based detection model (Wang, 2017) and LSTM-based RNN model for representation learning from word sequences (Rashkin et al., 2017), respectively, both using *only claim content without considering external resources*; 3) **DeClarE**: The word-level neural attention model for Debunking Claims with Interpretable Evidence (Popat et al., 2018) capturing *world-level* evidence from relevant articles; 4) **HAN**: Our full model based on Hierarchical Attention Networks, where coherence component uses Eq. 3; 5) **HAN-ba**: A variant of HAN with biaffine attention in Eq. 2; 6) **HAN-na**: Our reduced model with no attention but only using original sentence representations; 7) **HAN-nli**: A variant of HAN by replacing h_S^c in Eq. 8 with $\hat{h}_S^c$ for the output layer (see Section 4.4).

We implement our models and DeClarE with Theano[3], and use the original codes of other baselines. As DeClarE is not yet open-source, we consult with its developers for our implementation.

Results of Comparison

As shown in Table 2, CNN and LSTM using barely content of claims without considering external information are comparable with SVM which uses handcrafted features based on relevant article sentences. Among all the baselines, DeClarE performs the best because it not only learns to capture complex features effectively via the neural model, but also strengthens the learned features by attending on the salient words that are important for predicting the correct label.

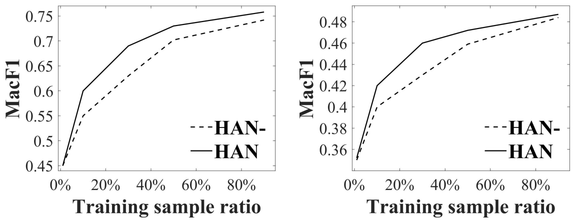
Our model can capture more accurate sentence-level evidence which convey the semantics more completely and deeply. The superiority is clear: HAN-na which considers sentence as evidence without using attention is already better than the baselines except DeClarE, implying the importance of sentence-level information. HAN-ba and HAN using attentions to embed sentence-level evidence consistently outperform DeClarE in large margin that is based on word-level attention.

HAN consistently outperforms HAN-ba on both datasets. This suggests that the co-attention considering claim for capturing sentence coherence is more effective to represent more accurate evidence. HAN-nli, however, fails to work and is even worse than DeClarE, which confirms our conjecture that veracity classification on news data differs from a NLI type of task like on FEVER (see Section 5.3) since news reports often openly remark the claim's veracity and involving the claim in the output layer may interfere the decision.

Ablation Study

To evaluate the impact of each component, we perform ablation tests based on the no-attention model **HAN-na** plus some component(s) which can be one or combination of the following attentions: 1) **ba** and 2) **ca** correspond to the coherence-based biaffine attention (Eq. 2) and co-attention

[3] http://deeplearning.net/software/theano/



(a) Snopes dataset (b) Politifact dataset

Figure 2: Results of **HAN** and **HAN-** (not pre-trained) under different sizes of training data.

(Eq. 3), respectively; 3) **ea**: entailment-based attention (Eq. 7).

As shown in Table 3, HAN-na plus each component alone improves the model, indicating their effectiveness for embedding sentence-level evidence. Furthermore, +ca consistently outperforms +ba, reaffirming the advantage of co-attention; +ea makes similar improvements over HAN-na as +ba and +ca did, suggesting that both types of attention are comparably helpful. Combining them hierarchically makes further improvements especially in the case of +ca+ea, implying that the two attention mechanisms are complementary.

We also examine the impact of pre-training on **HAN** in comparison with its performance without pre-training, namely **HAN-**. In Figure 2, we observe that the pre-training does not have much impact when we use the entire training set, but it clearly improves the model when only using certain proportions of the training data. This indicates that the fine-tuned coherence and entailment models are generally helpful for claim verification, especially when the sampled set is not sufficiently large for fully training the model.

Discussion

Regarding the gap between the published performance of DeClarE (Plain+Attn) (Popat et al., 2018) which is 0.79 on the Snopes dataset and that of our implementation of it which is 0.759, we conjecture the reason may be that DeClarE utilized an undisclosed strategy for balancing the training datasets that we could not easily replicate, while we trained all the systems in Table 2 on the original unbalanced dataset. We leave this for further investigation in future upon the availability of DeClarE source codes. On the PolitiFact dataset, since we adopt a three-way classification, it is thus not directly comparable with the original DeClarE performance which is based on two classes.

Claim: Comedian Bill Murray is running for president
Verdict: ***False***

1 It turns out it's not true and just the subject of a hoax article by a website parodying ABC news.
2 Bill Murry is not running for president, nor has he announced that fact from his hometown.
3 Unknown Internet prankster created fake website for NBC, ABC and Fox News running the headline "Bill Murray is running for president".
...
8 Murray made the announcement from his home in and he felt the 2016 presidential election seemed like the right time to go.
9 Paul Horner, a spokesman for the campaign, told reporters that he believes in Bill Murry for President.

Table 4: Examples of attended sentences ranked by the attention weight β_i that can explain the verdict.

Method	Acc.	Prec.	Rec.	F_1	FEVER
Fever-base	0.521	–	–	–	0.326
NSMN	0.697	0.286	**0.870**	0.431	**0.665**
HAN-nli	0.642	0.340	0.484	0.400	0.464
HAN-nli*	**0.720**	**0.447**	0.536	**0.488**	0.571
HAN*	0.475	0.356	0.471	0.406	0.365

Table 5: Results of different claim verification models on FEVER dataset (Dev set). The columns correspond to the predicted *label* accuracy, the *evidence* precision, recall, F1 score, and the FEVER score.

Case Study

Table 4 illustrates some top sentences embedded with a claim from Snopes dataset which is correctly detected as fake. We can see that 1) the top sentences have high topical overlap with both the claim and each other; 2) the highly ranked sentences play a major role in deciding the verdict, as they remark on the claim's veracity directly; 3) the lower sentences seem less important since they either repeat the claim or are very subjective. Providing such readable pieces of evidence to human fact-checker for verifying the claim can be helpful.

5.3 Experiments on FEVER Dataset

We compare the following systems on the public Dev set[4] of FEVER dataset: 1) **Fever-base**: The FEVER baseline (Thorne et al., 2018a) that is a pipeline for claim verification including 3 stages: document retrieval, sentence selection and textual entailment. 2) **NSMN**: The pipeline-based system named as UNC-NLP topping the FEVER shared task (Thorne et al., 2018b), which was later reported as using Neural Semantic Matching Networks (Nie et al., 2019). 3) **HAN-nli**: Our full

[4] The test set is not publicly available at the time of this work being done.

model trained using the FEVER task dataset. Note that similar to DeClarE our model assumes that the set of articles about each claim have been retrieved, while the FEVER task requires users search relevant Wikipages in the first place. Using FEVER, our method thus is not truly end-to-end in this setting. We utilize the document retrieval module of NSMN (Nie et al., 2019) to obtain the relevant Wikipages. 4) **HAN-nli***: For more fair comparison with NSMN which utilized the ground-truth sentences in the training set to train their sentence selector, we fine-tune the HAN-nli, namely HAN-nli*, by optimizing the square error loss between the entailment attention score b_i (see Eq. 7) and the -1/+1 value indicating whether s_i is selected as a piece of evidence in the ground truth. 5) **HAN***: The original HAN using Eq. 8 in the output layer and fine-tuned like HAN-nli*.

Table 5 shows that HAN-nli* is much better than the two baselines in terms of label accuracy and evidence F1 score. There are two reasons: 1) apart from the retrieval module, our model optimizes all the parameters end-to-end, while the two pipeline systems may result in error propagation; and 2) our evidence embedding method considers more complex facets such as topical coherence and semantic entailment, while NSMN just focuses on similarity matching between the claim and each sentence. HAN-nli seem already a decent model given its much better performance than Fever-base. This confirms the advantage of our evidence embedding method on the FEVER task.

NSMN achieves higher FEVER score and evidence recall than our method. However, the reason is straightforward: FEVER score favors recalling the annotated evidential sentences while one of the limitations of FEVER dataset is that the ground-truth sentences provided by human annotators were often incomplete (Thorne et al., 2018a,b). Our approach is not limited by selecting top-k sentences and may embed into evidence as many diverse sentences as the model requires. Compared to NSMN which aims to recall the top evidence sentences in FEVER's ground truth, our model achieves much higher Accuracy, Evidence Precision and F_1.

HAN* is ineffective, confirming that in FEVER task the claim content is needed in the output layer for the NLI to take effect since the evidence from Wikipedia typically does not contain direct remarks on the veracity of a claim.

Discussion

The pipeline-based system NSMN demonstrates superior evidence retrieval performance in terms of FEVER score. We emphasize that the essential objective of our model is not for evidence retrieval and ranking. Instead of ranking sentences into the top-k positions, we pay more attention on claim verification accuracy by embedding and aggregating the useful sentences as evidence like we have explained above. However, such discrepancy inspires us to investigate in the future an end-to-end approach to jointly model evidence retrieval and claim verification in a unified framework based on our sentence-level attention mechanism.

Finally, thanks to one of our reviewers, we learn about another two-stage model named **TwoWingOS** (Yin and Roth, 2018), which achieves a comparable FEVER score but a little bit higher accuracy than ours on FEVER task. The TwoWingOS applies a two-wing optimization approach to jointly optimizing sentence selection and veracity classification. The reasons regarding their higher performance might lie in that: 1) their input word embeddings are fine-tuned based on the context of the evidence and claim while ours are fixed during training; and 2) the document retrieval module of the TwoWingOS has demonstrated higher effectiveness than that of the NSMN (see rate (recall) and acc_ceiling (OFEVER) in Tables 2 in (Yin and Roth, 2018; Nie et al., 2019) for details).

6 Conclusions and Future Work

We propose a novel neural end-to-end framework for claim verification by learning to embed sentence-level evidence with a hierarchical attention mechanism. Our model strengthens the evidence representations by attending on the sentences that are not only topically coherent but can also semantically infer the target claim. The results on three public benchmark datasets confirm the advantages of our method. For the future work, beyond what we have mentioned, we plan to examine our model on different information sources. We will also try to incorporate relevant metadata into it, e.g., author profile, website credibility, etc.

Acknowledgment

This work was partly supported by Hong Kong RGC GRF (14232816, 14209416, 14204118), NSFC (61877020) and SCSE-SUG grant M4082038 at NTU.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90.

Samuel R 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.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR*, abs/1609.06038.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches. *Journal of Natural Language Engineering*, 4.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108.

Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2972–2978. AAAI Press.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3818–3824. AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1980–1989.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 130.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

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

Ankur Padia, Francis Ferraro, and Tim Finin. 2018. Team umbc-fever: Claim verification using semantic lexical resources. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 161–165.

Cesc C Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *Advances in neural information processing systems*, pages 73–81.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012. International World Wide Web Conferences Steering Committee.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics.

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.

Edward S Reed, Elliot Turiel, and Terrance Brown. 2013. Naive realism in everyday life: Implications for social conflict and misunderstanding. In *Values and Knowledge*, pages 113–146. Psychology Press.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *arXiv preprint arXiv:1901.06437*.

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.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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)*, volume 1, pages 809–819.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and verification (fever) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9. Association for Computational Linguistics.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 647–653.

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)*, volume 2, pages 422–426.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *CoRR*, abs/1611.01604.

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.

James O. Young. 2018. The coherence theory of truth. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):32.