Managing Uncertainty in KBs with Probabilistic Databases

Andrew McCallum

Department of Computer Science
University of Massachusetts Amherst

TAC KBP
November 15, 2011
Actionable Knowledge from Text

Text docs → Entity Extraction → Relation Extraction → Resolution (Coref) → KB

query → answer
Actionable Knowledge from Text
Actionable Knowledge from Text

Text docs

Entity Extraction

Entity Mentions
Wei Li
W. Li
Xinghua U.

Relation Extraction

Relation Mentions
Attends(Wei Li, Xinghua U.)

Resolution (Coref)

KB

query

answer
Actionable Knowledge from Text

Text docs → Entity Extraction → Relation Extraction → Resolution (Coref) → Structured Data → KB

Entity Mentions: Wei Li, W. Li, Xinghua U.
Relation Mentions: Attends(Wei Li, Xinghua U.)

Query → answer
Actionable Knowledge from Text

Entity Extraction

Relation Extraction

Resolution (Coref)

Structured Data

Text docs

Entity Mentions

Relation Mentions

Attends(Wei Li, Xinghua U.)

Wei Li
W. Li
Xinghua U.

Entitlements, Relations

KB

query

answer
Actionable Knowledge from Text

Text docs → Entity Extraction → Relation Extraction → Resolution (Coref) → Structured Data → KB

Entity Mentions: Wei Li, W. Li, Xinghua U.
Relation Mentions: Attends(Wei Li, Xinghua U.)

“truth”

answer
Actionable Knowledge from Text

Text docs → Entity Extraction (Wei Li, W. Li, Xinghua U.) → Relation Extraction (Attends(Wei Li, Xinghua U.)) → Resolution (Coref) → Structured Data → KB

90% 90% 90%
Actionable Knowledge from Text

Entity Extraction

Entity Mentions

Relation Extraction

Relation Mentions

Resolution (Coref)

Structured Data

KB

query

answer

Text docs

Text docs

90% × 90% × 90% = 72%
Information Extraction components aren’t perfect. Errors snowball.
Actionable Knowledge from Text

Text docs → Entity Extraction → Relation Extraction → Resolution (Coref) → KB

Entity Mentions → Relation Mentions → "truth" → KB

Structured Data
Actionable Knowledge from Text
Actionable Knowledge from Text

Joint inference shown to increase accuracy.
Actionable Knowledge from Text

Joint inference shown to increase accuracy.

1. How to represent & inject uncertainty from IE into DB?
2. Want to use DB contents to aid IE.
3. IE isn’t “one-shot.” Add new data later; redo inference. Want DB infrastructure to manage IE.
Actionable Knowledge from Text
Actionable Knowledge from Text

Text docs → Entity Extraction → Relation Extraction → Resolution (Coref) → Structured Data

p(Entity Mentions) → p(Relation Mentions) → p(Entities, Relations) → KB

Evidence

query

answer

p("truth")
Actionable Knowledge from Text

Text docs  \rightarrow \text{Entity Extraction} \rightarrow \text{Relation Extraction} \rightarrow \text{Resolution (Coref)} \rightarrow \text{Structured Data} \rightarrow \text{KB} \rightarrow \text{answer}

- p(Entity Mentions)
- p(Relation Mentions)
- p(Entities, Relations)
- p("truth")

Evidence flow:
- Text docs
- Structured Data
- KB

Query:
- query

Answer:
- answer
Actionable Knowledge from Text

“Truth is inferred, not observed.”
Epistemological Philosophy
“Truth is inferred, not observed.”

Epistemological Philosophy
Actionable Knowledge from Text

Human Edits:

✖ Traditional: Change DB record of truth
✔ Mini-document **“Nov 15: Alan said this was true”**
Actionable Knowledge from Text

Human Edits:
- ✗ Traditional: Change DB record of truth
- ✔ Mini-document “Nov 15: Alan said this was true”

Sometimes humans are wrong, disagree, out-of-date. Jointly reason about truth and editors’ reliability/reputation.
Actionable Knowledge from Text

- Entity Extraction (p(Entity Mentions))
- Relation Extraction (p(Relation Mentions))
- Resolution (Coref) (p(Entities, Relations))

Structured Data

- Query
- KB
- Answer

Evidence

Text docs

Human Edits
Actionable Knowledge from Text

Text docs → evidence

Human Edits → evidence

Structured Data → evidence

query

KB

answer

Text docs → Entity Extraction → p(Entity Mentions) → Relation Extraction → p(Relation Mentions) → Resolution (Coref) → p(Entities, Relations) → p("truth") → KB

p(Entities, Relations)
Actionable Knowledge from Text

Managing Evidence and Uncertainty
Bayesian Network

Directed, Generative Graphical Models

\[ p(a, b, c, d, e) = p(a) \cdot p(b|a) \cdot p(c|a) \cdot p(d|a, c) \cdot p(e|b, c) \]
Conditional Random Field (CRF)

Undirected graphical model, conditioned on some data variables

[Lafferty, McCallum, Pereira 2001]
Conditional Random Field (CRF)

Undirected graphical model, conditioned on some data variables

\[ p(y|x) = \frac{1}{Z_x} \prod_{f} \phi(x \in f, y \in f) \]
Conditional Random Field (CRF)

Undirected graphical model, conditioned on some data variables

\[
p(y|x) = \frac{1}{Z_x} \prod_{f} \phi(x \in f, y \in f)
\]

[Lafferty, McCallum, Pereira 2001]
Conditional Random Field (CRF)

Undirected graphical model, conditioned on some data variables

\[ p(y|x) = \frac{1}{Z_x} \prod_{f} \phi(x \in f, y \in f) \]

+ Tremendous freedom to use arbitrary features of input.
+ Predict multiple dependent variables (“structured output”)

[Lafferty, McCallum, Pereira 2001]
Information Extraction with Linear-chain CRFs

Logistic Regression analogue of a hidden Markov model

Graphical model

Finite state model

Today Morgan Stanley Inc announced Mr. Friday’s appointment.
### Information Integration

#### Database A (Schema A)

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>J.</td>
<td>Smith</td>
<td>222-444-1337</td>
</tr>
<tr>
<td>J.</td>
<td>Smith</td>
<td>444 1337</td>
</tr>
<tr>
<td>John</td>
<td>Smith</td>
<td>(1) 4321115555</td>
</tr>
</tbody>
</table>

#### Database B (Schema B)

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>U.S. 222-444-1337</td>
</tr>
<tr>
<td>John D. Smith</td>
<td>444 1337</td>
</tr>
<tr>
<td>J Smiht</td>
<td>432-111-5555</td>
</tr>
</tbody>
</table>

---

#### Schema Matching

<table>
<thead>
<tr>
<th>Schema A</th>
<th>Schema B</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Name</td>
<td>Name</td>
</tr>
<tr>
<td>Last Name</td>
<td>Phone</td>
</tr>
<tr>
<td>Contact</td>
<td></td>
</tr>
</tbody>
</table>

#### Coreference

- **John #1**
  - J. Smith
  - J. Smith
  - John Smith
  - John D. Smith

- **John #2**
  - John Smith
  - J Smiht

#### Canonicalization: Normalized DB

<table>
<thead>
<tr>
<th>Entity#</th>
<th>Name</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>523</td>
<td>John Smith</td>
<td>222-444-1337</td>
</tr>
<tr>
<td>524</td>
<td>John D. Smith</td>
<td>432-111-5555</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Schema Matching

\[ P(Y \mid X) = \frac{1}{Z_X} \prod_{y_i \in Y} \psi_v(y_i, x_i) \prod_{y_i, y_j \in Y} \psi_b(y_{ij}, x_{ij}) \]

\[ \psi(y_i, x_i) = \exp \left( \sum_k \lambda_k f_k(y_i, x_i) \right) \]
Schema Matching

\[ P(Y \mid X) = \frac{1}{Z_X} \prod_{y_i \in Y} \psi_w(y_i, x_i) \prod_{y_i, y_j \in Y} \psi_b(y_{ij}, x_{ij}) \]

\[ \psi(y_i, x_i) = \exp \left( \sum_k \lambda_k f_k(y_i, x_i) \right) \]

- \( x_6 \) is a set of attributes \{phone, contact, telephone\}
- \( x_7 \) is a set of attributes \{last name, last name\}
- \( f_{67} \) is a factor between \( x_6 \)/\( x_7 \)
- \( y_{67} \) is a binary variable indicating a match (no)
- \( f_7 \) is a factor over cluster \( x_7 \)
- \( y_7 \) is a binary variable indicating match (yes)
Coreference and Canonicalization
• $x_1$ is a set of mentions \{J. Smith, John, John Smith\}
• $x_2$ is a set of mentions \{Amanda, A. Jones\}
• $f_{12}$ is a factor between $x_1/x_2$
• $y_{12}$ is a binary variable indicating a match (no)
• $f_1$ is a factor over cluster $x_1$
• $y_1$ is a binary variable indicating match (yes)
• Entity/attribute factors omitted for clarity
Schema Matching

\[ P(Y \mid X) = \frac{1}{Z_X} \prod_{y_i \in Y} \psi_{y_i}(y_i, x_i) \prod_{y_i, y_j \in Y} \psi_{y_i, y_j}(y_i, x_i) \]

\[ \psi_{y_i}(y_i, x_i) = \exp \left( \sum_k \lambda_k f_k(y_i, x_i) \right) \]
\[
P(Y | X) = \frac{1}{Z_X} \prod_{y \in Y} \psi(y_i, x_i) \prod_{y_i, y_j \in Y} \psi(y_{ij}, x_{ij})
\]

\[
\psi(y_i, x_i) = \exp \left( \sum_k \lambda_k f_k(y_i, x_i) \right)
\]
Prob-DBs & IE in practice
Information Extraction into DB

Documents
Information Extraction into DB

Documents

Extraction & Matching

Database
Information Extraction into DB

Documents → Extraction & Matching → DB
Information Extraction into DB
Information Extraction into DB

Documents → Extraction & Matching → DB
Information Extraction into DB

Documents

Extraction & Matching

query
Information Extraction into DB

Documents

Extraction & Matching

query

query proc

answer
Information Extraction into Pr DB
Information Extraction into Pr DB

Documents → Extraction & Matching → Pr DB
Information Extraction into Pr DB
Information Extraction into Pr DB
Information Extraction into Pr DB
Information Extraction into Pr DB

Documents → Extraction & Matching → Database
Information Extraction into Pr DB

Documents → Extraction & Matching → Pr DB

query
Information Extraction into Pr DB

Documents

Extraction & Matching

query

proc

answer
Information Extraction into Pr DB

Documents

Extraction & Matching

query

answer

query proc
Information Extraction into Pr DB
The MCMC Alternative

Documents
Information Extraction into Pr DB
The MCMC Alternative

Documents
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & Matching
Information Extraction into Pr DB

The MCMC Alternative

Documents

Extraction & Matching

DB contains only one possible world at a time.
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & Matching
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & Matching
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & Matching

MH inference
Information Extraction into Pr DB

The MCMC Alternative

Documents

Extraction & Matching

MH inference
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & Matching

MH inference
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & Matching

query

MH inference
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & query

MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & Matching

query

MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Extraction & Matching

Documents

MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & query
Matching

MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & query

MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & query

MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & query
 MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Extraction & Matching

Documents

query

MH inference

SQL answer
Information Extraction into Pr DB
The MCMC Alternative

Documents

Extraction & query
Matching

MH inference

SQL answer

answer
Information Extraction into Pr DB

The MCMC Alternative

[Wick, McCallum, Miklau VLDB 2010]
Particle Filtering

Documents

Extraction & query
Matching

answer
Particle Filtering with compact representation

[Schultz, McCallum, Miklau 2010]

Documents

Extraction & query
Matching

answer
Current Experimental Work

- Prob Query Convergence
- Query-specific Sampling
- Database-Text Alignment
- Scalable Coref with Distributed Parallelism
- Probabilistic programming support with FACTORIE
Experiments (Queries)

Query 1
SELECT string
FROM token
WHERE label = "B-PER"

Query 2
SELECT count(*)
FROM token
WHERE label = "B-PER"

Query 3
SELECT T.doc_id
FROM token T
WHERE (SELECT COUNT(*)
    FROM token T1
    WHERE T1.label='B-PER' AND T.doc_id=T1.doc_id)
=(SELECT COUNT(*)
    FROM token T1
    WHERE T1.label="B-ORG" AND T.doc_id=T1.doc_id

Not even possible in many other prDBs
Query Evaluation Loss Over Time (Q1)

Query Evaluation Loss over Time (Query 1)

Legend
- △ Materialized Sampler
- ○ Naive Sampler

Marginal distribution squared error

10-1000x faster

With maintenance trick
Current Experimental Work

- Prob Query Convergence
- Query-specific Sampling
- Database-Text Alignment
- Scalable Coref with Distributed Parallelism
- Probabilistic programming support with FACTORIE
Query-specific Sampling

Given a query,
no need to run sampling on whole universe of data.

Want \textit{query-specific sampling}.

Amazingly: under-studied problem.
**Query-specific Sampling**

- Define *total variation score & influence trail score*.
- Prove

  **Proposition 1.** If \( p(i) = 1(i \neq l) \frac{1}{n-1} \) induces an MH kernel that neglects variable \( x_l \), then the expected total variation error \( \xi_{tv} \) of the resulting MH sampling procedure under the model is the total variation influence \( tv \).

![Diagram showing query-only, uniform, and our query-specific distributions]

On larger graphs effect will be significantly more dramatic.
Current Experimental Work

- Prob Query Convergence
- Query-specific Sampling
- Database-Text Alignment
- Scalable Coref with Distributed Parallelism
- Probabilistic programming support with FACTORIE
Structured ↔ Unstructured Alignment
Structured $\leftrightarrow$ Unstructured Alignment

- founded(Bill Gates, Microsoft)
- nationality(Steve Jobs, USA)
- ...

Freebase
Structured ↔ Unstructured Alignment

founded(Bill Gates, Microsoft)

nationality(Steve Jobs, USA)

...(...,..)

founded(D. L. Sifry, Technorati)?
Structured↔Unstructured Alignment

founded(Bill Gates, Microsoft)
nationality(Steve Jobs, USA)
(...,...)

Freebase

Microsoft was founded by Bill Gates

With Microsoft chairman Bill Gates soon relinquishing ...

D. L. Sifry, a founder of Technorati

founded(D. L. Sifry, Technorati)?
Structured↔Unstructured Alignment

Freebase

founded(Bill Gates, Microsoft)
nationality(Steve Jobs, USA)

Microsoft was founded by Bill Gates
With Microsoft chairman Bill Gates soon relinquishing ...

D. L. Sifry, a founder of Technorati

founded(D. L. Sifry, Technorati)?
Structured ↔ Unstructured Alignment

**Freebase**

- founded(Bill Gates, Microsoft)
- nationality(Steve Jobs, USA)

**Microsoft** was founded by Bill Gates

- With Microsoft chairman Bill Gates soon relinquishing ...

**D. L. Sifry**, a founder of Technorati

- founded(D. L. Sifry, Technorati)?

Joint model of entities & relations
Structured ↔ Unstructured Alignment

founded(Bill Gates, Microsoft)
nationality(Steve Jobs, USA)
...

Freebase

Johnt model of entities & relations

Microsoft was founded by Bill Gates
With Microsoft chairman Bill Gates soon relinquishing ...

D. L. Sifry, a founder of Technorati

founded(D. L. Sifry, Technorati)?
Structured ↔ Unstructured Alignment

Freebase

- founded(Bill Gates, Microsoft)
- nationality(Steve Jobs, USA)
- (....,...)
- founded(D.L. Sifry, Technorati)

Joint model of entities & relations

- Microsoft was founded by Bill Gates
- With Microsoft chairman Bill Gates soon relinquishing ...
- D. L. Sifry, a founder of Technorati
- founded(D. L. Sifry, Technorati)?
Model of Entities and Relations

- $T_{Joint}$
- $Y_{Microsoft}$
- $Y_{Bill Gates}$
- $Y_{USA}$
- $Y_{Bill Gates, Microsoft}$
- $Y_{Bill Gates, USA}$
- $T_{Bias}$
- $T_{Mention}$
- $X^1_{Bill Gates, Microsoft}$
- $X^2_{Bill Gates, Microsoft}$
- $X^1_{Bill Gates, USA}$

In order to extract relations from text, we need to introduce the joint template. This often results from patterns that lead us to extract that $R_i$. For this purpose, we define the conditional distribution $p(Y_i | \theta, \mathcal{D})$.

One feature of the graphical model we use FACTORIE is that it creates one factor per variable. For example, the sequence $r_1$ director of $r$ appears to be a candidate tuple. The sequence $r_2$ is a candidate tuple the sequence $r_1$ director of $r$ appears to be 2.

Joint Inference

Microsoft was founded by Bill Gates...

With Microsoft chairman Bill Gates soon relinquishing...

Bill Gates was born in the USA in 1955.

Cross-document
Relation Extraction Experiments

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Test Set</th>
<th>Precision @50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 years</td>
<td>1 year</td>
<td></td>
</tr>
<tr>
<td>334K entities, 10 types</td>
<td>~100k articles</td>
<td>Isolated: 0.78, Pipeline: 0.82, Joint: 0.94</td>
</tr>
<tr>
<td>488K relation instances, 54 types</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Relation Extraction Experiments

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Test Set</th>
<th>Precision @50</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 years</td>
<td>1 year</td>
<td></td>
</tr>
<tr>
<td>334K entities, 10 types</td>
<td>~100k articles</td>
<td></td>
</tr>
<tr>
<td>488K relation instances, 54 types</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Isolated</th>
<th>Pipeline</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.78</td>
<td>0.82</td>
<td>0.94</td>
</tr>
</tbody>
</table>

This work was done on top of a Probabilistic Database
Current Experimental Work

• Prob Query Convergence
• Query-specific Sampling
• Database-Text Alignment

• Scalable Coref with Distributed Parallelism
• Probabilistic programming support with FACTORIE
Entity Resolution

Entity resolution by CRF with pairwise factors

These two proposals can be evaluated (and accepted) in parallel!
Entity Resolution in Parallel by Map-Reduce

[Singh, S, P, McCallum, ACL, 2011]
Parallelism = faster

Accuracy versus Time

- $B^3$ F1
- Pairwise F1

Accuracy

Wallclock Running Time (ms)

Values: 1, 2, 5, 10, 50
Hierarchical Model

Super-entities infer good “data distribution”

Sub-entities infer good “block moves”

Sampling: Fix variables of two levels, sample the remaining level
Smart Parallelism = much faster

Accuracy versus Time

Accuracy

Wallclock Running Time (ms)

pairwise
super-entities
sub-entities
combined

$B^3$ F1
Pairwise F1
combined

0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

0.0
0.5
1.0
1.5
2.0
2.5
3.0

1e7
Parallelism for Probabilistic Databases

Central Data Repository
Current state of the variables

Inference Worker

Inference Worker

Inference Worker
Parallelism for Probabilistic Databases

Inference Worker

Central Data Repository

Current state of the variables

Inference Worker

Inference Worker

Inference Worker

How to evaluate proposals correctly and resolve conflicts?
Current Experimental Work

• Prob Query Convergence
• Query-specific Sampling
• Database-Text Alignment
• Scalable Coref with Distributed Parallelism

• Probabilistic programming support with FACTORIE
Probabilistic Programming Languages

• Make it easy to represent rich, complex models, using the full power of programming languages
  - data structures
  - control mechanisms
  - abstraction

Provides language to easily create new models
Our Approach to Probabilistic Programming

**FACTORIE**
http://code.google.com/p/factorie/

- **Object-oriented**: Variables, factors, inference & learning methods are objects,.. inheritance...

- **Embedded** in a general-purpose programming language.

- **Scalable** to billions of variables and factors. Optional DB back-end.

[McCallum, Rohanemanesh, Wick, Schultz, Singh, 2008]

Replacement for MALLET
Our Approach to Probabilistic Programming

“Imperatively-defined Factor Graphs”

Traditional **declarative** semantics of factor graphs, with some **imperative** definition of construction & operation.

- Imperatively defined jump functions
- Imperative variable value coordination
- Imperatively defined mapping from neighbor variables to features
- Imperatively defined model structure

```scala
// Joint segmentation, classification, coref on entities
// DATA TEMPLATES
class Document extends VariableSequence[Token]
class Token(word:String) extends CategoricalVariable(word)
class Mention extends SpanVariable[Token] {
  val entity = new RefVariable[Entity]
}
class Entity extends SetVariable[Mention] {
  var canonical:String = ""
  def add(m:Mention, d:DiffList) = {
    super.add(m,d); m.set(this,d)
    canonical = recomputeCanonical(members)
  }
  def remove(m:Mention, d:DiffList) = {
    super.remove(m,d); m.set(null,d)
    canonical = recomputeCanonical(members)
  }
}
// FACTOR TEMPLATES
  def unroll1 (m:Mention) = Factor(m, m.entity)
  def unroll2 (e:Entity) = for (mention <- e.mentions)
    yield Factor(mention, e)
  def statistics(m:Mention,e:Entity) =
    Bool(distance(m.string, e.canonical) < 0.5)
}
// INFERENCE
val sampler = new ProposalSampler[Mention] {
  def propose(m:Mention) = {
    // Move Mention m to a randomly-sampled Entity.
    entities.sample.add(m)
  }
}
val documents = loadData()
sampler.process(documents.mentions), 1)
```
TAC KBP

Uncertainty and Confidence
KBP Proposal

• Allow KBP systems to output a distribution over possible answers to the competition questions.

• Additional evaluation gives rewards proportional to the amount of probability assigned to the correct answer.
Prob-DB Open Questions

• How to represent uncertainty in DB?
  - Independence assumptions will bite you; but capturing all dependencies intractable.
  - Marginals? Samples? Particles?...

• How to do inference to answer queries?
  - Surprisingly tough. DB community struggling.
  - Query-specific sampling?

• How to do inference for integration/IE?
  - Sampling? Message passing; message compression?

• Scalability in (a) model complexity, (b) data.
  - Lots of juicy research at ML+systems intersect.
  - Distributed. Multi-score. ML inference guiding distribution.