Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

ClaimEval: Integrated and Flexible Framework for Claim Evaluation Using Credibility of Sources

Mehdi Samadi
Carnegie Mellon University
msamadi@cs.cmu.edu

Partha Talukdar
Indian Institute of Science
ppt@serc.iisc.in

Manuela Veloso
Carnegie Mellon University
veloso@cs.cmu.edu

Manuel Blum
Carnegie Mellon University
mblum@cs.cmu.edu

Abstract

The World Wide Web (WWW) has become a rapidly growing platform consisting of numerous sources which provide supporting or contradictory information about claims (e.g., *"Chicken meat is healthy"*). In order to decide whether a claim is true or false, one needs to analyze content of different sources of information on the Web, measure credibility of information sources, and aggregate all these information. This is a tedious process and the Web search engines address only part of the overall problem, viz., producing only a list of relevant sources. In this paper, we present ClaimEval, a novel and integrated approach which given a set of claims to validate, extracts a set of pro and con arguments from the Web information sources, and jointly estimates credibility of sources and correctness of claims. ClaimEval uses Probabilistic Soft Logic (PSL), resulting in a flexible and principled framework which makes it easy to state and incorporate different forms of prior-knowledge. Through extensive experiments on real-world datasets, we demonstrate ClaimEval's capability in determining validity of a set of claims, resulting in improved accuracy compared to state-of-the-art baselines.

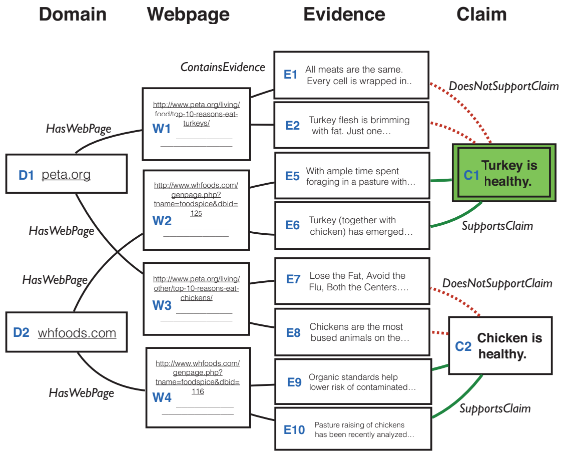
Domain
Webpage
Evidence
Claim
ContainsEvidence
HasWebPage
HasWebPage
HasWebPage
HasWebPage
D1 peta.org
D2 whfoods.com
W1 http://www.peta.org/living/food/top-10-reasons-eat-turkeys/
W2 http://www.whfoods.com/genpage.php?tname=foodspice&dbid=125
W3 http://www.peta.org/living/other/top-10-reasons-eat-chickens/
W4 http://www.whfoods.com/genpage.php?tname=foodspice&dbid=116
E1 All meats are the same. Every cell is wrapped in..
E2 Turkey flesh is brimming with fat. Just one…
E5 With ample time spent foraging in a pasture with…
E6 Turkey (together with chicken) has emerged…
E7 Lose the Fat, Avoid the Flu, Both the Centers….
E8 Chickens are the most bused animals on the…
E9 Organic standards help lower risk of contaminated…
E10 Pasture raising of chickens has been recently analyzed…
DoesNotSupportClaim
C1 Turkey is healthy.
SupportsClaim
DoesNotSupportClaim
C2 Chicken is healthy.
SupportsClaim

Figure 1: The Credibility Assessment (CA) Graph constructed to evaluate C2, while the user has already specified the other claim C1 to be true (green concentric rectangle). In addition to the claims, the CA graph also consists of nodes corresponding to *Domains*, *Webpages* and *Evidences*. While some evidences are supportive of the claims (solid green lines), others are not (red dotted line) (see Section for details). ClaimEval, the proposed system, estimates credibility of the domains based on the assessment available on claim C1, and combines that with the level of support from evidences originating from those domains to evaluate claim C2 (see Section for more details).

Introduction

The World Wide Web (WWW) and the Web search engines, such as Google, Bing, etc., that operate over millions of documents have made information readily available to everyone. Over the last few years, several large knowledge bases (KBs), such as Freebase, Yago, Google Knowledge Graph, etc., have also been developed which makes it possible to readily evaluate factoid claims (e.g., *"Paris is the capital of France"*). In spite of this democratization of information, evaluating correctness of non-factoid *claims* (e.g., *"Turkey meat is healthy"*), is still an open challenge. This is particularly challenging as two different webpages may contain conflicting evidences even related to a single claim. For example, while an animal rights website might not support meat-eating and thus term turkey meat as unhealthy, the website of a grocery store might claim otherwise. Additionally, a scientific paper focusing on this question might provide the most authoritative answer. So ideally, one would want to trust evidences contained in the credible source (the scientific paper) and ignore the other two. Hence, given a set of claims, one needs to identify relevant sources on the Web, extract supporting and contradictory evidences from those sources, estimate source credibility, and finally aggregate all these information to evaluate the given set of claims. This is a tedious and time consuming process, and current Web search engines only address the first aspect of this bigger problem, viz., identifying relevant sources. Thus, there is a growing need for an *integrated* approach for automatic extraction of relevant evidences and sources, estimation of information source credibility and utilizing those credibility estimates in claim evaluation.

Moreover, estimating information source credibility may be subjective (Bhattacharya, Devinney, and Pillutla 1998;

Gambetta 1990). In other words, we may have to take user preferences and context into account when estimating source credibility and evaluating claims. Let us motivate this through an example shown in Figure 1. In this example, there are two claims: *"Turkey is healthy"* (C1) and *"Chicken is healthy"* (C2). These claims are either supported (*SupportsClaim*) or refuted (*DoesNotSupportClaim*) by evidences originating from webpages of two domains, *peta.org*, website of an animal rights activists' group, and *whfoods.com*, a non-profit promoting healthy food. Additionally, the user has also indicated that she believes C1 is in fact true (shown by a green double-line rectangular). Given this initial information, we would like to evaluate whether claim C2 is true or false. In this case, we find that all the evidences originating from the domain *peta.org* are in contradiction with the user's assessment of claim C1. Hence, we would like to decrease the credibility of *peta.org* and reduce influence of all evidences originating from it while evaluating C2. We note that the evidences from domain *whfoods.com* are in agreement with the user's assessment of C1, and hence we should increase its credibility. We note that these credibility adjustments are claim (and user) specific. If instead of claims involving healthy meat options, they were focused on animal cruelty, then *peta.org* might have been a very credible source.

The above example shows that the correctness estimates of claims in the Credibility Assessment (CA) graph and the credibility of non-claim nodes can be calculated by propagating any available information about the correctness of claims or the credibility information of other nodes over the CA graph. Our intuition is that humans naturally do this type of credibility and correctness propagation based on *prior-knowledge* using a set of rules. The prior-knowledge specifies how the credibility inference should be performed across the CA graph. Ideally, in order to automatically calculate the credibility of a set of claims and sources, a credibility assessment approach should be able to take the prior-knowledge as an input and incorporate it in its credibility calculation.

Relevant prior work (Pasternack and Roth 2013; 2010) in this area have addressed only parts of this overall problem, and often with restricted forms of prior-knowledge. We address this gap in this paper, and make the following contributions:

1. We propose a novel and fully-integrated technique called **ClaimEval**. Input to ClaimEval is the prior credibility assessment knowledge, and a set of claims, where the truth values of a few of them is known. ClaimEval then evaluates the truth of a set of unlabeled claims by automatically crawling the relevant information from the Web, building the CA graph, calculating the credibility of sources, and incorporating the calculated credibility scores to validate the truth of the claims.

2. ClaimEval uses Probabilistic Soft Logic (PSL) (Kimmig et al. 2012; Broecheler, Mihalkova, and Getoor 2010), resulting in a *flexible* and *principled* framework for joint credibility estimation and claim evaluation. In contrast to prior approaches, ClaimEval has the following three properties: (a) ease of incorporation of prior knowledge, (b) guaranteed convergence, and (c) interpretable credibility scores. To the best of our knowledge, ClaimEval is the first such integrated system of its kind.

3. We present experimental results on real-world data demonstrating effectiveness of ClaimEval compared to the state-of-the-art approaches.

Related Work

Given a few labeled nodes and a graph where the edge weight represents the degree of similarity between the two connected nodes, Graph-based Semi-Supervised Learning (GSSL) algorithms classify the initially unlabeled nodes in the graph (Subramanya and Talukdar 2014). Since GSSL algorithms can handle only a specific edge type (i.e., node similarity), they are not applicable to the CA graph (Figure 1) and the setting of this paper as a CA graph consists of multiple edge types with differing semantics.

To overcome the limitations of GSSL algorithms, several fact-finding algorithms have been developed which consider a bipartite version of a graph similar to that of CA, and propagate information between nodes in a non-symmetric way. Algorithms such as Sums (Kleinberg 1999), TruthFinder (Yin, Han, and Yu 2007), Generalized-Investment (Pasternack and Roth 2011), Pooled-Investment (Pasternack and Roth 2010), AccuVote (Dong, Berti-Equille, and Srivastava 2009), Average.Log (Pasternack and Roth 2010), and 3-Estimates (Galland et al. 2010) have been developed that use different propagation update functions which specify how the information flows between the nodes in such a graph. Each of these fact-finding algorithms suffers from at least one of the following main disadvantages. First, the score assigned to claims are *biased* toward favoring potentially (non-credible) sources that assert many evidences. Second, the scores assigned to the nodes are not *interpretable*. Third, the convergence of these iterative algorithms are not guaranteed. Finally, and most importantly, incorporating additional prior-knowledge to guide how the credibility information should flow over the CA-like graph is not intuitive and easy. Prior assumptions usually are incorporated through a set of cumbersome and usually non-intuitive update functions.

Some of these limitations are partially addressed by some other previous work. Pasternack and Roth (Pasternack and Roth 2010) introduced a pipelined approach which takes beliefs output by a fact-finder as an input, and "corrects" those beliefs based on some prior-knowledge defined by the user. Unlike ClaimEval, (Pasternack and Roth 2010) cannot incorporate additional prior-knowledge such as how credibility should propagate over the graph. Additionally, and in contrast to ClaimEval, the semantics of belief scores in (Pasternack and Roth 2010) is specific to the external fact-finder that is iteratively used to assign scores to the nodes in the graph, and convergence of the algorithm is also not guaranteed. To address *bias* and *interpretability* limitations of fact-finders, probabilistic fact-finding approaches are introduced that model the joint probability of the sources and the claims (Pasternack and Roth 2013; Zhao et al. 2012; Wang et al. 2011). Although these approaches provide a transparent and interpretable model, the prior knowledge can be provided only in a node-centric manner. Incorporating more sophisticated types of prior-knowledge, such as adding different types of nodes or how information should propagate over the graph, requires non-trivial modeling changes. In contrast, ClaimEval offers a flexible framework which makes it possible to add such prior-knowledge using first-order logic rules and without requiring any changes in the model.

Moreover, the above approaches do not provide an integrated system that asserts the truthfulness of claims by extracting evidences from the unstructured sources of information on the Web. OpenEval (Samadi, Veloso, and Blum 2013) addresses this issue but it doesn't take source credibility into account. Similarly, Defacto (Lehmann et al. 2012) only uses PageRank as an estimation for the credibility of source. In contrast, ClaimEval identifies relevant sources, extracts evidences from them, estimates source credibility and uses those credibility scores for improved claim evaluation, all in an integrated system.

The notion of trust has been also addressed in other areas. *Reputation-based systems* are based on the notion of transitive trust and are used to measure credibility of entities such as webpages (Page et al.), people (Levien et al. 1998), and peers in network (Kamvar, Schlosser, and Garcia-Molina 2003). In *computational trust*, variations of probabilistic logic frameworks have been used to present a formal definition of trust and trust propagation, which can potentially be used in artificial agents to make trust-based decisions (Marsh 1994; Manchala 1998; Jøsang, Marsh, and Pope 2006). For *data integration*, different approaches are built to integrate the most complete and accurate records from diverse set of sources where the data sources are extremely heterogeneous (Zhao et al. 2012; Dong and Srivastava 2013; Li et al. 2014). In *crowdsourcing*, various techniques have been developed to automatically validate the quality of crowd answers, by identifying the faulty workers (Venanzi, Rogers, and Jennings ; Richardson and Domingos ; Wang et al. 2013; Nguyen et al. 2015). In *social networks*, Probabilistic Soft Logic (PSL) is used to model the trust in the social interactions (Huang et al. 2012; 2013). The truthfulness of *deep web data* has been studied by (Li et al. 2013), which surprisingly shows a large amount of inconsistency between the data provided by different sources. We point out that, while relevant, none of these prior work directly address the problem of claim evaluation in its full complexity as considered in this paper.

Our Approach: ClaimEval

ClaimEval performs joint estimation of source credibility and claim evaluation in one single model using Probabilistic Soft Logic (PSL). ClaimEval consists of the following two steps: (i) **Credibility Assessment(CA) Graph Construction** (Section), and (ii) **Joint Source Credibility Estimation and Claim Evaluation** (Section). In the next two sections, we describe each one of these two steps in greater detail. Algorithm 1 summarizes the process that ClaimEval follows for evaluating a claim.

Algorithm 1 ClaimEval - Evaluating Correctness of a Claim

```
Input: $\langle U, L, r, K \rangle$ /* $U$ is a set of unlabeled claims that should
    be evaluated, $L$ is a set of labeled claims (label is either true or
    false), $r$ is the category that $U$ and $L$ belong to, and $K$ is a set
    of first-order logic rules defining the prior credibility knowledge. */
Output: $\langle$label (True or False), confidence$\rangle$ for claims in $U$
$W$ $\leftarrow$Search the Web using Bing and extract webpages for claims in $L \cup U$
CAG $\leftarrow$ Construct layers of CA graph by parsing content of webpages in $W$
Connect nodes between different layers of CAG
Label edges that are connecting *Evidence* and *Claim* layers, using the evidence classifier trained for category $r$.
Label nodes in the *Claim* layer for which we know the truth label (i.e., claims that exist in $L$).
$R \leftarrow$ For all the nodes and edges in CAG, instantiate variables of the rules that are defined in prior knowledge $K$, Equation 4, and Equation 5.
$\hat{R}$ $\leftarrow$ Relax logical operators in $R$ using *Luka-siewicz* real-values operators
$\langle$label,confidence$\rangle$ $\leftarrow$ Convert $\hat{R}$ to an optimization problem using Probabilistic Soft Logic (PSL) and find the label (True or False) and confidence value of claims in $U$
```

Credibility Assessment (CA) Graph Construction

Let us consider the two claims: *"Turkey is healthy"* (C1) and *"Chicken is healthy"* (C2) as shown in Figure 1. In practice, in order to evaluate these claims, one would probably first try to identify sources (e.g., webpages) which are likely to contain evidence either in favor or against such claims. An evidence may be considered as a text snippet (e.g., a sentence or paragraph in a webpage) expressing opinion about the claim in a webpage. Overall, a *claim* is substantiated by one or more *evidences* which are contained in *webpages*, with webpages in turn contained within *domains*. We can put all this together in a multi-relational graph which we shall call a Credibility Assessment (CA) graph. A CA graph has four types of nodes: *Domain*, *Webpage*, *Evidence*, and *Claim*. Nodes are connected by the following types of edges: (1) *HasWebPage(Domain, WebPage)*, (2) *ContainsEvidence(WebPage, Evidence)*, (3) *DoesNotSupportClaim(Evidence, Claim)*, and (4) *SupportsClaim(Evidence, Claim)*. For example, in Figure 1, evidence node *"All meats are the same..."* (E1) *DoesNotSupportClaim* C1. Evidence E1 is found in webpage represented by webpage node W1. This particular webpage is from the *peta.org* domain which is represented by node D1 in the CA graph.

As an input, we assume that a set of categories of claims is given to our system. For each category, a set of *true* and *false* claims are provided (i.e., category instances). For example, *healthy food* is an example of a category, and $\{$*apple, broccoli*$\}$ and $\{$*mayonnaise, soda*$\}$ are, respectively, *true* and *false* claims provided for this category. For each category, we assume that a set of labeled and unlabeled claims are provided (e.g., a set of *true* and *false* claims). Our goal is to classify the set of unlabeled claims to either *true* or *false* classes, with a confidence value attached to the label.

Training Evidence Classifier: To build the CA graph, ClaimEval first learns an *evidence classifier* for each category of claims, which is later used to classifying evidences, for a given claim, as either *SupportsClaim* (*pro*) or *DoesNotSupportClaim* (*con*) classes. This step is performed only once during training for each category. The trained classifier is then used during the test time. To build the training data for the classifier, ClaimEval first iterates over all the claims for which the label is known. ClaimEval then converts each claim to a search query, where the search query is built from the claim and the name of the category. ClaimEval then searches the query on the Web (e.g., using Bing) and downloads the set of the highest ranked webpages. In each of the returned webpages, ClaimEval extracts all the paragraphs that contain the claim phrase. All the words in each paragraph is saved as an *evidence*. The extracted evidences get the same labels as the labels of claims. By assuming that the training data extracted from the search query are ground-truth, ClaimEval trains a classifier (e.g., SVM) for each category of claims. For the input to the classifier, each evidence is represented using standard bag-of-words model (bigrams and unigrams).

Constructing CA Graph: After training the evidence classifier, ClaimEval iterates over all the labeled and unlabeled claims, similar to the training process, and extracts a set of evidences for each claim. Using the trained evidence classifier, each of the extracted evidences is then assigned to either the *pro* or the *con* classes. The evidence layer in the CA graph is built using all the evidences extracted from the Web. Each evidence is connected to the claim for which it is extracted. The edge that connects an evidence to a claim, is classified as *SupportsClaim* (solid green line in Figure 1), if the evidence is supporting the claim, and otherwise is classified as *DoesNotSupportClaim* (red dotted line). The webpage layer is constructed from the webpages that are returned by the search engine. Each webpage is connected to the set of evidences extracted from the webpage. For each webpage, we extract the domain name (e.g., *whfoods.org*) and create a corresponding node in the domain layer. Each domain node is connected to its webpages in the webpage layer.

Initially, all the nodes in the graph do not have any value, except for the nodes whose true labels (i.e., user assessments) are known. Each node in the CA graph takes a value in the range [0,1]. The semantic of the values assigned to the nodes varies across the layers in the graph. For example, the value of a node in the webpage layer is interpreted as the degree of the credibility of the webpage, and the value of a node in the claim layer is interpreted as the confidence in the truth of a claim. Edges in the graph can be seen as transforming the meaning of *values* of nodes between the layers.

Joint Source Credibility Estimation & Claim Evaluation

When computing the credibility scores in the CA graph, we propagate the credibility and the correctness information across the different layers based on some *prior knowledge*, which is defined as a set of rules. We first list a set of such rules that specify how the credibility inference should be performed across the CA graph. We later explain how these rules are incorporated in our credibility assessment model.

Prior Knowledge for Credibility Assessment

- **Evidence ⇒ Claim**: Inferring correctness of a claim based on the credibility of an evidence:
  - EC1: Evidence is *credible* & evidence *supports* claim ⇒ claim is *true*.
  - EC2: Evidence is *credible* & evidence *doesn't support* claim ⇒ claim is *false*.
  - EC3: Evidence is not *credible*, then the evidence has no effect on the correctness of the claim.
- **Claim ⇒ Evidence**: Inferring credibility of an evidence based on the correctness of a claim:
  - CE1: Claim is *true* & evidence *supports* claim ⇒ evidence is *credible*.
  - CE2: Claim is *true* & evidence *doesn't support* claim ⇒ evidence is not *credible*.
  - CE3: Claim is *false* & evidence *supports* claim ⇒ evidence is not *credible*.
  - CE4: Claim is *false* & evidence *doesn't support* claim ⇒ evidence is *credible*.
- **Webpage ⇔ Evidence**: Inferring credibility of an evidence based on the credibility of a webpage, and vice versa:
  - WE1: Webpage is *credible* ⇔ evidence is *credible*.
  - WE2: Webpage is not *credible* ⇔ evidence is not *credible*.
- **Domain ⇔ Webpage**: Inferring credibility of a webpage from the credibility of a domain, and vice versa:
  - DW1: Domain is *credible* ⇔ webpage is *credible*.
  - DW2: Domain is not *credible* ⇔ webpage is not *credible*.

Encoding Prior Knowledge using First-Order Logic

We use first-order logic (FOL) to formally define a set of rules, based on the prior knowledge that we defined in the previous section. Each layer of the CA graph is represented by a logical *predicate*, and each node in the graph is an instance of the predicate. For example, the predicate Domain(*x*) is used to define nodes in the domain layer. In this case, predicate instance *Domain(peta.org)* has value 1, since *peta.org* is a domain in the domains layer, otherwise it would have taken the value 0. Similarly, predicates *Webpage(x)*, *Evidence(x)*, and *Claim(x)* are defined to represent nodes in the other layers of the graph.

In addition to the predicates that represent different types of the nodes in the graph, we use the following predicates to define edges: *HasWebpage(Domain, Webpage)*, *ContainsEvidence(Webpage, Evidence)*, *SupportsClaim(Evidence, Claim)*, *DoesNotSupportClaim(Evidence, Claim)*. Values of the two predicates connecting an Evidence with a Claim is computed by the evidence classifier as described in the previous section.

We represent each of the rules that are defined as part of the prior knowledge in first-order logic. For example, the EC1 prior knowledge rule from above may be represented using the following rule:

$$\begin{aligned} &\mathit{Evidence}(e) \wedge\ \mathit{Claim}(c) \wedge\ \mathit{Credible}(e) \\ &\quad \wedge\ \mathit{SupportsClaim}(e,c) \Rightarrow\ \mathit{Correct}(c) \end{aligned} \tag{1}$$

In this case, if the evidence is credible (*Credible(e)* has value 'true' or 1.0), *e* supports claim *c* (i.e., *SupportsClaim(e, c)* = 1), then *c* is labeled *true* (i.e., *Correct(c)* = 1). Other rules are defined similarly.

Although the logical rules precisely define the flow of information in the graph, their applicability is limited on real world examples, as the nodes in the graph can only take values 1 (*true*) or 0 (*false*). For example, assume that a claim *c* is supported by two evidences *a* and *b*, both of which are credible. If *a* is in favor of *c* and *b* is against *c*, then we can neither say *c* is *true* nor that *c* is *false*. Ideally, the value of claim *c* is between 0 and 1 depending on the relative credibility of evidences *a* and *b*. In the next section, we explain how to address this issue.

Going Beyond Binary Logical Operators using Probabilistic Soft Logic (PSL) Probabilistic Soft Logic (PSL) is a general purpose logical system that uses First-Order Logic (FOL) as its underlying logical language. We overview the PSL technique briefly in this section; for a more detailed exposition of PSL, we refer the reader to (Bach et al. 2012; Broecheler, Mihalkova, and Getoor 2010; Kimmig et al. 2012).

In PSL, the value of each ground atom is relaxed to take a soft truth-value in the interval [0,1], where 0 is interpreted as absolute false and 1.0 as absolute true. The soft-truth assignment allows us to define the degree of correctness and credibility for the nodes in the graph. For example, if we have two sources s_1 and s_2, where *Credible*(s_1) $>$ *Credible*(s_2), then we can infer that s_1 is more credible compared to s_2.

To handle the continuous truth values in the variables, PSL relaxes conjunction, disjunction, and negation logical operators by using *Luka-siewicz* real-valued logic operators, which are defined as follows:

$$\begin{aligned} a \mathbin{\hat{\wedge}} b &= \max\{0, a+b-1\} \\ a \mathbin{\hat{\vee}} b &= \max\{1, a+b\} \\ \hat{\neg}\, a &= 1-a \end{aligned}$$

Lukasiewicz t-norm operators, compared to other non-classical logic operators such as *product t-norm*, are suitable for our credibility assessment model since they linearly combine the values that they take. For example, consider this rule:

$$\textit{Credible(e)} \mathbin{\hat{\wedge}} \textit{SupportsClaim(e,c)} \Rightarrow \textit{Correct(c)}$$

Using the *Lukasiewicz* disjunction operation, we can write the body of this rule as:

$$\max\ \{0, \textit{Credible(e)} + \textit{SupportsClaim(e, c)} - 1\}$$

which evaluates to *true* when the resulting value is greater than certain threshold, say 0.5. We can roughly interpret it as: claim c is *correct* only when e is credible and supports claim c, each at least by a degree of 0.5 (sum should be greater than 1.5).

Given a set of rules, we first instantiate all the variables in the rules with respect to the CA graph that is constructed. Given these instantiated rules, our goal is to find values of the different nodes in the graph such that the total number of satisfied rules is maximized. For solving this optimization problem, PSL defines the satisfaction distance for each rule in the domain. Given a ground rule r, such as $r := P \Rightarrow Q$, r is satisfied if and only if $\hat{V}(Q) \geq \hat{V}(P)$, where $\hat{V}(X)$ is defined as the value of the ground logical expression X using *Lukasiewicz* operators (i.e., Q is at least as truthful as P). The rule's *distance to satisfaction* (denoted by $d(r)$) measures the satisfaction degree of each rule $r := P \Rightarrow Q$:

$$d(r) = \max\{0, \hat{V}(P) - \hat{V}(Q)\} \tag{2}$$

Given the distance function in Equation 2 and a set of ground rules, PSL finds values for all the ground predicate instances in order to minimize the total satisfaction degrees of all the ground rules. To do this, assume that I is the assignment of values to predicate instances, and $d_I(r)$ is the satisfaction degree of rule r given assignment I. Thus, given the set of ground rules R, the optimal assignment I^* may be obtained as follows,

$$I^* \ \leftarrow \ \arg\max_{I} \ \frac{1}{Z} \exp[-\sum_{r \in R} \lambda_r (d_I(r))^2] \tag{3}$$

where Z is the normalization factor, and λ_r is the weight for each ground rule r. Most Probable Explanation (MPE) inference algorithm may be used to optimize Equation 3.

When optimizing Equation 3, we allow only the values of *Credible* and *Correct* predicate instances to change, and the values of the rest of predicate instances remain fixed. Also, in order to make sure that the labeled claims will preserve their values during the optimization (i.e., constraints), we define the following rules:

$$\forall \text{ labeled claims } c,\ \textit{PosClaim}(c) \rightarrow\ \textit{Correct}(c) \tag{4}$$

$$\forall \text{ labeled claims } c,\ \textit{NegClaim}(c) \rightarrow \neg\ \textit{Correct}(c) \tag{5}$$

Predicates *PosClaims* and *NegClaims* respectively define the positive and negative set of labeled claims. Ideally, the above rules should have higher weight compared to the other rules defined as part of the prior knowledge, in order to make sure that during the optimization, the correct labels are assigned to the labeled data. The weight of rules can be tuned manually or can be learnt by using using the Maximum Likelihood Estimation (MLE) (Broecheler, Mihalkova, and Getoor 2010).

Experimental Evaluation

Setup: ClaimEval is tested on nine different sets of categories. Table 1 lists all the categories in the leftmost column. For each category, about 40 seed examples are used to train a SVM-based evidence classifier (used to annotate evidences with *pro* or *con* labels), and about 40 examples are used as test data. All the data sets, including the exact number of training examples for each category, and the methodology used to obtain the train and test data for each category are available at: www.ClaimEval.com. To train the evidence classifier, we use the top ten pages returned by Bing, and all the evidences extracted from the returned search pages. 0.5 is used as the confidence threshold for all experiments.

Baselines: We compare the accuracy of ClaimEval to the following Fact-Finding approaches: **Majority Vote**

Category	MV	GS	TF	AL	GI	PI	CE
Healthy Food	0.89	0.89	0.69	0.86	0.89	0.89	**0.91**
Company with Stock Growth	0.62	0.65	0.65	0.62	0.62	0.60	**0.72**
High Ranked Universities	0.80	0.82	0.73	**0.85**	0.80	0.80	**0.85**
Top CS Journals	0.74	0.71	**0.82**	0.67	0.71	0.71	0.79
Top CS Conferences	0.53	0.55	0.58	0.57	0.53	0.53	**0.68**
High GDP Growth	0.60	**0.70**	0.50	**0.70**	0.6	0.50	0.60
High HDI Growth	0.81	0.53	0.65	0.63	0.81	**0.86**	0.76
High Crime Rate Cities	0.67	0.63	0.67	0.60	**0.80**	**0.80**	**0.80**
Top Soccer Club Teams	0.65	0.62	0.65	0.65	**0.69**	0.62	**0.69**
Average	0.71	0.68	0.66	0.69	0.72	0.71	**0.76**

Table 1: The accuracy of Majority Vote (MV), Generalized Sums (GS), TruthFinder (TF), Average-Log (AL), Generalized Investment (GI), Pooled-Investment (PI), and ClaimEval (CE) techniques in predicting the truth values for different categories of claims. Maximum value of each row shown in bold. ClaimEval, the proposed system, achieves the best overall performance.

(MV), **Generalized Sums (GS)** which is based on Hubs and Authorities algorithm (Kleinberg 1999), **Truth Finder (TF)** (Yin, Han, and Yu 2007), **Average-Log (AL)** (Pasternack and Roth 2011), **Generalized Investment (GI)** (Pasternack and Roth 2011), **Pooled-Investment (PI)** (Pasternack and Roth 2010). All these baseline approaches operate over a bipartite graph construction (two layers: sources and claims) instead of CA graph. In order to compare ClaimEval with these approaches, we construct the equivalent bipartite graph from a CA graph by marginalizing over evidence nodes and dropping the domain nodes. Apart from MV, all other baselines make use of the available labeled claims.

Main Result: Table 1 shows the accuracy of baselines and ClaimEval for nine different categories. The experiments are obtained when 60 evidences are randomly extracted for each claim from the first 10 pages returned by Bing.

From Table 1 we observe that the performance of fact-finding algorithms may be category-specific, which is consistent with similar observations in (Pasternack and Roth 2011). Overall, ClaimEval, the proposed system, achieves best performance in 6 out of the 9 categories, and on average outperforming all other state-of-the-art baselines.

Change in Performance with Increasing Evidence Size: Figure 2 shows accuracy of different methods when increasing number of evidences are used. As we increase the number of evidences, ClaimEval consistently outperforms all other methods. Among the baselines, *Generalized Investment* is more successful in making use of the increased evidence size, and this is consistent with prior research (Pasternack and Roth 2011). *TruthFinder*'s performance peaks with 30 evidences, but its performance degrades with more evidences due to over-fitting.

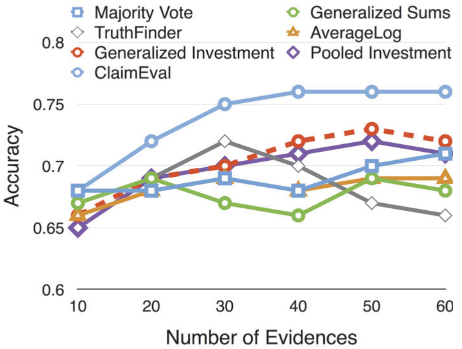

Figure 2: Performance of different systems when increasing amounts of evidence is available. ClaimEval, the proposed system (top-most plot), is best able to exploit additional evidence achieving best overall performance.

Qualitative Example: As described in the introduction, one of the main disadvantages of most of the fact-finding algorithms (including the baselines considered in this paper) is their bias towards increased credibility for sources that assert many claims.

Figure 3 shows an example of a CA graph, where a domain node 1 (through webpage node 3) is asserting overwhelmingly many evidences against claim node 93 (negative evidence is marked by dotted red line in the figure, as in Figure 1). Claim node 92 with concentric circles is known to be correct a-priori. Correctness of the other claim node 93 needs to be evaluated. Based on these input information, judgments of a human annotator is shown on top of each node. Most of the fact-finding algorithms (including MV, AL, and GS) are biased towards favoring domain nodes with many evidences. In this example, these algorithms overestimate the credibility of nodes 1 and 3, and incorrectly classify claim 93 as incorrect (as all evidences from source node 3 oppose claim 93). In contrast, ClaimEval is able to infer that nodes 1 and 3 have low credibility since they provide contradictory information regarding claim 92. ClaimEval then uses this reduced credibility to correctly classify claim 93 as true. This examples provides qualitative evidence of how ClaimEval is able to overcome bias of many existing fact-finding algorithms.

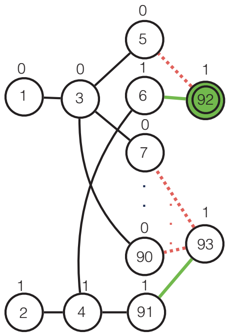

Figure 3: An example CA graph, with annotator judgments marked on top of each node. While baselines such as MV, GS and AL overfit by overtrusting domains with many evidences (e.g., nodes 1 and 3 here), ClaimEval is able to match annotator judgments.

Conclusion

In this paper, we proposed ClaimEval, an integrated, flexible, and principled system for claim evaluation. In contrast to previous approaches for claim evaluation, given a set of claims, ClaimEval identifies a set of relevant sources and evidences within them which might support or refute the claims, estimates credibility of those sources, and uses those credibility estimates to evaluate correctness of the claims – all in a single integrated system. ClaimEval uses Probabilistic Soft Logic (PSL) to flexibly incorporate various types of prior-knowledge from the user. Through extensive experiments on real-world datasets, we demonstrated ClaimEval's effectiveness over other state-of-the-art baselines. As part of future work, we hope to exploit the flexibility offered by ClaimEval and incorporate other types of prior-knowledge and user preferences and evaluate their effect on claim evaluation.

Acknowledgments

The authors would like to thank Uday Saini and Ankit Sharma for carefully reading a draft of this paper. This work is supported in part by gifts from Google and Accenture, and research grants from Pradeep Sindhu Fellowship and the National Science Foundation (under Grants CCF-1101668 and CCR-0122581).

References

Bach, S.; Broecheler, M.; Getoor, L.; and leary, D. O. 2012. Scaling mpe inference for constrained continuous markov random fields with consensus optimization. In *NIPS*.

Bhattacharya, R.; Devinney, T. M.; and Pillutla, M. M. 1998. A Formal Model of Trust Based on Outcomes. *The Academy of Management Review* 23(3):459–472.

Broecheler, M.; Mihalkova, L.; and Getoor, L. 2010. Probabilistic similarity logic. In *UAI*.

Dong, X., and Srivastava, D. 2013. Big data integration. In *ICDE*.

Dong, X. L.; Berti-Equille, L.; and Srivastava, D. 2009. Truth discovery and copying detection in a dynamic world. *Proc. VLDB Endow.* 2(1):562–573.

Galland, A.; Abiteboul, S.; Marian, A.; and Senellart, P. 2010. Corroborating information from disagreeing views. In *WSDM*.

Gambetta, D. 1990. *Trust: Making and Breaking Cooperative Relations*. B. Blackwell.

Huang, B.; Bach, S. H.; Norris, E.; Pujara, J.; and Getoor, L. 2012. Social group modeling with probabilistic soft logic. In *NIPS Workshop on Social Network and Social Media Analysis: Methods, Models, and Applications*.

Huang, B.; Kimmig, A.; Getoor, L.; and Golbeck, J. 2013. A flexible framework for probabilistic models of social trust. In *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP)*.

Jøsang, A.; Marsh, S.; and Pope, S. 2006. Exploring different types of trust propagation. In *Proceedings of the 4th International Conference on Trust Management*.

Kamvar, S. D.; Schlosser, M. T.; and Garcia-Molina, H. 2003. The eigentrust algorithm for reputation management in p2p networks. In *WWW*.

Kimmig, A.; Bach, S. H.; Broecheler, M.; Huang, B.; and Getoor, L. 2012. A short introduction to probabilistic soft logic. In *NIPS Workshop on Probabilistic Programming: Foundations and Applications*.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46(5):604–632.

Lehmann, J.; Gerber, D.; Morsey, M.; and Ngomo, A. N. 2012. Defacto - deep fact validation. In *ISWC*.

Levien, R.; Aiken, A.; Levien, R.; and Aiken, A. 1998. Attack resistant trust metrics for public key certification. In *In 7th USENIX Security Symposium*.

Li, X.; Dong, X. L.; Lyons, K.; Meng, W.; and Srivastava, D. 2013. Truth finding on the deep web: is the problem solved? In *VLDB*.

Li, Q.; Li, Y.; Gao, J.; Zhao, B.; Fan, W.; and Han, J. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD*.

Manchala, D. W. 1998. Trust metrics, models and protocols for electronic commerce transactions. In *Proceedings of the 18th International Conference on Distributed Computing Systems*.

Marsh, S. P. 1994. Formalising trust as a computational concept. Technical report, Department of Computing Science and Mathematics, University of Stirling.

Nguyen, Q. V. H.; Duong, C. T.; Weidlich, M.; and Aberer, K. 2015. Minimizing Efforts in Validating Crowd Answers. In *SIGMOD*.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Pasternack, J., and Roth, D. 2010. Knowing what to believe (when you already know something). In *COLING*.

Pasternack, J., and Roth, D. 2011. Making better informed trust decisions with generalized fact-finding. In *IJCAI*.

Pasternack, J., and Roth, D. 2013. Latent credibility analysis. In *WWW*.

Richardson, M., and Domingos, P. Building large knowledge bases by mass collaboration. In *Proceedings of the 2nd International Conference on Knowledge Capture*.

Samadi, M.; Veloso, M.; and Blum, M. 2013. Openeval: Web information query evaluation. In *AAAI*.

Subramanya, A., and Talukdar, P. P. 2014. Graph-based semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 8(4):1–125.

Venanzi, M.; Rogers, A.; and Jennings, N. R. Trust-based fusion of untrustworthy information in crowdsourcing applications. In *AAMAS 2013*.

Wang, D.; Abdelzaher, T.; Ahmadi, H.; Pasternack, J.; Roth, D.; Gupta, M.; Han, J.; Fatemieh, O.; Le, H.; and Aggarwal, C. C. 2011. On Bayesian interpretation of fact-finding in information networks. In *Information Fusion*.

Wang, D.; Abdelzaher, T.; Kaplan, L.; and Aggarwal, C. 2013. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In *ICDCS*.

Yin, X.; Han, J.; and Yu, P. S. 2007. Truth discovery with multiple conflicting information providers on the web. In *KDD*.

Zhao, B.; Rubinstein, B. I. P.; Gemmell, J.; and Han, J. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. *Proc. VLDB Endow.* 5(6).