The Wisdom of Minority: Unsupervised Slot Filling Validation based on Multi-dimensional Truth-Finding

Dian Yu[1], Hongzhao Huang[1], Taylor Cassidy[2,3], Heng Ji[1]
Chi Wang[4], Shi Zhi[4], Jiawei Han[4], Clare Voss[2], Malik Magdon-Ismail[1]

[1]Computer Science Department, Rensselaer Polytechnic Institute
[2]U.S. Army Research Lab [3]IBM T. J. Watson Research Center
[4]Computer Science Department, Univerisity of Illinois at Urbana-Champaign
[1]{yud2,huangh9,jih,magdon}@rpi.edu,
[2,3]{taylor.cassidy.ctr,clare.r.voss.civ}@mail.mil
[4]{chiwang1,shizhi2,hanj}@illinois.edu

Abstract

Information Extraction using multiple information sources and systems is *beneficial* due to multi-source/system consolidation and *challenging* due to the resulting inconsistency and redundancy. We integrate IE and truth-finding research and present a novel unsupervised *multi-dimensional truth finding* framework which incorporates signals from multiple sources, multiple systems and multiple pieces of evidence by knowledge graph construction through multi-layer deep linguistic analysis. Experiments on the case study of Slot Filling Validation demonstrate that our approach can find truths accurately (9.4% higher F-score than supervised methods) and efficiently (finding 90% truths with only one half the cost of a baseline without credibility estimation).

1 Introduction

Traditional Information Extraction (IE) techniques assess the ability to *extract information from individual documents in isolation*. However, similar, complementary or conflicting information may exist in multiple heterogeneous sources. We take the Slot Filling Validation (SFV) task of the NIST Text Analysis Conference Knowledge Base Population (TAC-KBP) track (Ji et al., 2011) as a case study. The Slot Filling (SF) task aims at collecting from a large-scale multi-source corpus the values ("slot fillers") for certain attributes ("slot types") of a query entity, which is a person or some type of organization. KBP 2013 has defined 25 slot types for persons (per) (*e.g.*, age, spouse, employing organization) and 16 slot types for organizations (org) (*e.g.*, founder, headquarters-location, and subsidiaries). Some slot types take only a single slot filler (*e.g.*, per:birthplace), whereas others take multiple slot fillers (*e.g.*, org:top employees).

We call a combination of query entity, slot type, and slot filler a ***claim***. Along with each claim, each system must provide the ID of a source document and one or more detailed context sentences as ***evidence*** which supports the claim. A ***response*** (*i.e.*, a claim, evidence pair) is *correct* if and only if the claim is true *and* the evidence supports it.

Given the responses produced by multiple systems from multiple sources in the SF task, the goal of the SFV task is to determine whether each response is true or false. Though it's a promising line of research, it raises two complications: (1) different information ***sources*** may generate claims that vary in trustability; and (2) a large-scale number of SF ***systems*** using different resources and algorithms may generate erroneous, conflicting, redundant, complementary, ambiguously worded, or inter-dependent claims from the same set of documents. Table 1 presents responses from four SF systems for the query entity ***Ronnie James Dio*** and the slot type ***per:city_of_death***. Systems A, B and D return *Los Angeles*

with different pieces of evidence [1] extracted from different information sources, though the evidence of System D does not decisively support the claim. System C returns *Atlantic City*, which is neither true nor supported by the corresponding evidence.

Such complications call for "*truth finding*": determining the *veracity* of multiple conflicting claims from various sources and systems. We propose a novel unsupervised multi-dimensional truth finding framework to study credibility perceptions in rich and wide contexts. It incorporates signals from multiple sources and systems, using linguistic indicators derived from knowledge graphs constructed from multiple evidences using multi-layer deep linguistic analysis. Experiments demonstrate that our approach can find truths accurately (9.4% higher F-score than supervised methods) and efficiently (find 90% truths with only one half cost of a baseline without credibility estimation).

System	Source	Slot Filler	Evidence
A	Agence France-Presse, News	Los Angeles	The statement was confirmed by publicist Maureen O'Connor, who said ***Dio*** died in ***Los Angeles***.
B	New York Times, News	Los Angeles	***Ronnie James Dio***, a singer with the heavy-metal bands Rainbow, Black Sabbath and Dio, whose semioperatic vocal style and attachment to demonic imagery made him a mainstay of the genre, died on Sunday in ***Los Angeles***.
C	Discussion Forum	Atlantic City	***Dio*** revealed last summer that he was suffering from stomach cancer shortly after wrapping up a tour in ***Atlantic City***.
D	Associated Press Worldstream, News	Los Angeles	***LOS ANGELES*** 2010-05-16 20:31:18 UTC ***Ronnie James Dio***, the metal god who replaced Ozzy Osbourne in Black Sabbath and later piloted the bands Heaven, Hell and Dio, has died, according to his wife and manager.

Table 1: Conflicting responses across different SF systems and different sources (query entity = ***Ronnie James Dio***, slot type = ***per:city_of_death***).

2 Related Work & Our Novel Contributions

Most previous SFV work (*e.g.*, (Tamang and Ji, 2011; Li and Grishman, 2013)) focused on filtering incorrect claims from multiple systems by simple heuristic rules, weighted voting, or costly supervised learning to rank algorithms. We are the first to introduce the truth finding concept to this task.

The "truth finding" problem has been studied in the data mining and database communities (*e.g.*, (Yin et al., 2008; Dong et al., 2009a; Dong et al., 2009b; Galland et al., 2010; Blanco et al., 2010; Pasternack and Roth, 2010; Yin and Tan, 2011; Pasternack and Roth, 2011; Vydiswaran et al., 2011; Ge et al., 2012; Zhao et al., 2012; Wang et al., 2012; Pasternack and Roth, 2013)). Compared with the previous work, our truth finding problem is defined under a unique setting: each *response* consists of a claim and supporting evidence, automatically generated from unstructured natural language texts by a SF *system*. The judgement of a *response* concerns both the truth of the claim and whether the *evidence* supports the claim. This has never been modeled before. We mine and exploit rich linguistic knowledge from multiple lexical, syntactic and semantic levels from evidence sentences for truth finding. In addition, previous truth finding work assumed most claims are likely to be true. However, most SF systems have hit a performance ceiling of 35% F-measure, and false responses constitute the majority class (72.02%) due to the imperfect algorithms as well as the inconsistencies of information sources. Furthermore, certain truths might only be discovered by a minority of good systems or from a few good sources. For example, 62% of the true responses are produced only by 1 or 2 of the 18 SF systems.

3 MTM: A Multi-dimensional Truth-Finding Model

MTM Construction

A response is trustworthy if its claim is true and its evidence supports the claim. A trusted source always supports true claims by giving convincing evidence, and a good system tends to extract trustworthy responses from trusted sources. We propose a *multi-dimensional truth-finding model (MTM)* to incorporate and compute multi-dimensional credibility scores.

[1] Hereafter, we refer to "pieces of evidence" with the shorthand "evidences". Note that SF systems may include multiple sentences as "evidence" within their responses.

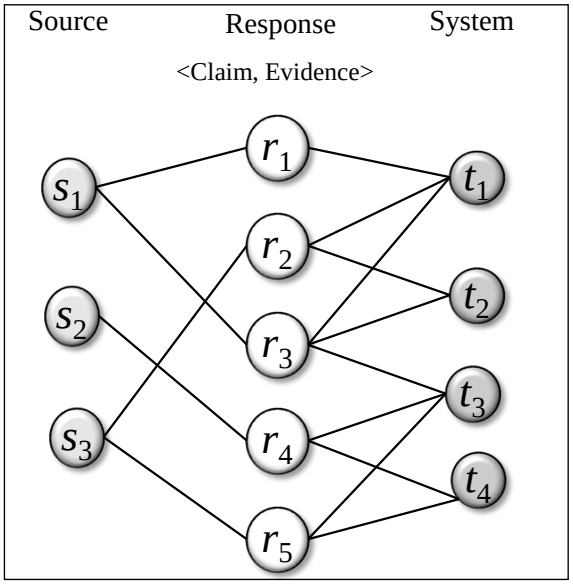

Figure 1: Heterogeneous networks for MTM.

Consider a set of responses $R = \{r_1, \ldots, r_m\}$ extracted from a set of sources $S = \{s_1, \ldots, s_n\}$ and provided by a set of systems $T = \{t_1, \ldots, t_l\}$. A heterogeneous network is constructed as shown in Fig. 1. Let weight matrices be $W^{rs}_{m\times n} = \{w^{rs}_{ij}\}$ and $W^{rt}_{m\times l} = \{w^{rt}_{ik}\}$. A link $w^{rs}_{ij} = 1$ is generated between r_i and s_j when response r_i is extracted from source s_j, and a link $w^{rt}_{ik} = 1$ is generated between r_i and t_k when response r_i is provided by system t_k.

Credibility Initialization

Each source is represented as a combination of publication venue and genre. The credibility scores of sources S are initialized uniformly as $\frac{1}{n}$, where n is the number of sources. Given the set of systems $T = \{t_1, \ldots, t_l\}$, we initialize their credibility scores $c^0(t)$ based on their interactions on the predicted responses. Suppose each system t_i generates a set of responses R_{t_i}. The similarity between two systems t_i and t_j is defined as $similarity(t_i, t_j) = \frac{|R_{t_i} \cap R_{t_j}|}{\log(|R_{t_i}|)+\log(|R_{t_j}|)}$ (Mihalcea, 2004). Then we construct a weighted undirected graph $G = \langle T, E\rangle$, where $T(G) = \{t_1, \ldots, t_l\}$ and $E(G) = \{\langle t_i, t_j\rangle\}$, $\langle t_i, t_j\rangle = similarity(t_i, t_j)$, and apply the TextRank algorithm (Mihalcea, 2004) on G to obtain $c^0(t)$.

We got negative results by initializing system credibility scores uniformly. We also got negative results by initializing system credibility scores using system metadata, such as the algorithms and resources the system used at each step, its previous performance in benchmark tests, and the confidence values it produced for its responses. We found the quality of an SF system depends on many different resources instead of any dominant one. For example, an SF system using a better dependency parser does not necessarily produce more truths. In addition, many systems are actively being improved, rendering previous benchmark results unreliable. Furthermore, most SF systems still lack reliable confidence estimation.

The initialization of the credibility scores for responses relies on deep linguistic analysis of the evidence sentences and the exploitation of semantic clues, which will be described in Section 4.

Credibility Propagation

We explore the following heuristics in MTM.

HEURISTIC 1: *A response is more likely to be true if derived from many trustworthy sources. A source is more likely to be trustworthy if many responses derived from it are true.*

HEURISTIC 2: *A response is more likely to be true if it is extracted by many trustworthy systems. A system is more likely to be trustworthy if many responses generated by it are true.*

Input: A set of responses (R), sources (S) and systems (T).
Output: Credibility scores ($c(r)$) for R.

Initialize the credibility scores $c^0(s)$ for S as $c^0(s_i) = \frac{1}{|S|}$;
Use TextRank to compute initial credibility scores $c^0(t)$ for T;
Initialize the credibility scores $c^0(r)$ using linguistic indicators (Section 4);
Construct heterogeneous networks across R, S and T;
$k \leftarrow 0$, diff $\leftarrow 10e6$;
while $k <$ MaxIteration and diff $>$ MinThreshold **do**
    Use Eq. (1) to compute $c^{k+1}(s)$;
    Use Eq. (2) to compute $c^{k+1}(t)$;
    Use Eq. (3) to compute $c^{k+1}(r)$;
    Normalize $c^{k+1}(s)$, $c^{k+1}(t)$, and $c^{k+1}(r)$;
    diff $\leftarrow \sum(|c^{k+1}(r) - c^k(r)|)$;
    $k \leftarrow k + 1$
end while

Algorithm 1: Multi-dimensional Truth-Finding.

Based on these two heuristics we design the following credibility propagation approach to mutually reinforce the trustworthiness of linked objects in MTM.

By extension of Co-HITS (Deng et al., 2009), designed for bipartite graphs, we develop a propagation method to handle heterogeneous networks with three types of objects: *source, response* and *system*. Let the weight matrices be W^{rs} (between responses and sources) and W^{rt} (between responses and systems), and their transposes be W^{sr} and W^{tr}. We can obtain the transition probability that vertex s_i in S reaches vertex r_j in R at the next iteration, which can be formally defined as a normalized weight $p_{ij}^{sr} = \frac{w_{ij}^{sr}}{\sum_k w_{ik}^{sr}}$ such that $\sum_{r_j \in R} p_{ij}^{sr} = 1$. We compute the transition probabilities p_{ji}^{rs}, p_{jk}^{rt} and p_{kj}^{tr} in an analogous fashion.

Given the initial credibility scores $c^0(r)$, $c^0(s)$ and $c^0(t)$, we aim to obtain the refined credibility scores $c(r)$, $c(s)$ and $c(t)$ for responses, sources, and systems, respectively. Starting with sources, the update process considers both the initial score $c^0(s)$ and the propagation from connected responses, which we formulated as:

$$c(s_i) = (1 - \lambda_{rs})c^0(s_i) + \lambda_{rs} \sum_{r_j \in R} p_{ji}^{rs} c(r_j) \tag{1}$$

Similarly, the propagation from responses to systems is formulated as:

$$c(t_k) = (1 - \lambda_{rt})c^0(t_k) + \lambda_{rt} \sum_{r_j \in R} p_{jk}^{rt} c(r_j) \tag{2}$$

Each response's score $c(r_j)$ is influenced by both linked sources and systems:

$$c(r_j) = (1 - \lambda_{sr} - \lambda_{tr})c^0(r_j) + \lambda_{sr} \sum_{s_i \in S} p_{ij}^{sr} c(s_i) + \lambda_{tr} \sum_{t_k \in T} p_{kj}^{tr} c(t_k) \tag{3}$$

where $\lambda_{rs}, \lambda_{rt}, \lambda_{sr}$ and $\lambda_{tr} \in [0, 1]$. These parameters control the preference for the propagated over initial score for every type of random walk link. The larger they are, the more we rely on link structure[2]. The propagation algorithm converges (10 iterations in our experimental settings) and a similar theoretical proof to HITS (Peserico and Pretto, 2009) can be constructed. Algorithm 1 summarizes MTM.

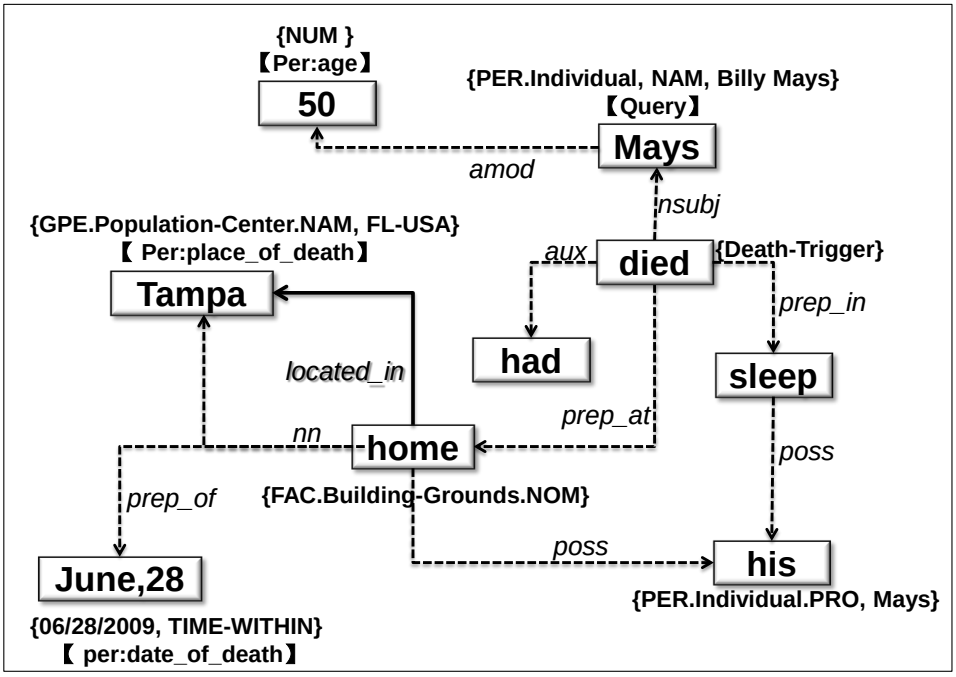

Figure 2: Knowledge Graph Example.

4 Response Credibility Initialization

Each evidence along with a claim is expressed as a few natural language sentences that include the query entity and the slot filler, along with semantic content to support the claim. We analyze the evidence of each response in order to initialize that response's credibility score. This is done using heuristic rules defined in terms of the binary outputs of various *linguistic indicators* (Section 4.1).

4.1 Linguistic Indicators

We encode linguistic indicators based on deep linguistic knowledge acquisition and use them to determine whether responses provide supporting clues or carry negative indications (Section 4.3). These indicators make use of linguistic features on varying levels - surface form, sentential syntax, semantics, and pragmatics - and are defined in terms of knowledge graphs (Section 4.2). We define a heuristic rule for each slot type in terms of the binary-valued linguistic indicator outputs to yield a single binary value (1 or 0) for each response. If a response's linguistic indicator value is 1, the credibility score of a response is initialized at 1.0, and 0.5 otherwise.

4.2 Knowledge Graph Construction

A semantically rich knowledge graph is constructed that links a query entity, all of its relevant slot filler nodes, and nodes for other intermediate elements excerpted from evidence sentences. There is one knowledge graph per sentence.

Fig. 2 shows a subregion of the knowledge graph built from the sentence: "***Mays***, ***50***, *died in his sleep at his* ***Tampa*** *home the morning of* ***June 28***.". It supports 3 claims: [*Mays, per: city_of_death, Tampa*], [*Mays, per: date_of_death, 06/28/2009*] and [*Mays, per: age, 50*].

Formally, a knowledge graph is an annotated graph of entity mentions, phrases and their links. It must contain one query entity node and one or more slot filler nodes. The annotation of a node includes its entity type, subtype, mention type, referent entities, and semantic category (though not every node has each type of annotation). The annotation of a link includes a dependency label and/or a semantic relation between the two linked nodes.

[2]We set $\lambda_{rs} = 0.9$, $\lambda_{sr} = 0.1$, $\lambda_{rt} = 0.3$ and $\lambda_{tr} = 0.2$, optimized from a development set. See Section 5.1.

The knowledge graph is constructed using the following procedure. First, we annotate the evidence text using dependency parsing (Marneffe et al., 2006) and Information Extraction (entity, relation and event) (Li et al., 2013; Li and Ji, 2014). Two nodes are linked if they are deemed related by one of the annotation methods (*e.g.*, [*Mays, 50*] is labeled with the dependency type *amod*, and [*home, Tampa*] is labeled with the semantic relation *located_in*). The annotation output is often in terms of syntactic heads. Thus, we extend the boundaries of entity, time, and value mentions (*e.g.*, people's titles) to include an entire phrase where possible. We then enrich each node with annotation for entity type, subtype and mention type. Entity type and subtype refer to the role played by the entity in the world, the latter being more fine-grained, whereas mention type is syntactic in nature (it may be pronoun, nominal, or proper name). For example, "*Tampa*" in Fig. 2 is annotated as a *Geopolitical (entity type) Population-Center (subtype) Name (mention type)* mention. Every time expression node is annotated with its normalized reference date (*e.g.*, "*June, 28*" in Fig. 2 is normalized as "*06/28/2009*").

Second, we perform co-reference resolution, which introduces implicit links between nodes that refer to the same entity. Thus, an entity mention that is a nominal or pronoun will often be co-referentially linked to a mention of a proper name. This is important because many queries and slot fillers are expressed only as nominal mentions or pronouns in evidence sentences, their canonical form appearing elsewhere in the document.

Finally, we address the fact that a given relation type may be expressed in a variety of ways. For example, "*the face of*" indicates the membership relation in the following sentence: "*Jennifer Dunn was the face of the Washington state Republican Party for more than two decades*." We mined a large number of trigger phrases for each slot type by mapping various knowledge bases, including Wikipedia Infoboxes, Freebase (Bollacker et al., 2008), DBPedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007), into the Gigaword corpus[3] and Wikipedia articles via distant supervision (Mintz et al., 2009)[4]. Each intermediate node in the knowledge graph that matches a trigger phrase is then assigned a corresponding semantic category. For example, "*died*" in Fig. 2 is labeled a *Death-Trigger*.

4.3 Knowledge Graph-Based Verification

We design linguistic indicators in terms of the properties of nodes and paths that are likely to be bear on the response's veracity. Formally, a *path* consists of the list of nodes and links that must be traversed along a route from a query node to a slot filler node.

Node indicators contribute information about a query entity or slot filler node in isolation, that may bear on the trustworthiness of the containing evidence sentence. For instance, a slot filler for the ***per:date_of_birth*** slot type must be a time expression.

Node Indicators

1. *Surface*: Whether the slot filler includes stop words; whether it is lower cased but appears in news. These serve as negative indicators.

2. *Entity type, subtype and mention type*: For example, the slot fillers for "*org:top_employees*" must be person names; and fillers for "*org:website*" must match the url format. Besides the entity extraction system, we also exploited the entity attributes mined by the NELL system (Carlson et al., 2010) from the KBP source corpus.

Each path contains syntactic and/or semantic relational information that may shed light on the manner in which the query entity and slot filler are related, based on dependency parser output, IE output, and trigger phrase labeling. Path indicators are used to define properties of the context in which which query-entity and slot-filler are related in an evidence sentence. For example, whether the path associated with a claim about an organization's top employee includes a title commonly associated with

[3] http://catalog.ldc.upenn.edu/LDC2011T07

[4] Under the distant supervision assumption, sentences that appear to mention both entities in a binary relation contained in the knowledge base were assumed to express that relation.

decision-making power can be roughly represented using the *trigger phrases* indicator.

Path Indicators

1. *Trigger phrases*: Whether the path includes any trigger phrases as described in Section 4.2.

2. *Relations and events*: Whether the path includes semantic relations or events indicative of the slot type. For example, a "*Start-Position*" event indicates a person becomes a "*member*" or "*employee*" of an organization.

3. *Path length*: Usually the length of the dependency path connecting a query node and a slot filler node is within a certain range for a given slot type. For example, the path for "*per:title*" is usually no longer than 1. A long dependency path between the query entity and slot filler indicates a lack of a relationship. In the following evidence sentence, which does not entail the "*per:religion*" relation between "*His*" and the religion "*Muslim*", there is a long path ("*his-poss-moment-nsubj-came-advcl-seized-militant-acmod-Muslim*"): "His *most noticeable moment in the public eye came in 1979, when* Muslim *militants in Iran seized the U.S. Embassy and took the Americans stationed there hostage.*".

Detecting and making use of interdependencies among various claims is another unique challenge in SFV. After initial response credibility scores are calculated by combining linguistic indicator values, we identify responses that have potentially conflicting or potentially supporting slot-filler candidates. For such responses, their credibility scores are changed in accordance with the binary values returned by the following indicators.

Interdependent Claims Indicators

1. *Conflicting slot fillers*: When fillers for two claims with the same query entity and slot type appear in the same evidence sentence, we apply an additional heuristic rule designed for the slot type in question. For example, the following evidence sentence indicates that compared to "*Cathleen P. Black*", "*Susan K. Reed*" is more likely to be in a "*org:top_employees/members*" relation with "*The Oprah Magazine*" due to the latter pair's shorter dependency path: "*Hearst Magazine's President* Cathleen P. Black *has appointed* Susan K. Reed *as editor-in-chief of the U.S. edition of* The Oprah Magazine.". The credibility scores are accordingly changed (or kept at) 0.5 for responses associated with the former claim, and 1.0 for those associated with the latter.

2. *Inter-dependent slot types*: Many slot types are inter-dependent, such as "*per:title*" and "*per:employee_of*", and various family slots. After determining initial credibility scores for each response, we check whether evidence exists for any implied claims. For example, given initial credibility scores of 1.0 for two responses supporting the claims that (1)"*David*" is "*per:children*" of "*Carolyn Goodman*" and (2)"*Andrew*" is "*per:sibling*" of "*David*", we check for any responses supporting the claim that (3)"*Andrew*" is "*per:children*" of "*Carolyn Goodman*", and set their credibility scores to 1.0. For example, a response supporting this claim included the evidence sentence, "*Dr.* Carolyn Goodman*, her husband, Robert, and their son,* David*, said goodbye to David's brother,* Andrew.".

5 Experimental Results

This section presents the experiment results and analysis of our approach.

5.1 Data

The data set we use is from the TAC-KBP2013 Slot Filling Validation (SFV) task, which consists of the merged responses returned by 52 runs (regarded as systems in MTM) from 18 teams submitted to the Slot Filling (SF) task. The source collection has 1,000,257 newswire documents, 999,999 web documents

Methods	Precision	Recall	F-measure	Accuracy	Mean Average Precision
1.Random	28.64%	50.48%	36.54%	50.54%	34%
2.Voting	42.16%	70.18%	52.68%	62.54%	62%
3.Linguistic Indicators	50.24%	70.69%	58.73%	72.29%	60%
4.SVM (3 + System + Source)	56.59%	48.72%	52.36%	75.86%	56%
5.MTM (3 + System + Source)	53.94%	**72.11%**	**61.72%**	**81.57%**	**70%**

Table 2: Overall Performance Comparison.

and 99,063 discussion forum posts, which results in 10 different sources (combinations of publication venues and genres) in our experiment. There are 100 queries: 50 person and 50 organization entities. After removing redundant responses within each single system run, we use 45,950 unique responses as the input to truth-finding. Linguistic Data Consortium (LDC) human annotators manually assessed all of these responses and produced 12,844 unique responses as ground truth. In order to compare with state-of-the-art supervised learning methods for SFV (Tamang and Ji, 2011; Li and Grishman, 2013), we trained a SVMs classifier [5] as a counterpart, incorporating the same set of linguistic indicators, sources and systems as features. We picked 10% (every 10th line) to compose the development set for MTM and the training set for the SVMs. The rest is used for blind test.

5.2 Overall Performance

Table 2 shows the overall performance of various truth finding methods on judging each response as true or false. MTM achieves promising results and even outperforms supervised learning approach. Table 3 presents some examples ranked at the top and the bottom based on the credibility scores produced by MTM.

	Response Ranked by MTM							
		Claim				Source		System Rank
		Query Entity	Slot Type	Slot Filler	Evidence			
Top Truths	T1	China Banking Regulatory Commission	org:top members/employees	Liu Mingkang	**Liu Mingkang**, the chairman of the China Banking Regulatory Commission	Central News Agency of Taiwan	News	15
	T2	Galleon Group	org:founded by	Raj Rajaratnam	Galleon Group, founded by billionaire **Raj Rajaratnam**	New York Times	News	9
	T3	Mike Penner	per:age	52	L.A. Times Sportswriter Mike Penner, **52**, Dies	New York Times	News	1
	T4	China Banking Regulatory Commission	org:alternate names	CBRC	...China Banking Regulatory Commission said in the notice. The five banks ... according to **CBRC**.	Xinhua,	News	5
	T5	Stuart Rose	per:origin	Briton	Bolland, 50, will replace **Briton** Stuart Rose at the start of 2010.	Agence France-Presse	News	3
Bottom False Claims	F1	American Association for the Advancement of Science	org:top members employees	Freedman	erica.html > American Library Association, President: Maurice **Freedman** < http://www.aft.org > American Federation of Teachers ...	Google	Newsgroup	4
	F2	Jade Goody	per:origin	Britain	because Jade Goody's the only person to ever I love **Britain**	Discussion Forum		3
	F3	Don Hewitt	per:spouse	Swap	...whether "Wife **Swap**" on ABC or "Jon & Kate" on TLC	New York Times	News	7
	F4	Council of Mortgage Lenders	org:website	www.cml.org.uk	he purchases in the U.K. jumped by 16 percent in April, suggesting the property market slump may have bottomed out	Associated Press World-stream	News	18
	F5	Don Hewitt	per:alternate names	Hewitt Mchen	US DoMIna THOMPson LACtaTe haVeD [3866 words]	Google	Newsgroup	13

Table 3: Top and Bottom Response Examples Ranked by MTM.

[5] We used the LIBSVM toolkit (Chang and Lin, 2011) with Gaussian radial basis function kernel.

We can see that majority voting across systems performs much better than random assessment, but its accuracy is still low. For example, the true claim *T5* was extracted by only one system because most systems mistakenly identified "*Briton Stuart Rose*" as a person name. In comparison, MTM obtained much better accuracy by also incorporating multiple dimensions of source and evidence information.

Method 3 using linguistic indicators alone, already achieved promising results. For example, many claims are judged as truths through trigger phrases (*T1* and *T5*), event extraction (*T2*), coreference (*T4*), and node type indicators (*T3*). On the other hand, many claims are correctly judged as false because their evidence sentences did not include the slot filler (*F1*, *F4*, *F5*) or valid knowledge paths to connect the query entity and the slot filler (*F2*, *F3*). The performance gain (2.99% F-score) from Method 3 to Method 5 shows the need for incorporating system and source dimensions. For example, most truths are from news while many false claims are from newsgroups and discussion forum posts (*F1*, *F2*, *F5*).

The SVMs model got very low recall because of the following two reasons: (1) It ignored the inter-dependency between multiple dimensions; (2) the negative instances are dominant in the training data, so the model is biased towards labeling responses as false.

5.3 Truth Finding Efficiency

Table 3 shows that some truths (*T1*) are produced from low-ranked systems whereas some false responses from high-ranked systems (*F1*, *F2*). Note that systems are ranked by their performance in KBP SF task. In order to find all the truths, human assessors need to go through all the responses returned by multiple systems. This process was proven very tedious and costly (Ji et al., 2010; Tamang and Ji, 2011).

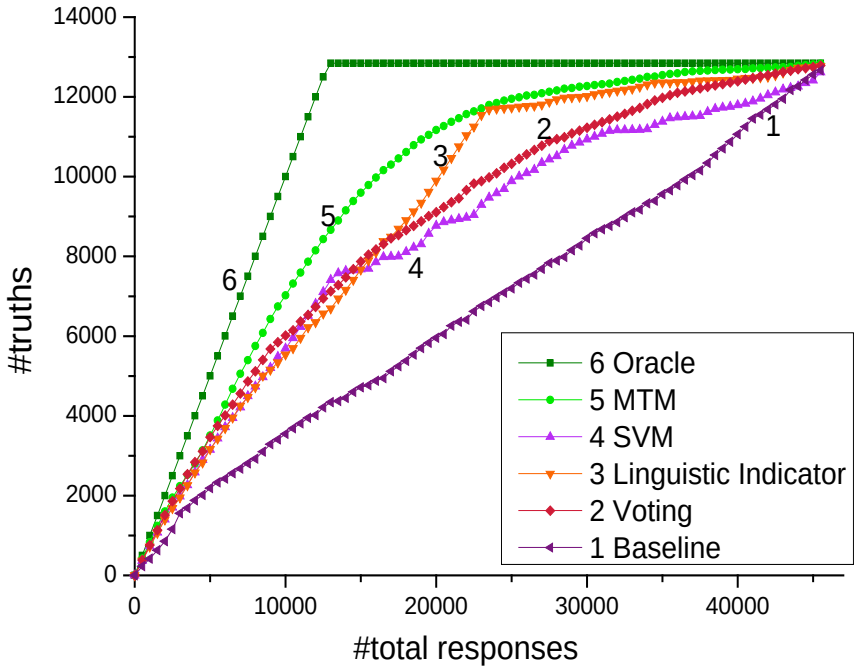

Figure 3: Truth Finding Efficiency.

Our MTM approach can expedite this process by ranking responses based on their credibility scores and asking human to assess the responses with high credibility first. Traditionally, when human assess responses, they follow an alphabetical order or system IDs in a "passive learning" style. This is set as our baseline. For comparison, we also present the results using only linguistic indicators, using voting in which the responses which get more votes across systems are assessed first, and the oracle method assessing all correct responses first. Table 2 shows our model can successfully rank trustworthy responses at high positions compared with other approaches.

Fig. 3 summarizes the results from the above 6 approaches. The common end point of all curves represents the cost and benefit of assessing all system responses. We can see that the baseline is very

inefficient at finding the truths. If we employ linguistic indicators, the process can be dramatically expedited. MTM provides further significant gains, with performance close to the Oracle. With only half the cost of the baseline, MTM can already find 90% truths.

5.4 Enhance Individual SF Systems

Finally, as a by-product, our MTM approach can also be exploited to validate the responses from each individual SF system based on their credibility scores. For fair comparison with the official KBP evaluation, we use the same ground-truth in KBP2013 and standard precision, recall and F-measure metrics as defined in (Ji et al., 2011). To increase the chance of including truths which may be particularly difficult for a system to find, LDC prepared a manual key which was assessed and included in the final ground truth. According to the SF evaluation setting, F-measure is computed based on the number of unique true claims. After removing redundancy across multiple systems, there are 1,468 unique true claims. The cutoff criteria for determining whether a response is true or not was optimized from the development set.

Fig. 4 presents the F-measure scores of the best run from each individual SF system. We can see that our MTM approach consistently improves the performance of almost all SF systems, in an absolute gain range of [-1.22%, 5.70%]. It promotes state-of-the-art SF performance from 33.51% to 35.70%. Our MTM approach provides more gains to SF systems which mainly rely on lexical or syntactic patterns than other systems using distant supervision or logic rules.

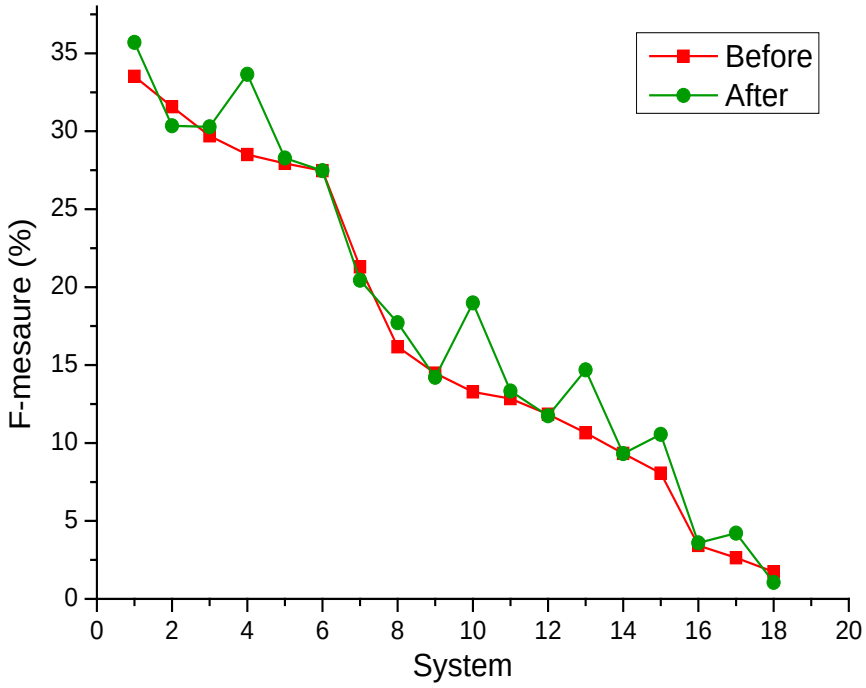

Figure 4: Impact on Individual SF Systems.

6 Conclusions and Future Work

Truth finding has received attention from both Natural Language Processing (NLP) and Data Mining communities. NLP work has mostly explored linguistic analysis of the content, while Data Mining work proposed advanced models in resolving conflict information from multiple sources. They have relative strengths and weaknesses. In this paper we leverage the strengths of these two distinct, but complementary research paradigms and propose a novel unsupervised multi-dimensional truth-finding framework incorporating signals both from multiple sources, multiple systems and multiple evidences based on knowledge graph construction with multi-layer linguistic analysis. Experiments on a challenging SFV

task demonstrated that this framework can find high-quality truths efficiently. In the future we will focus on exploring more inter-dependencies among responses such as temporal and causal relations.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA), the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193, U.S. National Science Foundation grants IIS-0953149, CNS-0931975, IIS-1017362, IIS-1320617, IIS-1354329, U.S. DARPA Award No. FA8750-13-2-0041 in the Deep Exploration and Filtering of Text (DEFT) Program, IBM Faculty Award, Google Research Award, DTRA, DHS and RPI faculty start-up grant. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proc. the 6th International Semantic Web Conference*.

L. Blanco, V. Crescenzi, P. Merialdo, and P. Papotti. 2010. Probabilistic models to reconcile complex data from inaccurate data sources. In *Proc. Int. Conf. on Advanced Information Systems Engineering (CAiSE'10)*, Hammamet, Tunisia, June.

K. Bollacker, R. Cook, and P. Tufts. 2008. Freebase: A shared database of structured general human knowledge. In *Proc. National Conference on Artificial Intelligence*.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*.

C. Chang and C. Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

H. Deng, M. R. Lyu, and I. King. 2009. A generalized co-hits algorithm and its application to bipartite graphs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 239–248, New York, NY, USA. ACM.

X. L. Dong, L. Berti-Equille, and D. Srivastavas. 2009a. Integrating conflicting data: The role of source dependence. In *Proc. 2009 Int. Conf. Very Large Data Bases (VLDB'09)*, Lyon, France, Aug.

X. L. Dong, L. Berti-Equille, and D. Srivastavas. 2009b. Truth discovery and copying detection in a dynamic world. In *Proc. 2009 Int. Conf. Very Large Data Bases (VLDB'09)*, Lyon, France, Aug.

A. Galland, S. Abiteboul, A. Marian, and P. Senellart. 2010. Corroborating information from disagreeing views. In *Proc. ACM Int. Conf. on Web Search and Data Mining (WSDM'10)*, New York, NY, Feb.

L. Ge, J. Gao, X. Yu, W. Fan, and A. Zhang. 2012. Estimating local information trustworthiness via multi-source joint matrix factorization. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 876–881. IEEE.

H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. 2010. An overview of the tac2010 knowledge base population track. In *Proc. Text Analytics Conf. (TAC'10)*, Gaithersburg, Maryland, Nov.

H. Ji, R. Grishman, and H.T. Dang. 2011. Overview of the tac 2011 knowledge base population track. In *Text Analysis Conf. (TAC) 2011*.

X. Li and R. Grishman. 2013. Confidence estimation for knowledge base population. In *Proc. Recent Advances in Natural Language Processing (RANLP)*.

Q. Li and H. Ji. 2014. Incremental joint extraction of entity mentions and relations.

Q. Li, H. Ji, and L. Huang. 2013. Joint event extraction via structured prediction with global features.

M. D. Marneffe, B. Maccartney, and C. D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC*, pages 449,454.

R. Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proc. ACL2004*.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. ACL2009*.

J. Pasternack and D. Roth. 2010. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 877–885. Association for Computational Linguistics.

J. Pasternack and D. Roth. 2011. Making better informed trust decisions with generalized fact-finding. In *Proc. 2011 Int. Joint Conf. on Artificial Intelligence (IJCAI'11)*, Barcelona, Spain, July.

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

E. Peserico and L. Pretto. 2009. Score and rank convergence of hits. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 770–771. ACM.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA. ACM Press.

S. Tamang and H. Ji. 2011. Adding smarter systems instead of human annotators: Re-ranking for slot filling system combination. In *Proc. CIKM2011 Workshop on Search & Mining Entity-Relationship data*, Glasgow, Scotland, UK, Oct.

VG Vydiswaran, C.X. Zhai, and D. Roth. 2011. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 974–982. ACM.

D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. 2012. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proc. ACM/IEEE Int. Conf. on Information Processing in Sensor Networks (IPSN'12)*, pages 233–244, Beijing, China, April.

X. Yin and W. Tan. 2011. Semi-supervised truth discovery. In *Proc. 2011 Int. World Wide Web Conf. (WWW'11)*, Hyderabad, India, March.

X. Yin, J. Han, and P. S. Yu. 2008. Truth discovery with multiple conflicting information providers on the Web. *IEEE Trans. Knowledge and Data Engineering*, 20:796–808.

B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. 2012. A Bayesian approach to discovering truth from conflicting sources for data integration. In *Proc. 2012 Int. Conf. Very Large Data Bases (VLDB'12)*, Istanbul, Turkey, Aug.