RPI-BLENDER TAC-KBP2013 Knowledge Base Population System

Dian Yu, Haibo Li, Taylor Cassidy, Qi Li, Hongzhao Huang, Zheng Chen and Heng Ji
Computer Science Department
Rensselaer Polytechnic Institute
jih@rpi.edu

Yaozhong Zhang and Dan Roth
Computer Science Department
Univerisity of Illinois at Urbana-Champaign
danr@illinois.edu

Abstract

This year the RPI-BLENDER team participated in the following four tasks: English Entity Linking, Regular Slot Filling, Temporal Slot Filling and Slot Filling Validation. The major improvement was made for Regular Slot Filling and Slot Filling validation. We developed a fresh system for both tasks. Our approach embraces detailed linguistic analysis and knowledge discovery, and advanced knowledge graph construction and truth-finding algorithms.

1 Introduction

This is the fourth year we participated in KBP evaluation. Looking back the development of various KBP tasks since 2009, Entity Linking has produced the largest impact on the Natural Language Processing (NLP) and data mining communities. Entity Linking for clean newswire data seems to be a solved problem - the best system achieved F-measure close to 85% B-cubed+ F-measure (Ji et al., 2011). In addition, we have achieved top 1 performance on temporal slot filling and top 2 on cross-lingual entity linking in previous years. In contrast, Regular Slot Filling (SF) has been the most challenging task in KBP since 2009. To the date there have been very few significant publications on this task at major venues. Our team did not make improvement on our SF system during 2011 and 2012. Therefore this year we invested our 90% efforts on developing a slot filling system from scratch. Since most of our systems for Entity Linking and Temporal Slot Filling were based on extending our previous approaches, in this paper we will focus on describing the new techniques that we have developed this year, for English Slot Filling and Slot Filling Validation (Section 3).

2 Entity Linking

Our Entity Linking system generally followed our 2012 system (Chen and Ji, 2011). We submitted seven runs for the Entity Linking task, two of which did not use KB text, and the other five used KB text, none of which used external entity KBs such as Freebase, DBPedia, and Wikilinks.

Regarding query expansion and query reformulation, we used Wikipedia redirect and disambiguation pages for query expansion, Lucene for candidate KB retrieval and document/KB retrieval, an ACE Information Extraction system (Li et al., 2012; Li et al., 2013) for entity mention extraction and coreference resolution, Lucene Spell Checker for misspelling correction in queries, acronym expansion using patterns, GPE name expansion using GPE dictionaries. We limited the maximum number of retrieved KB candidate entities to be 100.

Regarding KB ranking algorithms, we used two unsupervised learning methods including popularity based, TF-IDF based document/KB similarity, and three supervised including maximum entropy based pointwise ranking, SVM pointwise ranking and ListNet listwise ranking. The detailed descriptions of these ranking algorithms are in (Chen and Ji, 2011).

For NIL clustering, last year we did not achieve success by applying our advanced collaborative clustering methods, so this year we only applied two simple clustering algorithms: one-in-one which

assigned a cluster id for each NIL query, and all-in-one, which assigned a cluster id for all NIL queries with the same name after expansion. In addition, we enhanced our query reformulation method, especially for GPE name expansion. After such query reformulation, the query ambiguity level has been significantly reduced on the training data sets. Therefore, we hypothesize that assigning a single cluster id for all queries which share the same reformulated and expanded names will achieve reasonable performance.

In addition, we focused on improving our system for informal genres. For example, there are often many misspellings in web blogs and discussion forum posts. This year, we made specific efforts on enhancing misspelling correction (similar to many popular search engine's misspelling correction features) based on the method described in (Lalwani et al.,). This approach provided better query matching results. However, we found that the B-cubed+ F-measure degraded from 65.5% to 63.1% by adding features from Wikipedia texts due to the informality and noise in the discussion forum posts. We believe that in the future it's crucial to enhance entity linking for informal genres by advanced mis-spelling correction and text normalization techniques.

3 Slot Filling & Slot Filler Validation

In this section, we will focus on describing our novel methods for slot filling and slot filler validation.

3.1 Motivation

We call a combination of query entity, slot type, and slot filler as a *claim*. A system is given a partial claim consisting of a query entity and a slot type, and must return a complete claim which includes a slot-filler. Along with each claim, each system must provide the ID of a document and some 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.

Extracting true claims from multiple sources, though a promising line of research, raises two complications: (1) different information sources (*e.g.*, CNN vs. Twitter (Morris et al., 2012; Zubiaga and Ji, 2013)) may generate claims with varied trustability; and (2) various SF systems may generate erroneous, conflicting, redundant, complementary, ambiguously worded, or inter-dependent claims from the same set of documents because they may be built using a diverse set of algorithms on different data sets and resources.

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 providers (*i.e.* systems or humans). The "truth finding" problem has been studied in the data mining and database communities (*e.g.*, (Yin et al., 2008; Galland et al., 2010; Dong et al., 2009a; Dong et al., 2009b; Blanco et al., 2010; Pasternack and Roth, 2011; Ge et al., 2012; Zhao et al., 2012; Wang et al., 2012; Pasternack and Roth, 2010; Yin and Tan, 2011)). It is also closely related to crowdsourcing (Howe, 2006), since one critical task in crowdsourcing is to corroborate knowledge from human annotators of various levels of expertise and reliability (*e.g.*, (Smyth et al., 1995; Whitehill et al., 2009; Kasneci et al., 2011; Bachrach et al., 2012; Zhou et al., 2012)), as well as to truth propagation (Jøsang et al., 2006). Nevertheless, 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 SF systems. They tend to produce errors due to both the imperfect algorithms they employ as well as the inconsistencies of information sources, and thus pose the following new challenges.

- We require not only high-confidence claims but also trustworthy evidence to verify them. Furthermore, the evidence is expressed in unstructured natural language, therefore deep understanding is

[1] Hereafter, we refer to "pieces of evidence" as "evidences". Note that one evidence may contain multiple evidence sentences.

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***)

needed to verify claims. To the best of our knowledge, this will be the first attempt to mine and use rich knowledge from multiple lexical, syntactic and semantic levels from evidence for truth finding.

- Previous truth finding work assumed most claims are likely to be true. However, most SF systems have hit a performance of 35% F-measure. That false responses constitute the majority class invites truth finding strategies based on negative indicators.
- Most of the previous methods relied on the "wisdom of the crowd" (*i.e.*, the majority will make correct claims, most of the time). In our task, certain implicit truths might only be discovered by a minority of good systems or from a few good sources, thus the majority response is not always the most trustworthy.
- The performance of a system or source may vary over time: new SF systems may enter the network, whereas older ones may exit or their reliability may fluctuate, or the trustability of information sources may fluctuate when facing dynamic, volatile events (*e.g.*, searching for Boston Marathon bombing victims).
- Systems, sources and claims may be dependent on each other: multiple systems may share similar resources, sources may be forwarding or commenting each other; claims may be dependent on one another (*e.g.*, age vs. birth date).

Most previous slot filling work (*e.g.*, (Chen et al., 2010; Sun et al., 2011; Tamang and Ji, 2011; Surdeanu et al., 2012; Min et al., 2013; Roth and Klakow, 2013; McNamee et al., 2013; Li and Grishman, 2013)) focused on analyzing the text between the query entity and the slot filler in a single sentence, without considering global context evidence or properties of the information source itself (*e.g.*, genre). The most related work to this study is on filtering incorrect claims from multiple systems by simple heuristic rules, voting, or costly supervised learning to rank algorithms (*e.g.*, (Tamang and Ji, 2011; Li and Grishman, 2013)).

In contrast, we study credibility perceptions in richer and wider contexts. A novel minimally-supervised multi-dimensional truth finding framework is proposed. It incorporates signals from multiple sources, multiple systems, hard constraints and soft features by construction of a knowledge graph from multiple evidences using multi-layer deep linguistic analysis. Experiments demonstrate that our approach can find truths accurately (11.06% higher accuracy than supervised methods) and efficiently (find 90% truths with only one half cost of a baseline without credibility estimation). In addition, it significantly enhances the state-of-the-art SF systems independent of their algorithms and resources.

3.2 MTM: A Multi-dimensional Truth-Finding Model

For quality truth-finding, we propose a novel *multi-dimensional truth-finding model (MTM)* to incorporate and compute multi-dimensional credibility scores. Consider a set of responses $R = \{r_1, \ldots, r_m\}$ provided by a set of sources $S = \{s_1, \ldots, s_n\}$ and extracted by a set of systems $T = \{t_1, \ldots, t_l\}$. A heterogeneous network is constructed as shown in Fig. 1. Let weight matrices $W^{rs}_{m\times n} = \{w^{rs}_{ij}\}$ 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.

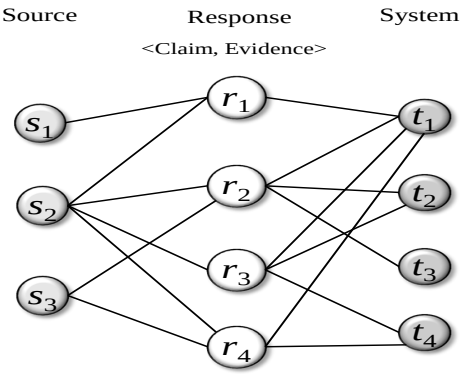

Figure 1: Multi-dimensional Truth-finding Model

The following heuristics are explored in MTM.

Heuristic 1: *A response is more likely to be true if it is provided by many trustworthy sources. A source is more likely to be trustworthy if many responses it provides are correct.*

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 it extracts are correct.*

Similar heuristics are explored in previous truth-finding studies. The major differences between our setting and theirs are (1) we evaluate responses, which are pairs of claims and evidences, rather than just claims; and (2) the fraction of false responses is much higher than that of truths. It is therefore critical to assign reliable initial credibility scores as opposed to random scores as in previous work.

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). Besides string matching, we also use a coreference resolution system(references omitted for blind review) to compute the overlap between the elements in two responses.

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 TextRank algorithm (Mihalcea, 2004) on G to obtain $c^0(t)$.

We obtained negative results by attempting to incorporate system metadata into credibility initialization, such as the algorithms and resources the system uses at each step, its previous performance in benchmark tests, and the confidence values it produces for each response. 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 progressing: thus the previous benchmark results are not reliable. Furthermore, most SF systems still lack of reliable confidence estimation.

Each source is represented as a combination of publication venue and genre, the credibility scores of sources S are initialized as a uniformed value $\frac{1}{n}$. The initialization of the credibility scores for responses relies on deep linguistic analysis on the evidence sentences and the exploitation of semantic clues, which will be described in Section 3.3.

Credibility propagation. To mutually reinforce the trustworthiness of linked objects, credibility scores are propagated across our source-response-system network iteratively. By extension of Co-HITS (Deng et al., 2009), designed for bipartite graphs, we develop a novel 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 transpose 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 step, which can be formally de-

fined 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 $c(t)$. Starting with sources, the update process considers both the initial score $c^0(s)$ and the propagation from connected responses, which can be formulated as:

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

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

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

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

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

where $\lambda_{rs}, \lambda_{rt}, \lambda_{sr}$ and $\lambda_{tr} \in [0, 1]$. They control the preference to the propagation over the initial score for every type of random walk link. The larger they are, the more we rely on link structure[2].

3.3 Response Credibility Initialization

As stated in Section 3.1, a unique challenge of our problem is to provide evidence along with a claim. Each evidence is expressed as a few natural language sentences that include the query entity and the slot filler; as well as appropriate semantic content to support the claim. We analyze the multiple evidences returned by multiple SF systems for each claim in order to initialize the credibility score for each response. There are two types of signals pertaining to response credibility: (1) *hard constraints* (section 3.3.1), which pertain to the propositional content of a evidence sentences, *i.e.*, whether a sentence declaratively indicates the truth or falsity of a claim; and (2) *soft features* (section 3.3.2), which consist of coarse-grained clues that implicitly indicate the likelihood that an evidence sentence contains any supportive information. They are then combined in an Support Vector Machines (SVMs)-based classifier to initialize the credibility scores of claims (Section 3.3.3).

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

3.3.1 Hard Constraints

We encode the following hard constraints based on deep linguistic knowledge acquisition and use them to assess responses based on supporting clues or negative indications.

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.

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*].

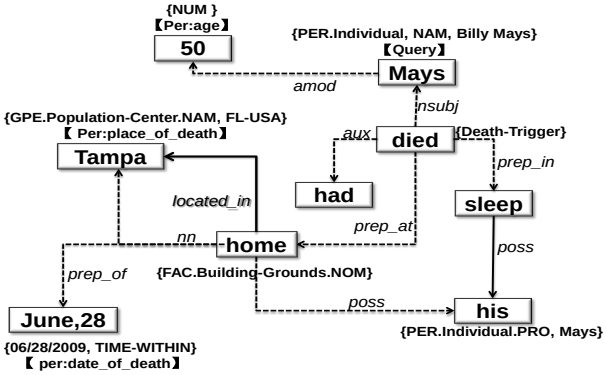

Figure 2: Knowledge Graph Example

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 all these types of annotations). The annotation of a link includes a dependency label and a semantic relation between the two linked nodes.

The knowledge graph is constructed using the following procedure. First, we annotate the evi-

dence text using dependency parsing (Marneffe et al., 2006) and Information Extraction (entity, relation and event) (references omitted for blind review). 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 IE 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.*, person 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 coreference 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 proper name. This is important because many queries and slot fillers are expressed as nominal mentions or pronouns in evidence sentences. For example, from the following sentences: "*Almost overnight, he became fabulously rich, Giuliani Partners. His consulting partners included seven of those who were with him on 9/11, and in 2002 Alan Placa, his boyhood pal, went to work at the firm.*", we will not be able to infer the relation in the knowledge graph between "*Giuliani Partners*" and "*Alan Placa*" if there is no co-referential link between "*Giuliani Partners*" and "*the firm*". Both the relation and co-referential link can be inferred. In fact, coreference resolution has played a major role in improving the recall of slot filling (Ji et al., 2011).

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[3] 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 Gigaword corpus[4] and Wikipedia articles through distant supervision (Mintz et al., 2009)[5]. In addition, we mined a disease list extracted from medical ontologies for the slot type "*per:cause_of_death*". 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 as *Death-Trigger*.

Knowledge Graph-Based Verification We design hard constraints in terms of the properties of nodes and paths that pertain to the propositional content of evidence sentences. 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. Each slot filler node possesses properties that may indicate whether it is a valid semantic type for a given slot type. For example, a claim about birth date is valid only if its slot filler is a time expression. 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. For example, whether a claim about an organization's top employee includes a title commonly associated with decision-making power. We verify both nodes and edges contained in each path as follows.

Node Constraints

1. *Surface*: Whether the slot filler includes stop words; whether it is lower cased and appears in news. These serve as negative constraints.
2. *Entity type, subtype and mention type*: For example, the slot fillers for "*org:top_employees*" must be person names; and that for "*org:website*" must match the url format.
3. *Whether the slot filler is a commenter or reporter*: This serves as a negative constraint.

[3] They will be publicly shared if the paper gets accepted.

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

[5] 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

4. *Knowledge base match*: Whether the claim exists in external knowledge bases which are manually constructed (Wikipedia Infoboxes, Freebase, DBPedia and YAGO) or automatically mined from Wikipedia articles.

Path Constraints

1. *Trigger phrases*: Whether the path includes any trigger phrases as described in section 3.3.1.
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. The following evidence does not entail that the "*per:religion*" relation holds between the person "*His*" refers to and the religion "*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.*", because the dependency path between the query entity and slot filler indicates they are syntactically distant: "*his-poss-moment-nsubj-came-advcl-seized-militant-acmod-Muslim*".
4. *Position of a particular node/edge type in the path*: For example, the dependency path for "*per:place_of_birth*" usually ends with "*prep_in*" or "*prep_at*"; the dependency path for "*per:employee_or_member_of*" often starts with "*nsubj*" and ends with "*dobj*".

The interdependency among various claims is another unique challenge in slot filling. A claim is further verified by checking whether there exists a conflicting slot filler or slot type with stronger support from the same evidence sentence.

Interdependent Claims

1. *Conflicting slot fillers*: For example, the following evidence "*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.*" 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 their shorter dependency path.
2. *Inter-dependent slot types*: Many slot types are inter-dependent, such as "*per:title*" and "*per:employee_of*", and various family slots. Consider the following sentence " *Dr. Carolyn Goodman, her husband, Robert, and their 17-year-old son, David, said goodbye to David's brother, Andrew, who was 20.*" We can all but confirm "*Andrew*" stands in the "*per:children*" relation to "*Carolyn Goodman*" because of two other claims with explicit trigger phrases: "*David*" is "*per:children*" of "*Carolyn Goodman*" and "*Andrew*" is "*per:sibling*" of "*David*".

3.3.2 Soft Features

The constraints described above capture the deep syntactic and semantic knowledge that is likely to pertain to the propositional content of claims. In addition, we can also capture some more coarse-grained shallow signals. For example, whether an evidence sentence is overall clean and informative may indicate how likely it contains trustable clues. These signals are soft features since a single signal of this type cannot guarantee high quality. They are divided into the following four categories.

1. *Cleanness*: We evaluate the cleanness of an evidence sentence by the number of special characters, the number of words, and the average length of words. The evidence sentence including a truth is likely to be clean.
2. *Informativeness*: We evaluate the informativeness of an evidence sentence by the number of capitalized words, numbers and time expressions. The evidence sentence including a truth is likely to be informative.
3. *Local Knowledge Graph*: For each slot filler node, we construct its local knowledge graph by including context nodes which are reachable by traversing at most two dependency edges. By analysing the size of the local knowledge graphs, we can have a general estimation of the high-level latent semantic role and degree of informativeness that the slot filler plays in an evidence sentence.
4. *Voting*: As discussed in Section 3.1, majority vot-

ing directly on claims may not be effective because certain implicit truths might only be discovered by a minority of good systems or from a few good sources. However, if an evidence sentence is extracted by many systems and it includes many on-topic keywords, then the sentence is more likely to include truths. Therefore we incorporate the following two additional soft features: 1) We compute the frequency of a sentence submitted by all systems as evidence; and 2) We generate a topical keyword list by choosing the top 6% most frequently used words in all evidence sentences, excluding stop words (*e.g.*, 'is' and 'and'). Then we compute the percentage of topical keywords in each evidence sentence.

3.3.3 Combining the Signals

In order to systematically incorporate all the evidence signals to initialize the response credibility scores, we employ a supervised classifier based on SVMs (Chang and Lin, 2011) with the default Gaussian radial basis function kernel instead of heuristic rules. Since the two kinds of signals are of very different nature, we use the binary classification result (true or false) based on heuristic rules combining the hard constraints as one single feature. We feed it together with the various types of soft features into the classifier. The probability scores from this classifier are used as initial credibility scores for responses in MTM as described in Section 3.2.

3.4 Experiments and Discussions

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

3.4.1 Data

The data set we use is from the TAC-KBP2013 Slot Filling Validation (SFV) track, which consists of the merged responses returned by 52 runs from 18 teams submitted to the regular slot filling track. The source collection has 1,000,257 newswire documents, 999,999 web documents 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 truths. We picked 10% (every 10th line) to compose the training set for the SVMs classifier and the development set for MTM, and used the rest for blind test.

Methods	Acuracy
1.Random	49.90%
2.Voting	62.54%
3.Hard Constraints	72.29%
4.MTM (3 + System + Source)	79.20%
5.MTM (4 + Soft Features)	**83.35**%

Table 2: Overall Accuracy.

3.4.2 Truth Finding Accuracy

Table 2 shows the accuracy of various truth finding methods on judging each response as true or false. We use the standard accuracy as the evaluation metric. Table 3 presents some examples ranked at the top and the bottom based on the credibility scores produced by MTM.

We can see that majority voting across systems performs 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, our truth finding approaches obtained dramatically better accuracy by also incorporating multiple dimensions of source and evidence information.

Method 3 using hard constraints alone, with the learning framework similar to the state-of-the-art slot filling validation (Tamang and Ji, 2011; Li and Grishman, 2013), 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 constraints (*T3*). On the other hand, many claims are successfully judged as false because the evidences do not include the slot filler (*F1*, *F4*, *F5*) or valid knowledge paths to connect the query entity and slot filler (*F2*, *F3*).

The performance gain (6.91%) from Method 3 to Method 4 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). Method 5, which integrates all dimensions, outperforms Method 4 (4.15% further gain), proving that soft features are effective at filtering many false claims with noisy evidence (*F5*).

3.4.3 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*). 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). Our MTM approach can expedite this truth finding 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 hard constraints, using voting in which the responses which get more votes across systems are assessed first, and the oracle method assessing all correct responses first.

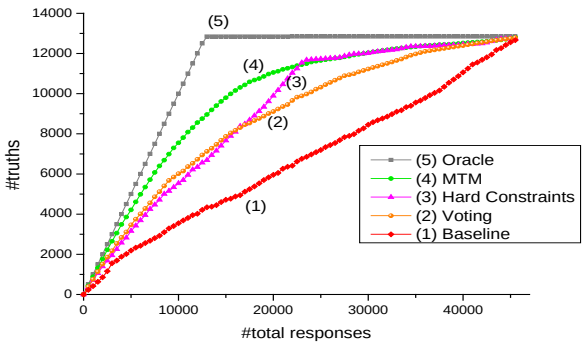

Figure 3: Truth Finding Efficiency

Fig. 3 summarizes the results from the above 5 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 hard constraints, the process can be dramatically expedited. The full MTM provides further significant gains, with performance close to the Oracle. With only half of the time of the baseline, MTM can already find 90% truths.

3.4.4 Enhance Individual SF Systems

Finally, as a by-product, our MTM approach can also be exploited to validate the responses from each individual 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.

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%.

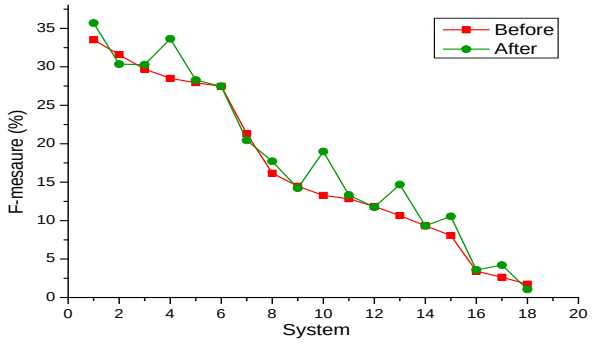

Figure 4: Impact on Individual SF Systems

Fig. 5 present the precision and recall scores as we apply various cut-offs to the claim credibility scores. It indicates that MTM can produce reliable credibility estimation in order to find the best set of truths. To conclude, our MTM approach can serve as an effective post-processing validator for SF systems, independent of the specific algorithms and resources they have adopted. In fact, our approach achieved competitive results in KBP 2013 SFV evaluation.

3.4.5 Enhance Merged SF System

Fig. 6 presents the F-measure of the merged SF system. We can see that our MTM approach con-

		Response						
		Claim			Evidence	Source		System Rank
		Query Entity	Slot Type	Slot Filler				
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	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	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	me 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 Claim Examples Ranked by MTM

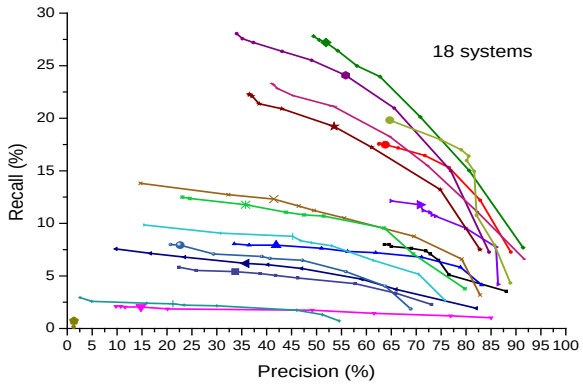

Figure 5: Individual System Performance with Credibility Thresholding (the Points with Best F Scores are Marked)

sistently improves the performance. However, we still missed 1200 correct responses using hard constraints only, and 1900 correct responses in the MTM propagation step. One reason is that some soft features are not robust. For example, we assess the cleanness of an evidence sentence by the number of words. However, in many cases, queries and fillers are not in different sentences or separated by several long clauses. In addition, we should better distinct the various granularities of hard and soft constraints and apply them to systems with different credibility scores. For example, we can apply those absolutely correct constraints such as the requiring each response of per:age to be a number in some certain range to strong systems, while apply soft constraints to weaker systems.

3.4.6 Discussion: Remaining Challenges

Despite the promising success on applying truth finding to enhance slot filling, a lot of challenges remain unsolved. The following summarize the ma-

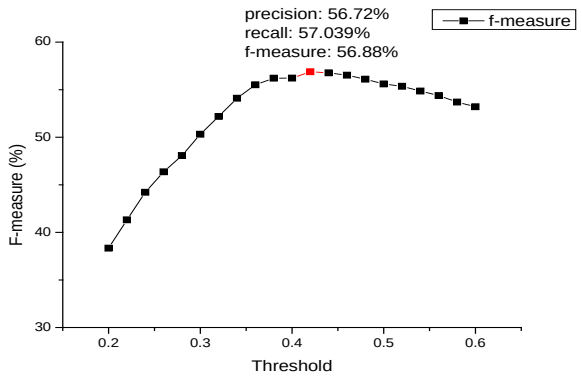

Figure 6: Performance of MTM Overall Run

jor challenges, which reflect both "classical" single-document IE bottlenecks as well as the new challenges from this new setting of truth finding.

Coreference Resolution Errors: Due to the low performance of nominal and pronoun coreference resolution, quite a few knowledge paths are missing and some nodes are replaced with wrong referent entities.

Directions of Knowledge Paths: Except the enriched annotations from relation and event extraction, most knowledge paths are un-directional and thus not sufficient to distinguish whether a slot type is "*org:parent*" or "*org:subsidiary*", "*org:members*" or "*org:member_of*", etc..

Vague Evidence: Some evidence sentences are too vague to judge, even for human annotators. For example, it's difficult to determine whether the "*per:statesorprovinces_of_residence*" of "*Dionne Warwick*" is "*N.J.*" from the following evidence: "*The list says that the state is owed $2,665,305 in personal income taxes by singer Dionne Warwick of South Orange, N.J.*"

Implicit Evidence: Some truths are implicitly expressed. For example, the following evidence "*Until last week, Palin was relatively unknown outside Alaska, and as facts have dribbled out about her, the McCain campaign has insisted that its examination of her background was thorough and that nothing that has come out about her was a surprise.*" indicates the '*per:places_of_residence*" of "*Palin*" is "*Alaska*".

4 Temporal Slot Filling

The core algorithm of our temporal slot filling (TSF) system is described in (Ji et al., 2013). We submitted five runs for the TSF task, none of which used the web during evaluation. We used Wikipedia redirects for query expansion, Lucene for document and sentence retrieval, and Stanford Core NLP toolkit for name tagging, coreference resolution, part-of-speech tagging, dependency parsing, temporal expression extraction and normalization. We used our ACE Information Extraction system (Li et al., 2012; Li et al., 2013) to extract nominal mentions, relations and events. We used Freebase to obtain training tuples for distant supervision. We performed feature elimination based on hypothesis testing and example re-labeling based on lasso regression to improve the quality and speed of training, as described in (Ji et al., 2013). Given a TSF query, we performed query expansion followed by document retrieval based on TSF query document content. Note that by using information from the entire document we aim to limit instances of the query-entity/slot-filler pair that don't express the TSF query slot-filling relation (i.e. those instances that violate the distant supervision assumption). All documents are processed as outlined above, and we then retrieve sentences containing the query-entity and slot-filler, making use of co-reference resolution, and assume retrieved sentences are instances of the TSF query's slot-filling relation. If a retrieved sentence contains a temporal expression *t*, the triple *¡query-entity, slot-filler, t¿* is assigned an intermediate label by both flat and structural SVM classifiers, where intermediate labels are selected from Beginning, Ending, Within, Beg_and_end, or None. The flat classifier uses surface lexical features and shallow dependency features, while the structural classifier makes use of full dependency paths, POS tags, and ACE relation/event information. If a sentence contains no temporal expressions the classification is performed setting *t* = document creation time.

In addition, we submitted two additional runs using temporal reasoning with event ordering constraints, based on collaboration with UIUC (Do et al., 2012). We begin with the assumption that any relations involving person entities should occur between this person's birth date and death date, and an

employment/membership relation should occur between an organization's start and dissolving dates. We combined output from our ACE Information Extraction system (Li et al., 2012; Li et al., 2013), semantic role labeling (Srikumar and Roth, 2011), and timeline creation systems (Do et al., 2012) to detect provenance for and date of death for TSF query-entities in TSF query-documents and documents considered similar to TSF query documents. If it was determined based on a document *d* that a TSF query-entity *q* died during a given interval (a, b), we add an output $t_4 = b$ for each slot-filling relation that involved *q*, citing relevant parts of *d* as provenance. Given our current architecture it is difficult for our system to learn to infer that death implies (i) the end of current fluent relations and (ii) a cutoff point for the end of previous fluent relations, so this common-sense assumption was hard-coded. Consider the classification instance $< Hariri, PrimeMinister,$ 2005-02-14 $>$. Given the sentence, *"The UN special tribunal was established on March 1, 2009 to try suspects that planned, facilitated and executed the assassination of former Lebanese Prime Minister Rafiq Hariri, who was killed along with 22 others in a suicide bombing in Beirut on Feb. 14, 2005"*, systems unaware of the entity existence-based constraints described above might return Within, whereas After_End or End would yield a better 4-tuple. Our baseline system labels the classification instance as NONE, indicating no relationship between per:title(Hariri, Prime Minister) and 2005-02-14. The final evaluation results showed that 42 changes were improvements while only 4 were harmful. Overall this approach provided 0.9% gain in F-measure.

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*.

Y. Bachrach, T. Minka, J. Guiver, and T. Graepel. 2012. How to grade a test without knowing the answers – a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. Int. Conf. on Machine Learning (ICML'12)*, Edinburgh, Scotland, June.

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*.

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

Z. Chen and H. Ji. 2011. Collaborative ranking: a case study on entity linking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 771–781. Association for Computational Linguistics.

Z. Chen, S. Tamang, A. Lee, X. Li, W. Lin, J. Artiles, M. Snover, M. Passantino, and H. Ji. 2010. Cuny-blender tac-kbp2010 entity linking and slot filling system description. In *Proc. Text Analytics Conf. (TAC'10)*, Gaithersburg, Maryland, Nov.

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.

Q. X. Do, W. Lu, and D. Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 677–687. Association for Computational Linguistics.

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.

J. Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

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*.

H. Ji, T. Cassidy, Q. Li, and S. Tamang. 2013. Tackling Representation, Annotation and Classification Challenges for Temporal Knowledge Base Population. In *Journal of Knowledge and Information Systems*.

A. Jøsang, S. Marsh, and S. Pope. 2006. Exploring different types of trust propagation. In *Trust management*, pages 179–192. Springer.

G. Kasneci, J. V. Gael, D. H. Stern, and T. Graepel. 2011. Cobayes: Bayesian knowledge corroboration with assessors of unknown areas of expertise. In *Proc. ACM Int. Conf. on Web Search and Data Mining (WSDM'11)*, Hong Kong, Feb.

M. Lalwani, N. Bagmar, and S. Parikh. Efficient algorithm for auto correction using n-gram indexing.

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

Q. Li, H. Li, H. Ji, W. Wang, J. Zheng, and F. Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proc. 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*, Maui, Hawaii, Oct.

Q. Li, H. Ji, and L. Huang. 2013. Joint event extraction via structured prediction with global features. In *ACL*, pages 73–82.

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

P. McNamee, J. Mayfield, T. Finin, T. Oates, D. Lawrie, T. Xu, and D. W. Oard. 2013. Kelvin: a tool for automated knowledge base construction. In *Proc. NAACL2013*.

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

B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proc. NAACL2013*.

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

M. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. 2012. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proc. Computer Supported Cooperative Work (CSCW'12)*, Seattle, Washington, Feb.

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.

B. Roth and D. Klakow. 2013. Combining generative and discriminative model scores for distant supervision. In *Proc. EMNLP2013*.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. 1995. Inferring ground truth from subjective labelling of venus images. In *Proc. Neural Information Processing Systems*, pages 1085–1092, Denver, CO, Nov.

V. Srikumar and D. Roth. 2011. A joint model for extended semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 129–139. Association for Computational Linguistics.

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

A. Sun, R. Grishman, B. Min, and W. Xu. 2011. Nyu 2011 system for kbp slot filling. In *Proc. Text Analysis Conference (TAC2011)*.

M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proc. EMNLP2012*.

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.

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.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. Whose vote should count more: Optimal integration of labelers of unknown expertise. In *Proc. of Neural Information Processing Systems*, pages 2035–2043, Vancouver, B.C., Canada, Dec.

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.

D. Zhou, J. C. Platt, S. Basu, and Y. Mao. 2012. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems 25*, Lake Tahoe, Nevada, Dec.

A. Zubiaga and H. Ji. 2013. Tweet, but verify: Epistemic study of information verification on twitter. *Social Network Analysis and Mining*.