A Survey of Truth Discovery in Information Extraction

Xuan Wang, Yu Zhang and Yinyin Chen
04/26/2018

University of Illinois at Urbana-Champaign
Outline

Truth Discovery

Information Extraction

Truth Finding in Information Extraction
Truth Discovery
Iterative Methods

**TruthFinder** [Yin et al., 2008]

- Confidence of facts $\Leftrightarrow$ Trustworthiness of web sites

**Investment** [Pasternack and Roth, 2010]

- Sources “invest” their reliabilities among claims, so that each candidate value $v$ receives the votes:

\[
\text{vote}(v) = G \left( \sum_{s \in S_v} \frac{w_s}{|V_s|} \right)
\]

- Sources collects their credit back from the identified truths:

\[
w_s = \sum_{f \in V_s} \left( \text{vote}(v) \frac{w_s/|V_s|}{\sum_{s' \in S_v} w_{s'}/|V_{s'}|} \right).
\]
Graphical Model Based Methods

**Latent Truth Model (LTM)** [Wang et al., 2012]
- Categorical truths
- Inference: Collapsed Gibbs sampling

**Gaussian Truth Model (GTM)** [Zhao and Han, 2012]
- Numerical truths
- Inference: EM algorithm, Gibbs sampling
Comparison

**Difference:**
- Iterative methods are *simple and straightforward* in formulation, implementation, and interpretation.
- Graphical model based methods, as Bayesian methods, are able to incorporate *prior knowledge* and estimate the *statistical distributions*.

**Similar** internal mechanism, and a general procedure is summarized:

1. Start with initial source weights.
2. Repeatedly estimate the truth and source weights following the general principles:
   - sources with higher weights contribute more in estimating the truth;
   - sources providing more truth receive higher weights.
3. Stop until certain criterion is achieved.
Advanced Methods

- **Multi-source sensing model (MSS)** [Qi et al., 2013]
  - dealt with source dependency.
- **Regular EM** [Wang et al., 2012]
  - focused on the application of social sensing.
- **CopyCEF** [Dong et al., 2009]
  - for dynamic source discovery and copying relationship detecting.
- **Semi-Supervised Truth Discovery (SSTF)** [Yin and Tan, 2011]
  - is a semi-supervised truth finding.
Information Extraction
Information Extraction

Main Purpose

• Extract tuples in the form of \((\text{head entity}, \text{relation}, \text{tail entity})\), or \((e_h, r, e_t)\).

Sub-tasks

• Named Entity Recognition (NER)
• Relation Extraction (RE)
• Open-Domain Information Extraction (OpenIE)
Named Entity Recognition

Goal

- Discover token spans of certain types from a given corpus

Methods

- Fully-supervised
  - BiLSTM-CRF [Lample et al., 2016, Ma and Hovy, 2016]
- Weakly-supervised
  - label propagation [Talukdar and Pereira, 2010]
- Distantly-supervised
  - ClusType [Ren et al., 2015]
  - link entity mentions with a knowledge base and infer the types of unlinkable mentions
Relation Extraction

Goal

- Predict the relation between two detected entities

Methods

- Pattern-based
  - PATTY: tokens along the dep path [Nakashole et al., 2013]
- Distributional
  - TransE: entity and relation embeddings [Bordes et al., 2013]
- Hybrid
  - REPEL: co-training of two modules [Qu et al., 2018]
Open-Domain IE

Goal

- Directly extract tuples using linguistic features or sentence structures, without given types

Methods

- Pattern-based
  - MetaPAD: context-aware segmentation and synonymous pattern grouping [Jiang et al., 2017]

- Clause-based
  - ClausIE: clause type analysis based on dependency parse, chunks, and POS tags [Del Corro and Gemulla, 2013]

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Clause type</th>
<th>Example</th>
<th>Derived clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$:</td>
<td>SV</td>
<td>AE died.</td>
<td>(AE, died)</td>
</tr>
<tr>
<td>$S_2$:</td>
<td>SV$_c$A</td>
<td>AE remained in Princeton.</td>
<td>(AE, remained in Princeton)</td>
</tr>
<tr>
<td>$S_3$:</td>
<td>SV$_c$C</td>
<td>AE is smart.</td>
<td>(AE, is, smart)</td>
</tr>
<tr>
<td>$S_4$:</td>
<td>SV$_m$O</td>
<td>AE has won the Nobel Prize.</td>
<td>(AE, has won, the Nobel Prize)</td>
</tr>
<tr>
<td>$S_5$:</td>
<td>SV$_{dt}$O$_i$O</td>
<td>RSAS gave AE the Nobel Prize.</td>
<td>(RSAS, gave, AE, the Nobel Prize)</td>
</tr>
<tr>
<td>$S_6$:</td>
<td>SV$_{ct}$OA</td>
<td>The doorman showed AE to his office.</td>
<td>(The doorman, showed, AE, to his office)</td>
</tr>
<tr>
<td>$S_7$:</td>
<td>SV$_{ct}$OC</td>
<td>AE declared the meeting open.</td>
<td>(AE, declared, the meeting, open)</td>
</tr>
</tbody>
</table>
Truth Finding in Information Extraction
Truth Finding in Information Extraction

Data Source

- Unstructured data v.s. structured data
- Text data is more noisy.
- Text data brings more information, e.g., uncertainty, evidence.

Conflict Information

- Text corpora
- Information extractor
Truth Finding in Information Extraction

**Attention Mechanism** [Luo et al., 2017, Lin et al., 2016]

- Give a larger weight to specific part of the input in computing the final output.

**Heterogeneous Supervision** [Liu et al., 2017]

- Infer the truth label from heterogeneous supervision (e.g., knowledge base and domain knowledge).

**Multi-dimensional Truth Discovery** [Yu et al., 2014]

- Incorporate signals from multiple sources, multiple systems and multiple pieces of evidence.

**Truth Existence in Truth Discovery** [Zhi et al., 2015]

- True answers are excluded from the candidate answers provided by all sources.
Slot Filling Validation (SFV) task [Yu et al., 2014]

<table>
<thead>
<tr>
<th>System</th>
<th>Source</th>
<th>Slot Filler</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Agence France-Presse, News</td>
<td>Los Angeles</td>
<td>The statement was confirmed by publicist Maureen O’Connor, who said <em>Dio</em> died in <em>Los Angeles</em>.</td>
</tr>
<tr>
<td>B</td>
<td>New York Times, News</td>
<td>Los Angeles</td>
<td><em>Ronnie James Dio</em>, 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 <em>Los Angeles</em>.</td>
</tr>
<tr>
<td>C</td>
<td>Discussion Forum</td>
<td>Atlantic City</td>
<td><em>Dio</em> revealed last summer that he was suffering from stomach cancer shortly after wrapping up a tour in <em>Atlantic City</em>.</td>
</tr>
<tr>
<td>D</td>
<td>Associated Press Worldstream, News</td>
<td>Los Angeles</td>
<td><em>LOS ANGELES</em> 2010-05-16 20:31:18 UTC <em>Ronnie James Dio</em>, 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.</td>
</tr>
</tbody>
</table>

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

Multi-dimensional truth-finding model (MTM)
[Yu et al., 2014]

- Construct a heterogeneous information network for source, system and response.
- Initialize credibility score of the sources, system and response.
- Propagate the credibility scores through the heterogeneous network until converge, i.e. the change of response credibility is below a minimum threshold.

Figure 1: Heterogeneous networks for MTM.
Thank you for your attention!


**Metapad: Meta pattern discovery from massive text corpora.**
In *KDD’17*, pages 877–886. ACM.

**Neural architectures for named entity recognition.**

**An attention-based bilstm-crf approach to document-level chemical named entity recognition.**  
*Bioinformatics*, 1:8.

**End-to-end sequence labeling via bi-directional lstm-cnns-crf.**  


Clustype: Effective entity recognition and typing by relation phrase-based clustering.
In KDD’15, pages 995–1004. ACM.

Experiments in graph-based semi-supervised learning methods for class-instance acquisition.


Zhao, B. and Han, J. (2012).  
**A probabilistic model for estimating real-valued truth from conflicting sources.**  
*QDB’12.*

Zhi, S., Zhao, B., Tong, W., Gao, J., Yu, D., Ji, H., and Han, J. (2015).  
**Modeling truth existence in truth discovery.**  
In *KDD’15*, pages 1543–1552. ACM.