Evidence-based Trustworthiness

Yi Zhang **Zachary G. Ives** **Dan Roth**
Department of Computer and Information Science
University of Pennsylvania
{yizhang5, zives, danroth}@cis.upenn.edu

Abstract

The information revolution brought with it *information pollution*. Information retrieval and extraction help us cope with abundant information from diverse sources. But some sources are of anonymous authorship, and some are of uncertain accuracy, so how can we determine what we should actually believe? Not all information sources are equally trustworthy, and simply accepting the majority view is often wrong.

This paper develops a general framework for estimating the trustworthiness of information sources in an environment where multiple sources provide claims and supporting evidence, and each claim can potentially be produced by multiple sources. We consider two settings: one in which information sources directly assert claims, and a more realistic and challenging one, in which claims are inferred from evidence provided by sources, via (possibly noisy) NLP techniques. Our key contribution is to develop a family of probabilistic models that jointly estimate the trustworthiness of sources, and the credibility of claims they assert. This is done while accounting for the (possibly noisy) NLP needed to infer claims from evidence supplied by sources. We evaluate our framework on several datasets, showing strong results and significant improvement over baselines.

1 Introduction

The emergence of social networks and news aggregators — combined with ill-informed posts, deliberate efforts to create and spread sensationalized information, and a strongly polarized political environment — makes it very difficult to establish what is really known. Therefore, *fact checking* seeks to assess whether the claim is true or false, or to provide a confidence level for the claim given textual evidence (Hassan et al., 2017; Wang, 2017; Wang et al., 2018). A typical fact checking pipeline consists of document retrieval, sentence-level evidence selection, and textual entailment stages (Thorne et al., 2018). However, this pipeline is *local* in that it applies to a given claim. The missing step here is to assess the *trustworthiness of the sources* producing the claims and evidence. This is a global step that, in principle, accounts for all claims made by a source and all sources making a claim.

Previous work has studied how to estimate the trustworthiness or credibility of information sources for *fact-finding* (Vydiswaran et al.; Pasternack and Roth, 2013), *truth discovery* (Dong et al.; Pochampally et al., 2014; Dong et al., 2015; Li et al., 2016) and *crowdsourcing* (Sabou et al., 2012; Hovy et al., 2013; Gao et al., 2015). Usually, given a list of conflicting facts, e.g. "source s asserts claim c", or "annotator x labels data item t by label y", we detect the true claims or correct labels for the data item by resolving conflicts, and then compute the trustworthiness of sources.

However, many sources do not directly assert claims, but rather generate articles as *evidence*, expecting readers to infer claims from this evidence. In practice, given a claim of interest, people may search for related articles from multiple sources and collect evidence for the claim; they can then determine the veracity of the claim by deciding whether the evidence found supports or refutes the claim. However, most existing work that attempted to study trustworthiness of sources assumed that sources make assertions directly. Even when intermediate text was accounted for (Vydiswaran et al.; Nakashole and Mitchell, 2014), it was assumed that clean evidence and clear connections between evidence and conflicting claims are provided, disregarding the fact that NLP systems attempting to support these tasks are noisy.

This paper considers two situations when eval-

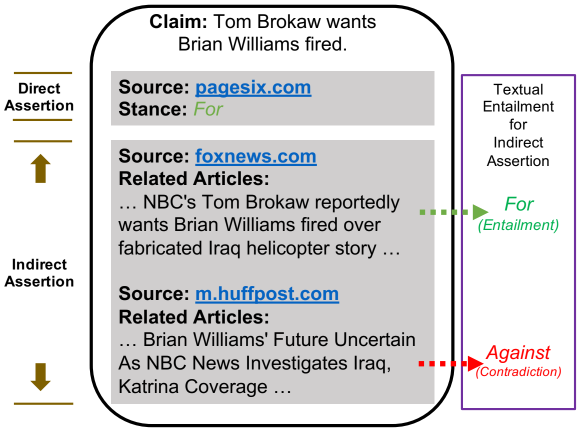

Claim: Tom Brokaw wants Brian Williams fired.
Direct Assertion
Source: pagesix.com
Stance: For
Indirect Assertion
Source: foxnews.com
Related Articles:
… NBC's Tom Brokaw reportedly wants Brian Williams fired over fabricated Iraq helicopter story …
Source: m.huffpost.com
Related Articles:
… Brian Williams' Future Uncertain As NBC News Investigates Iraq, Katrina Coverage …
Textual Entailment for Indirect Assertion
For
(Entailment)
Against
(Contradiction)

Figure 1: Claim with assertions from multiple sources (from `http://www.emergent.info/`). Direct assertions specify their stance; indirect assertions provide related articles, and we can leverage (noisy) text entailment tools to collect their stances. We want to assess whether to believe the stance and articles.

uating the trustworthiness of information sources: (1) the source directly asserts claims, and (2) the source indirectly asserts claims by proposing evidence. The first case is similar to previous work; the second case is more challenging but more important in practice. Both cases are depicted in Figure 1. A multitude of sources is given and each may assert multiple claims or propose multiple pieces of evidence. At the same time, multiple claims are observed, some of which are directly asserted by sources and some are supported by evidence.

Our goals are to identify true claims and to estimate the trustworthiness of each source. The key challenge is that this global inference task is influenced by the knowledge of which claims are made by which sources; however, establishing links – from evidence generated by a source to claims – requires NLP techniques such as textual entailment (TE) (Dagan et al., 2013). Such TE tools, which assess whether a given textual evidence (premise) entails a given claim (hypothesis), are often noisy — making the evaluation of sources more difficult.

The key contributions of this work are as follows: (1) It proposes a probabilistic model, **JELTA**, which jointly estimates the credibility of claims and the trustworthiness of sources, when claims are made by sources directly, indirectly, or both. (2) Our framework incorporates a TE model as part of the global inference framework as a way to link evidence (and thus, sources) to claims. (3) This is the first work to distinguish between direct and indirect assertions made by information sources.

Our experiments on both synthetic and natural datasets show solid results that are significantly better than baselines.

2 Trustworthiness Analysis

Our goal is to evaluate the trustworthiness of information sources by detecting the true claims while accounting for noise in the links between claims and evidence for them. While direct assertions are straightforward to deal with (since it is clear which source generates which claim), the challenge is to incorporate "noisy assertions" into our problem formulation. We first describe our setting, and then elaborate on the probabilistic modeling.

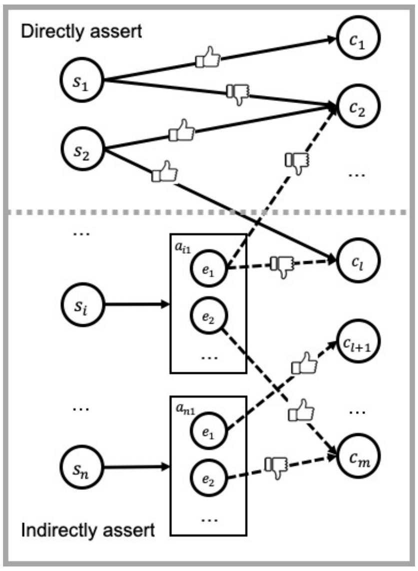


Figure 2: Our solution considers two settings: (1) source s_i directly asserts multiple claims c_j; (2) the source provides evidence e_k by multiple articles, and the proposed evidence can support or refute claims via some noisy NLP tool.

2.1 Noisy Assertions

We are given a set of claims to validate and a text corpus (pieces of evidence) generated by multiple sources that are believed to have generated the claims. Given the claim text, we issue a set of searches over the corpus, to find evidence in support of the claims. The result is a a set of (noisy) assertions. A (noisy) assertion consists of a claim, a sentence in the corpus, and a label ("entailment", "contradiction", "neutral"). The claim is a real world input we attempt to determine the truth value of. E.g., in Figure 1, "`Tom Brokaw` `wants Brian Williams fired`" is such a claim. An assertion, on the other hand, is an artifact of our framework. As we search the corpus generated by the sources for evidence supporting the claim, we identify candidate sentences ('Related Articles" in the figure) and use a pretrained textual entailment model (e.g. the decomposable attention model (Parikh et al., 2016)) to provide an entailment label and complete the triple (claim, sentence, label). The generation of noisy assertions as described above follows a typical fact-checking pipeline mentioned in Thorne et al. (2018).

Table 1: Notation Table

Notation	Description
s	an information source
c	a claim
m	a mutually exclusion set of claims
e	an evidence
y_m	the true claim in m
$b_{s,c}$	The (observed) probability of c asserted by s
$b_{s,e,c}$	The (observed) probability of c asserted by e from s
$w_{s,m}$	if s asserts claims of m
$w_{s,e}$	if s provides e
$w_{e,m}$	if e supports or refutes claims of m
H_s	the probability s makes an honest claim
P_s	the (hidden) probability s produces a true evidence
R_s	the probability s recalls a true evidence (true-positive rate)
Q_s	the probability s recalls a false evidence (false-positive rate)
X_i	Set of all (observed) direct assertions
X_d	Set of all (observed) indirect assertions
Y	Set of all true claims

Given noisy assertions, Figure 2 illustrates our problem setting. Overall, there are two situations. In the upper part of the figure, we show the case in which information sources make *direct assertions*: the source directly states that some claims are true or false. The alternative case, indicated in the lower part of the figure, involves the source *indirectly asserting claims* by making noisy assertions: the source first generates articles that contain sentences, and the sentences may entail or refute related claims. An entailment tool can then be used to assert the claims to be true or false, based on those sentences. A claim can be supported by multiple sources or multiple pieces of evidence from different sources. We now propose our model, JELTA, which handles both cases described above.

2.2 Fundamentals

Our probabilistic model denotes an information source as $s \in S$, a claim as $c \in C$, and m as a *mutual exclusion set* of claims (exactly one of the claims in each mutual exclusion set is true). Here m is a fact to be checked, and c is a statement that m is true or false. $w_{s,m}$, $w_{s,e}$ and $w_{e,m}$ are binary indicators — respectively telling us if s asserts claims of m, if s provides evidence e, and if evidence e supports claims in m. We denote evidence $e \in E$, and for each entailment result, we use $b_{s,c}$ and $b_{s,e,c}$ to represent the observed probability that s asserts c and s provides e to assert c respectively. Here, $\sum_{c \in m} b_{s,c} = 1$ and $\sum_{c \in m} b_{s,e,c} = 1$. We summarize our notation in Table 1.

2.3 JELTA

Our work models a joint distribution that reflects a "story" of how sources generate observations. Intuitively, given an estimation of the verdict of the claims and the factors, including the trustworthiness of sources providing claims and evidence, we want to maximize the probability that we can observe the claims and evidence.

We represent the verdict of a claim m as a latent variable, y_m, and associate a parameter H_s with each source s, reflecting the probability of s telling the truth, which we use to measure the trustworthiness of s. We now describe how y_m and H_s are used to compute the probability of observing the claims and evidence. Starting with the probability that source s makes a direct assertion: intuitively, if s asserts a true claim $c = \hat{c}$ in m, then the probability that s asserts c is H_s, the probability s telling the truth. Otherwise, s chooses uniformly from other claims in m with probability $\frac{1-H_s}{|m|-1}$.

Besides the term H_s, we require another (hidden) factor related to s, namely, the probability of s telling the truth as evidence. We denote this as P_s, the *precision* of s generating evidence. Here we allow P_s can be different with H_s, since providing true evidence for a true claim is more difficult than just providing a true claim. However, considering that those all reflect the trustworthiness of s, we assume they share a similar distribution over sources in our problem.

P_s can then be represented by two other parameters, R_s and Q_s (Dong et al., 2015). These represent the true- and false-positive rates of s producing evidence, respectively. We denote γ as the

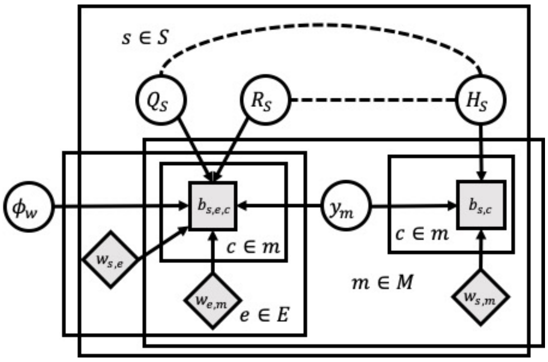


Figure 3: Plate diagram for probabilistic model describing the generation of direct and indirect assertions. Shaded parts are the observations, y_m is the latent variable, and H_s, R_s, Q_s and ϕ_w are groups of parameters. Dotted lines describe the interactions between H_s and R_s, Q_s.

probability of a claim being true, then P_s can be represented by Q_s and R_s as:

$$P_s = \frac{\gamma R_s}{\gamma R_s + (1-\gamma) Q_s} \tag{1}$$

We assume that the probability of the claim being true or false is equal. Since H_s and P_s share similar distributions, H_s relates to Q_s and R_s as follows:

$$H_s \sim P_s = \frac{R_s}{R_s + Q_s} \tag{2}$$

Now we discuss how to compute the probability of observing the noisy assertions. Intuitively, when source s wants to assert a claim with the NLP tool (textual entailment model) by proposing evidence: if s wants to support a true claim indirectly, s will recall true evidence with probability R_s. This requires the NLP tool to do textual entailment correctly, otherwise s will also uniformly choose false or unrelated evidence.

This paper considers the simplest way to generate a false claim or false evidence, and the choice may not always follow random sampling in practice. Our prior work Pasternack and Roth (2013) discusses some other options, which could alternatively be used here.

In the remainder of this section, we formally model those processes.

Direct Assertion. Modeling the generative process of direct assertions by sources is very similar to Simple LCA (Pasternack and Roth, 2013). As above, if the claim $c \in m$ asserted by s is the true claim y_m, the probability of observing the source s asserting claim c of m is H_s. Otherwise, the probability of s asserting a false claim of m is $\frac{1-H_s}{|m|-1}$.

Therefore, the joint probability of the observation over X_d and y_m can be modeled as follows:

$$P(y_m,\ X_d|H_s) = P(y_m)\Big(H_s^{b_{s,y_m}} \prod_{c \in m \setminus y_m} (\frac{1-H_s}{|m|-1})^{b_{s,c}}\Big)^{w_{s,m}} \tag{3}$$

Then, given all sources S and $\theta = \{H_s\}$, we can write the full joint of direct observations as:

$$P(Y, X_d|\theta) = \prod_m P(y_m) \prod_s \Big(H_s^{b_{s,y_m}} (\frac{1-H_s}{|m|-1})^{1-b_{s,y_m}}\Big)^{w_{s,m}} \tag{4}$$

Note that we simplify the expression by leveraging $\sum_{c \in m \setminus y_m} b_{s,c} = 1 - b_{s,y_m}$.

Indirect Assertion. Here the sources provide articles containing possible evidence rather than making direct assertions. Besides the parameters Q_s and R_s, the observation also depends on the noisy entailment results given by the textual entailment model. Therefore, we introduce a function $\phi_w(e, m, c) \in \mathbb{R}^1$ to measure the reliability of an entailment result. Here $\phi_w(e, m, c)$ is a linear combination of feature values in a sigmoid function, so that we can scale it to $[0, 1]$:

$$\phi_w(e, m, c) = \frac{\exp(\sum_i w_i z_i)}{1 + \exp(\sum_i w_i z_i)} \tag{5}$$

where z_i is a feature for each given $\langle e, m, c \rangle$, and $w = \{w_i\}$ are the weights of each z_i learned by our model.

For each observation $\langle s, e, m, c \rangle$, the source generates true evidence with probability R_s, and with probability of $\phi_w(e, m, c)$, the proposed evidence e supports claim c of m. This means that we have probability of $R_s \cdot \phi_w(e, m, c)$ to observe the tuple when $c = y_m$. If $c \neq y_m$, either the source does not provide true evidence, or the entailment model provides an unreliable entailment result — which means we have probability of $\frac{1}{N}\big(1 - P_s \cdot \phi_w(e, m, c)\big)$ to observe a false evidence-claim pair. Here N is the total number of such false evidence-claim pairs.

Therefore, the joint probability of the observation over y_m and X_i (indirect assertion observations) is as follows:

$$
\begin{aligned}
&P(y_m, X_i|R_s, Q_s, W) = \\
&\quad P(y_m)\prod_e \Bigg(\Big(R_s \cdot \phi_w(e,m,c)\Big)^{b_{s,e,y_m}} \\
&\quad \Big(\frac{1-\frac{R_s}{R_s+Q_s}\cdot\phi_w(e,m,c)}{N}\Big)^{1-b_{s,e,y_m}}\Bigg)^{w_{s,e},w_{e,m}}
\end{aligned} \tag{6}
$$

Here, we also use $\sum_{c\in m\backslash y_m} b_{s,e,c} = 1 - b_{s,e,y_m}$. Then, given all sources S and $\theta = \{Q_s, R_s, W\}$, the full joint probability of indirect assertions is:

$$
\begin{aligned}
&P(Y,X_i|\theta) = \\
&\quad \prod_m P(y_m)\prod_s\prod_e \Bigg(\Big(R_s \cdot \phi_w(e,m,c)\Big)^{b_{s,e,y_m}} \\
&\quad \Big(\frac{1-\frac{R_s}{R_s+Q_s}\cdot\phi_w(e,m,c)}{N}\Big)^{1-b_{s,e,y_m}}\Bigg)^{w_{s,e}w_{e,m}}
\end{aligned} \tag{7}
$$

Joint Modeling. Now to consider direct and indirect assertions together, we multiply Equations 4 and 7 together with two hyper-parameters, η_d and η_i, which give different weights to direct and indirect assertions. If $\eta_d > \eta_i$, this means we believe that a direct assertion is more accurate than an indirect assertion, and vice versa.

Therefore, observing that all sources propose their evidence and make their assertions independently, and taking $\theta = \{H_s, R_s, Q_s, W\}$, we can write the full joint as:

$$
\begin{aligned}
&P(X,Y|\theta) = \prod_m P(y_m)\prod_s \Bigg(H_s^{b_{s,y_m}} \\
&(\frac{1-H_s}{|m|-1})^{1-b_{s,y_m}}\Bigg)^{w_{s,m}\eta_d} \prod_e \Bigg(\Big(R_s\cdot\phi\Big)^{b_{s,e,y_m}} \\
&\Big(\frac{1-\frac{R_s}{R_s+Q_s}\cdot\phi}{N}\Big)^{1-b_{s,e,y_m}}\Bigg)^{w_{s,e}w_{e,m}\eta_i}
\end{aligned} \tag{8}
$$

where $\phi = \phi_w(e,m,c)$ for abbreviation. Meanwhile, since $H_s \sim \frac{R_s}{R_s+Q_s}$, we model it by minimizing their KL divergence. Therefore, we also minimize:

$$
\begin{aligned}
\mathbb{E}_{H_s}[\log\frac{H_s}{P_s}] &= \sum_s H_s \log\frac{H_s}{P_s} \\
&= \sum_s H_s \log\frac{H_s(R_s+Q_s)}{R_s}
\end{aligned} \tag{9}
$$

2.4 Inference

The true claim, y_m, is a latent variable that is unknown in our problem, so we solve this approximately by using the EM algorithm (Dempster et al., 1977) to first estimate the true claim, then find the maximum a posterior point estimate of the parameters. Therefore, the E-step is $\forall m$:

$$
P(y_m = c|X,\theta^t) = \frac{P(y_m = c|X,\theta^t)}{\sum_{v\in m} P(y_m = v|X,\theta^t)} \tag{10}
$$

In the M-step, besides maximizing the posterior of parameters, we should also consider the interactions between H_s and R_s, Q_s. We include it as a regularization term with a parameter λ that controls the importance of the interactions. Thus, the M-step is as follows:

$$
\begin{aligned}
\theta^{t+1} = \text{argmax}_\theta \mathbb{E}_{Y|X,\theta^t}&[\log P(X,Y|\theta)P(\theta)] \\
&- \lambda\mathbb{E}_{H_s}[\log\frac{H_s}{P_s}]
\end{aligned} \tag{11}
$$

Since there are no closed form solutions for those parameters, we use gradient ascent to solve them parameter-by-parameter. We leave the computation of derivatives to the appendix.

2.5 Measuring Entailment Results

In our model, $\phi_w(e,m,c)$ evaluates the reliability of an entailment result given by the entailment model. As we described in Section 2.3, $\phi_w(e,m,c)$ is a sigmoid function of a linear combination of feature values, and we include following features:

Entailment Score. For each prediction of the given entailment model, the model will predict a label, i.e. *entailment*, *contradiction* or *neutral* as well as a score to support its conclusion.

Text Similarity. This feature is computed by the cosine similarity between numerical representations of the evidence and the claim. In this work, we use tf-idf and Glove (Pennington et al., 2014) to represent sentences respectively. To represent a sentence, we use the pre-trained Glove [1] with a simple method proposed in (Arora et al., 2017).

Entity Similarity. We identify entities for each pair of evidence and claim, and compute the overlap of entities by jaccard similarity and entity similarity by NESim (Do et al., 2009) as two features.

3 Experimental Evaluation

We evaluate the effectiveness of our joint model JELTA and compare it with baselines. We first de-

[1] https://nlp.stanford.edu/projects/glove/

scribe our datasets and the methods we compare with, then elaborate on the results.

3.1 Experimental Settings

Data Sets We use both synthetic and natural datasets to evaluate our models.

Synthetic Dataset: FEVER. We use the training file of FEVER[2] to create the synthetic dataset. FEVER is a dataset for verification of claims. We augment FEVER with sources and other information using following steps.

Step 1: Assign Veracity for Claims. Fever provides evidence-claim pairs with their textual entailment labels. Considering our running example, Fever provides sentence pairs such as *"...NBC's Tom Brokaw reportedly..."* and *"Tom Brokaw wants Brian Williams fired."* as evidence and claim. For each experimental round, we sample 200 claims from those pairs, then randomly assign half as true and half as false.These labels will be the ground truth of claims' veracity.

Step 2: Create Sources with Accuracy. Next, we create sources with corresponding accuracy as the ground truth of trustworthiness. In our each experimental round, we create 200 sources and for each source s_i, we associate an accuracy denoted as H_{s_i}. To generate H_{s_i}, we sample a decimal number from a normal distribution ($\mu = 0.5, \sigma = 1$) in $[0, 1]$. A normal distribution is used here because we assume most sources mix true and false claims, and a few of them are highly trustworthy or totally unbelievable.

Step 3: Associate Sources with Evidence and Claims. The last step is to assign claims and evidence to each source. In our experiments, each source makes 30 assertions. Each source s_i, with probability H_{s_i}, picks a true claim; otherwise it picks a false claim. For evidence, since we assume that the distribution of precision generating evidence over sources shares a similar distribution with $\{H_{s_i}\}$, the source s_i picks a piece of evidence either supporting a true claim or refuting a false claim with $H_{s_i} + \epsilon$, where epsilon is a small Gaussian noise ($\mu = 0, \sigma = 1$). Considering the running example, if claim *"Tom Brokaw wants Brian Williams fired."* is associated with *"True"* in Step 1, and Fever provides the pair with label *"Entailment"*, *"...NBC's Tom Brokaw reportedly..."* is therefore a piece of evidence supporting a true claim. Otherwise s_i picks a piece of evidence supporting a false claim or refuting a true claim. Note we assume that each source provides more pieces of evidence than claims, and set the ratio of direct assertions to indirect assertions as $\frac{1}{4}$ in our expeirments.

We run 10 rounds of experiments and report the average performance.

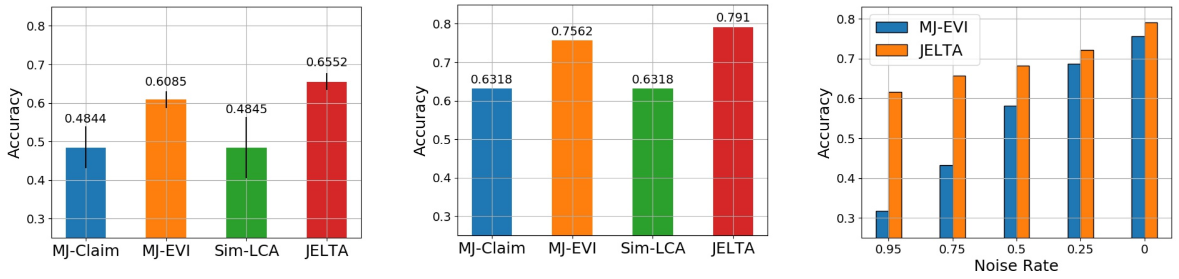
Natural Dataset: Emergent. We use Emergent (Ferreira and Vlachos, 2016) directly; it is derived from a digital journalism project for rumor debunking. It contains 300 rumored claims and 2,595 associated news articles from different websites, collected and labeled by journalists with an estimate of their veracity (true, false or unverified). We eliminated the claims that are unverified, leaving 201 claims and 589 effective sources. For each source, the dataset provides the claims it supports or refutes, which we use as direct assertions. It also provides the articles generated by the source, and we use them as possible evidence repository that may support or refute the claims. The ground truth of the trustworthiness is generated by computing the accuracy of sources based on the veracity label provided.

Entailment Model. We need a textual entailment model to tell us if evidence (a sentence) supports or refutes a claim. We use a pre-trained decomposable attention model (Parikh et al., 2016) with Elmo embedding (Peters et al., 2018) trained on the SNLI dataset [3]. The model's performance is not good on either FEVER or Emergent: when we use majority voting over the evidence to estimate the veracity of a claim, the accuracy is under 40%. To improve the textual entailment model, we adapt the pre-trained model with additional training data. For the experiment on *FEVER*, we randomly sample 100 training examples from labeled development dataset of FEVER. (There is no overlap between the additional training data and our created test data.) For the experiments on *Emergent*, we construct additional training data by article headlines and article headline stance provided by Emergent. Here, the article headline is generated by each article, and the dataset tells us if the article headline can support or refute the claim, which is a good source of additional training data.

Metrics To evaluate the performance of our method as well as the baselines, we evaluate

[2] http://fever.ai/data.html

[3] https://nlp.stanford.edu/projects/snli/

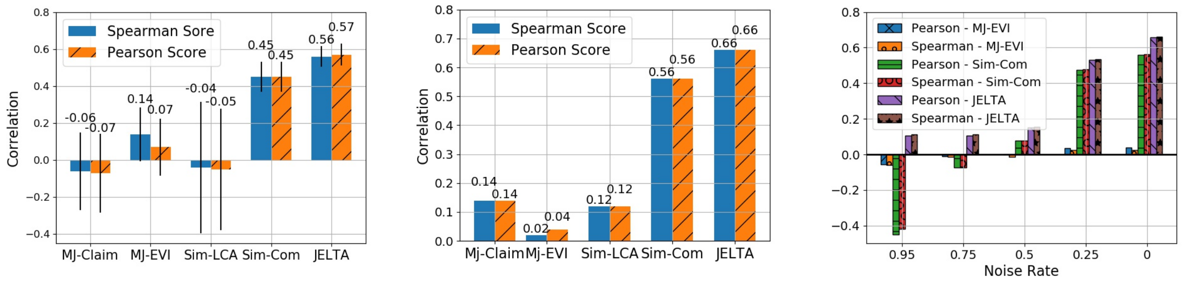


(a) Veracity accuracy of claims on FEVER

(b) Veracity accuracy of claims on Emergent

(c) Veracity varying noise rate in textual entailment results

Figure 4: The performance of estimating veracity of claims by different methods. On both FEVER and Emergent, JELTA achieves the highest accuracy, and a low standard deviation in the 10 rounds evaluation on FEVER. (c) reports the accuracy variation when we add different rate of noise in the textual entailment results, and JELTA is consistently better.



(a) Evaluating correlation estimation on FEVER

(b) Evaluating correlation estimation on Emergent

(c) Correlation varying noise rate in textual entailment results

Figure 5: The performance of estimating the trustworthiness of sources by Peasrson and Spearman score. The methods considering both direct and indirect assertions are much better than those considering one of them only, and JELTA can achieve an additional improvement. (c) reports the performance varying different rate of noises being added to the entailment results, and JELTA is also consistently better than other methods.

(1) the accuracy of the estimated veracity of the claims, (2) the accuracy of the estimated trustworthiness of the source. Here, we evaluate trustworthiness by two typical correlation scores, the Spearman correlation coefficient and Pearson correlation coefficient (Fieller et al., 1957). Spearman's correlation assesses monotonic relationships, whereas Pearson's correlation is the covariance of two random variables — thus when computing Pearson's correlation, we normalize the estimated accuracy of the sources.

Baselines

MJ-Claim. In this case, we only consider direct assertions made by sources, and for each claim we collect all related assertions and do majority vote to estimate the veracity of the claims. Once we get an estimation of their veracity, we can compute the accuracy for each source.

MJ-EVI. We only consider indirect assertions in this case. With the textual entailment model output, each evidence provided by the article will either support, refute or abstain the claim. Here, we also use majority vote to estimate the veracity of the claims, and use the mean ratio between the number of evidence supporting the true claim and the total number of evidence for each claim to estimate the trustworthiness of the source.

Sim-LCA. We leverage the model proposed in (Pasternack and Roth, 2013) to estimate the credibility of the sources. Here, the model only considers direct assertions.

Sim-Com. We propose a simple solution that considers both direct and indirect assertions. Here, we use MJ-EVI to estimate the truth of the claims, based on which we calculate the accuracy for each source. Note that the results are the same compared with MJ-EVI when estimating the veracity of the claims, while the estimation of trustworthiness over sources is different.

3.2 Results

Accuracy of Veracity. Figure 4 (a) reports (for each method) the accuracy of estimation for the claims' veracity, over our synthetic dataset. JELTA achieves the highest accuracy, by around 5%, and shows a low standard deviation over the 10 rounds of experiments. Figure 4 (b) reports the accuracy on Emergent. We again observe a 4% improvement in accuracy compared with MJ-EVI, and around 16% improvement vs the methods considering direct assertions only. It makes sense that evidence-based method (leveraging indirect assertions) can beat the claim-based method (leveraging direct assertions only) by using more information to reduce potential noise. However, using sources and claims only is more noisy, especially with many bad information sources. Figure 4 (a) shows that when the distribution of sources changes, the performance of MJ-Claim and Sim-LCA also varies a lot: their performance greatly depends on the distribution of trustworthiness over sources. Besides offering higher accuracy, evidence-based methods are more robust to varying sources' trustworthiness.

Trustworthiness Estimation. Figure 5 (a) reports the performance by Pearson and Spearman score, for each method's estimate of the trustworthiness of each source on FEVER. JELTA's accuracy is consistently better than other baselines, whenever we use the Spearman or Pearson score to compute the correlation between the estimation and the ground truth. JELTA also has a lower standard deviation over different rounds. This result is consistent with the results shown when we are estimating the veracity of claims. It reveals that evidence-based methods are relatively more stable than the methods considering direct assertions only. MJ-Claim and Sim-LCA highly depend on the trustworthiness distribution over sources. If most of the sources are more trustworthy, we can both estimate the true claims more accurately and better estimate the trustworthiness of sources; and vice versa. That is why both MJ-Claim and Sim-LCA have high standard deviations over different rounds. Based on the results of MJ-EVI, we can observe that simply calculating accuracy by estimated "correct" evidence cannot achieve a highly correlated estimation of sources: the entailment tool provides noisy evidence. However, Sim-Com, which directly counts estimated "correct" claims by MJ-EVI, can improve the estimation. Thus, if we can estimate the veracity of claims accurately, estimating the trustworthiness by claims is more accurate than doing that by noisy evidence. This is also why we can significantly improve the performance by joint modeling. Intuitively, we use evidence to better estimate the veracity of claims, and leverage claims to better estimate the trustworthiness of sources, in an iterative fashion. Figure 5 (b) leads to similar conclusions. Since there are more trustworthy sources, the performance of claim-based methods is better than MJ-EVI.

Influence of textual entailment model. Figures 4 and 5 show that our method, which jointly considers direct and indirect assertions, significantly improves the estimation. Among different factors, evidence contributes the most when estimating the veracity of claims, which can also help the estimation of the trustworthiness. However, the usefulness of evidence highly depends on the quality of the NLP tool. To quantify the amount of noise introduced, we report the Pearson and Spearman score varying a noise rate r. Given r, for each entailment result, with probability r, we will flip the answer of the textual entailment. For example, if the result is "entailment", we will change it randomly to either "contradiction" or "neutral", and vice versa. The results are shown in (c) of Figure 4 and 5. As noise increases, the accuracy, Pearson and Spearman score drop lower. However, the JELTA method is consistently better than the alternatives. JELTA's accuracy decreases more slowly, and its correlation remains positive, even though we flip 95% of the entailment results. This demonstrates that jointly considering direct and indirect assertions can better avoid the skewness caused by either evidence or claims.

4 Related Work

Evaluating the trustworthiness of sources has been studied for *fact-finding*, *truth discovery* and *crowdsourcing*. In the context of fact-finding (Vydiswaran et al.; Pasternack and Roth, 2013) and truth discovery (Yin et al., 2008; Dong et al.; Zhao et al., 2012; Li et al., 2014; Pochampally et al., 2014; Dong et al., 2015; Li et al., 2016), the solutions estimate the trustworthiness or credibility of sources, by resolving the conflicts of claims provided by multiple sources. The claims are usually in structured form, and conflicting values can be easily captured without noise. Works in (Vydiswaran et al.; Nakashole and Mitchell, 2014; Popat et al., 2017) further take text into consideration, however, in (Vydiswaran et al.; Nakashole and Mitchell, 2014), they still depend on a structured input form and thus the connection between evidence and conflicting claims are given, which is usually not practical. Popat et al. (2017) leverages text as evidence to do fact-checking, while their estimation of credibility of sources neglects the reliability of sources generating evidence. In crowdsourced labeling (Sabou et al., 2012; Hovy et al., 2013; Gao et al., 2015), the system is given noisy labels which are annotated by different annotators. The input is again in structured form, and there is no evidence to consider. This is a limited setting compared with our problem. Our problem is also related to *fact-checking* (Wang et al., 2018; Thorne et al., 2018; Yin and Roth, 2018; Zhao et al., 2018), however they only consider if the evidence can support the claim without tracking the source of the claim and evidence.

5 Conclusions and Future Work

This paper studied the problem of estimating the trustworthiness of given information sources. The sources make direct claims or indirect claims by generating evidence that implies these claims.

We proposed a probabilistic framework, JELTA, which jointly considers both kinds of assertions to better estimate claims' veracity and sources' trustworthiness. We evaluated JELTA over both synthetic and real datasets, and our results show significant improvements over baselines.

While we presented the framework here as applying to claims with two truth values, we believe that this framework can apply more broadly. For example, rather than considering a claim as being True or False, (Chen et al., 2019) suggests that one needs to view a claim from a diverse, yet comprehensive, set of *perspectives*. Our framework can be extended to deal with sources that generate a spectrum of perspectives, each with a stance relative to claim and with evidence supporting it. We leave this for future work.

Acknowledgments

This work is partly supported by a Google gift and by DARPA, under agreement number HR0011-18-2-0052.

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *ICLR*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. In *NAACL*.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzoto. 2013. Recognizing textual entailment: Models and applications.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Quang Do, Dan Roth, Mark Sammons, Yuancheng Tu, and V Vydiswaran. 2009. Robust, light-weight approaches to compute lexical similarity. Computer Science Research and Technical Reports.

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9):938–949.

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.

Edgar C Fieller, Herman O Hartley, and Egon S Pearson. 1957. Tests for rank correlation coefficients. i. *Biometrika*, 44(3/4):470–481.

Jing Gao, Qi Li, Bo Zhao, Wei Fan, and Jiawei Han. 2015. Truth discovery and crowdsourcing aggregation: A unified perspective. *Proceedings of the VLDB Endowment*, 8(12):2048–2049.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1187–1198. ACM.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16.

Ndapandula Nakashole and Tom M Mitchell. 2014. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1009–1019.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Jeff Pasternack and Dan Roth. 2013. Latent credibility analysis. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1009–1020. ACM.

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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. 2014. Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 433–444. ACM.

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.

Marta Sabou, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, page 17. ACM.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

VG Vydiswaran, ChengXiang Zhai, and Dan Roth. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 525–533. International World Wide Web Conferences Steering Committee.

Wenpeng Yin and Dan Roth. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. *arXiv preprint arXiv:1808.03465*.

Xiaoxin Yin, Jiawei Han, and S Yu Philip. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808.

Bo Zhao, Benjamin IP Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561.

Shuai Zhao, Bo Cheng, Hao Yang, et al. 2018. An end-to-end multi-task learning model for fact checking. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 138–144.

A Appredix

A.1 Inference

To infer the value of latent variables and parameters in our model, we use EM algorithm to first estimate the true claim, and then find the maximum a posterior point estimate of the parameters. As shown in Section 2.4, given parameters θ^t and X, E-step is easy to compute, while the M-step is more complicated. Since there are no closed form solutions for those parameters, we use gradient ascent to solve them and do them parameter-by-parameter.

For H_s, we have:

$$\frac{\partial P(X,Y|\theta)}{\partial H_s} = \frac{\sum_m \sum_{y_m} P(y_m|X,\theta^t) w_{s,m} \eta_d (b_{s,y_m} - H_s)}{H_s - {H_s}^2} + \lambda\Big(\log\frac{R_s}{R_s+Q_s} - \log H_s - 1\Big) \tag{12}$$

Then, for R_s, Q_s and W, the derivatives are as follows:

$$\frac{\partial P(X,Y|\theta)}{\partial R_s} = \sum_m \sum_{y_m} P(y_m|X,\theta^t)\eta_i \sum_e w_{s,e} w_{e,m} \Big[\frac{b_{s,e,y_m}}{R_s} + \frac{(1-b_{s,e,y_m})\phi_w(e,m,c)Q_s}{R_s(Q_s+R_s)\phi_w(e,m,c) - (Q_s+R_s)^2}\Big] + \lambda \cdot H_s \frac{Q_s}{R_s(Q_s+R_s)} \tag{13}$$

$$\frac{\partial P(X,Y|\theta)}{\partial Q_s} = \sum_m \sum_{y_m} P(y_m|X,\theta^t)\eta_i \sum_e w_{s,e} w_{e,m} \frac{(1-b_{s,e,y_m})\phi_w(e,m,c)R_s}{(Q_s+R_s)^2 - R_s(Q_s+R_s)\phi_w(e,m,c)} - \lambda \cdot \frac{H_s}{R_s+Q_s} \tag{14}$$

$$\frac{\partial P(X,Y|\theta)}{\partial w_i} = \sum_m \sum_{y_m} P(y_m|X,\theta^t)\eta_i \sum_s \sum_e w_{s,e} w_{e,m} \Big[b_{s,e,y_m} + \frac{H_s(1-b_{s,e,y_m})\phi_w(e,m,c)}{H_s \cdot \phi_w(e,m,c) - 1}\Big]\Big(1 - \phi_w(e,m,c)\Big) z_i \tag{15}$$