Lecture 12: Information Extraction
This Lecture

- How do we represent information for information extraction?
- Semantic role labeling / abstract meaning representation
- Relation extraction
- Slot filling
- Open Information Extraction
Representing Information
“World” is a set of entities and predicates

<table>
<thead>
<tr>
<th>person</th>
<th>president</th>
<th>stab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brutus</td>
<td>Obama</td>
<td>Brutus Caesar</td>
</tr>
<tr>
<td>Caesar</td>
<td>Bush</td>
<td>...</td>
</tr>
<tr>
<td>Obama</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Bush</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Example credit: Asad Sayeed
Semantic Representations

- “World” is a set of entities and predicates

<table>
<thead>
<tr>
<th>person</th>
<th>president</th>
<th>stab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brutus</td>
<td>Obama</td>
<td>Brutus Caesar</td>
</tr>
<tr>
<td>Caesar</td>
<td>Bush</td>
<td>...</td>
</tr>
<tr>
<td>Obama</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bush</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

- Statements are logical expressions that evaluate to true or false

Example credit: Asad Sayeed
"World" is a set of entities and predicates

<table>
<thead>
<tr>
<th>person</th>
<th>president</th>
<th>stab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brutus</td>
<td>Obama</td>
<td>Brutus Caesar</td>
</tr>
<tr>
<td>Caesar</td>
<td>Bush</td>
<td></td>
</tr>
<tr>
<td>Obama</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bush</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Statements are logical expressions that evaluate to true or false

Brutus stabs Caesar

Example credit: Asad Sayeed
Semantic Representations

- “World” is a set of entities and predicates

<table>
<thead>
<tr>
<th>person</th>
<th>president</th>
<th>stab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brutus</td>
<td>Obama</td>
<td>Brutus</td>
</tr>
<tr>
<td>Caesar</td>
<td>Bush</td>
<td>Caesar</td>
</tr>
<tr>
<td>Obama</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Bush</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Statements are logical expressions that evaluate to true or false

\[
\text{Brutus stabs Caesar} \quad \text{stab(Brutus, Caesar) } \Rightarrow \text{ true}
\]
“World” is a set of entities and predicates

- person
  - Brutus
  - Caesar
  - Obama
  - Bush
  - ...
- president
  - Obama
  - Bush
  - ...
- stab
  - Brutus
  - Caesar
  - ...

Statements are logical expressions that evaluate to true or false

\[ Brutus \text{ stabs } Caesar \quad \Rightarrow \quad \text{stab(Brutus, Caesar)} \Rightarrow \text{true} \]

\[ Caesar \text{ was stabbed} \]
“World” is a set of entities and predicates

- person
  - Brutus
  - Caesar
  - Obama
  - Bush
  - ...

- president
  - Obama
  - Bush
  - ...

- stab
  - Brutus Caesar
  - ...

Statements are logical expressions that evaluate to true or false

- Brutus stabs Caesar
  - stab(Brutus, Caesar) => true

- Caesar was stabbed
  - ∃x stab(x, Caesar) => true

Example credit: Asad Sayeed
Neo-Davidsonian Events

Brutus stabbed Caesar with a knife at the theater on the Ides of March

Example credit: Asad Sayeed
Neo-Davidsonian Events

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

\[ \exists e \text{ stabs}(e, \text{Brutus, Caesar}) \]
Neo-Davidsonian Events

Brutus stabbed Caesar with a knife at the theater on the Ides of March

\[ \exists e \text{ stabs}(e, \text{Brutus, Caesar}) \land \text{with}(e, \text{knife}) \]
Neo-Davidsonian Events

Brutus stabbed Caesar with a knife at the theater on the Ides of March

\[ \exists e \, \text{stabs}(e, \text{Brutus}, \text{Caesar}) \land \text{with}(e, \text{knife}) \land \text{location}(e, \text{theater}) \]
Neo-Davidsonian Events

*Brutus stabbed Caesar with a knife at the theater on the Ides of March*

\[ \exists e \text{ stabs}(e, \text{Brutus, Caesar}) \land \text{with}(e, \text{knife}) \land \text{location}(e, \text{theater}) \land \text{time}(e, \text{Ides of March}) \]
Neuro-Davidsonian Events

Brutus stabbed Caesar with a knife at the theater on the Ides of March

$\exists e \text{ stabs}(e, \text{Brutus, Caesar}) \land \text{with}(e, \text{knife}) \land \text{location}(e, \text{theater})$

$\land \text{time}(e, \text{Ides of March})$

- Lets us describe events as having properties
Neo-Davidsonian Events

Brutus stabbed Caesar with a knife at the theater on the Ides of March

$$\exists e \text{ stabs}(e, \text{Brutus}, \text{Caesar}) \land \text{with}(e, \text{knife}) \land \text{location}(e, \text{theater}) \land \text{time}(e, \text{Ides of March})$$

- Lets us describe events as having properties
- Unified representation of events and entities:

Example credit: Asad Sayeed
Neo-Davidsonian Events

Brutus stabbed Caesar with a knife at the theater on the Ides of March

\[\exists e \text{ stabs}(e, \text{Brutus, Caesar}) \land \text{with}(e, \text{knife}) \land \text{location}(e, \text{theater}) \land \text{time}(e, \text{Ides of March})\]

- Lets us describe events as having properties
- Unified representation of events and entities:

  *some clever driver in America*
Neo-Davidsonian Events

Brutus stabbed Caesar with a knife at the theater on the Ides of March

\[ \exists e \text{ stabs}(e, \text{Brutus, Caesar}) \land \text{with}(e, \text{knife}) \land \text{location}(e, \text{theater}) \land \text{time}(e, \text{Ides of March}) \]

- Lets us describe events as having properties
- Unified representation of events and entities:

some clever driver in America

\[ \exists x \text{ driver}(x) \land \text{clever}(x) \land \text{location}(x, \text{America}) \]
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

\[ \exists e \text{ sign}(e, \text{Barack Obama}) \land \text{patient}(e, \text{ACA}) \land \text{time}(e, \text{Tuesday}) \]
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

\[ \exists e \text{ sign}(e, \text{Barack Obama}) \land \text{patient}(e, \text{ACA}) \land \text{time}(e, \text{Tuesday}) \]
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

\[ \exists e \text{ sign}(e, \text{Barack Obama}) \land \text{patient}(e, \text{ACA}) \land \text{time}(e, \text{Tuesday}) \]
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

$$\exists e \text{ sign}(e, \text{Barack Obama}) \land \text{patient}(e, \text{ACA}) \land \text{time}(e, \text{Tuesday})$$
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

Need to impute missing information, resolve coreference, etc.
Barack Obama signed the Affordable Care Act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

∃e sign(e, Barack Obama) ∧ patient(e, ACA) ∧ time(e, Tuesday)
Other Challenges

Bob and Alice were friends until he moved away to attend college
Other Challenges

Bob and Alice were friends until he moved away to attend college

\[ \exists e_1 \exists e_2 \ \text{friends}(e_1, \text{Bob, Alice}) \land \text{moved}(e_2, \text{Bob}) \land \text{end\_of}(e_1, e_2) \]
Other Challenges

Bob and Alice were friends until he moved away to attend college

\[ \exists e_1 \exists e_2 \text{ friends}(e_1, \text{Bob}, \text{Alice}) \land \text{moved}(e_2, \text{Bob}) \land \text{end_of}(e_1, e_2) \]

- How to represent temporal information?
Other Challenges

Bob and Alice were friends until he moved away to attend college

$\exists e_1 \exists e_2$ friends($e_1$, Bob, Alice) $\land$ moved($e_2$, Bob) $\land$ end_of($e_1$, $e_2$)

- How to represent temporal information?

Bob and Alice were friends until around the time he moved away to attend college
Other Challenges

Bob and Alice were friends until he moved away to attend college

\[ \exists e_1 \exists e_2 \text{ friends}(e_1, \text{Bob, Alice}) \land \text{moved}(e_2, \text{Bob}) \land \text{end_of}(e_1, e_2) \]

- How to represent temporal information?

Bob and Alice were friends until around the time he moved away to attend college

- Representing truly open-domain information is very complicated! We don’t have a formal representation that can capture everything
(At least) Three Solutions
Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...
(At least) Three Solutions

- Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...

- Slot filling: specific ontology, populate information in a predefined way
Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...

Slot filling: specific ontology, populate information in a predefined way

(Earthquake: magnitude=8.0, epicenter=central Italy, ...)
(At least) Three Solutions

- Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...

- Slot filling: specific ontology, populate information in a predefined way
  
  (Earthquake: magnitude=8.0, epicenter=central Italy, ...)

- Entity-relation-entity triples: focus on entities and their relations (note that prominent events can still be entities)
Crafted annotations to capture some subset of phenomena: predicate-argument structures (semantic role labeling), time (temporal relations), ...

Slot filling: specific ontology, populate information in a predefined way

(Earthquake: magnitude=8.0, epicenter=central Italy, ...)

Entity-relation-entity triples: focus on entities and their relations (note that prominent events can still be entities)

(Lady Gaga, singerOf, Bad Romance)
Open IE

- Entity-relation-entity triples aren’t necessarily grounded in an ontology
- Extract strings and let a downstream system figure it out
Open IE

- Entity-relation-entity triples aren’t necessarily grounded in an ontology
- Extract strings and let a downstream system figure it out

*Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.*
Entity-relation-entity triples aren’t necessarily grounded in an ontology

Extract strings and let a downstream system figure it out

Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

(Barack Obama, signed, the Affordable Care act)
Barack Obama signed the Affordable Care act on Tuesday. He gave a speech later that afternoon on how the act would help the American people. Several prominent Republicans were quick to denounce the new law.

(Barack Obama, signed, the Affordable Care act)
(Several prominent Republicans, denounce, the new law)
IE: The Big Picture

- How do we represent information? What do we extract?
  - Semantic roles
  - Abstract meaning representation
  - Slot fillers
  - Entity-relation-entity triples (fixed ontology or open)
Semantic Role Labeling/
Abstract Meaning Representation
Semantic Role Labeling
Semantic Role Labeling

Figure from He et al. (2017)
Semantic Role Labeling

- Identify predicate, disambiguate it, identify that predicate’s arguments

Figure from He et al. (2017)
Semantic Role Labeling

- Identify predicate, disambiguate it, identify that predicate’s arguments
- Verb roles from Propbank (Palmer et al., 2005)

Gold

Housing starts are expected to quicken a bit from August’s pace

Figure from He et al. (2017)
Semantic Role Labeling

- Identify predicate, disambiguate it, identify that predicate’s arguments
- Verb roles from Propbank (Palmer et al., 2005)

![Diagram](image)

quicken:

**Arg0-PAG:** *causer of speed-up*
**Arg1-PPT:** *thing becoming faster* (vnrole: 45.4-patient)
**Arg2-EXT:** *EXT*
**Arg3-DIR:** *old speed*
**Arg4-PRD:** *new speed*

Figure from He et al. (2017)
Semantic Role Labeling

Figure from He et al. (2017)
Semantic Role Labeling

- Identify predicates (*love*) using a classifier (not shown)
Semantic Role Labeling

- Identify predicates (love) using a classifier (not shown)
- Identify ARG0, ARG1, etc. as a tagging task with a BiLSTM conditioned on love

Figure from He et al. (2017)
Semantic Role Labeling

- Identify predicates (love) using a classifier (not shown)
- Identify ARG0, ARG1, etc. as a tagging task with a BiLSTM conditioned on love
- Other systems incorporate syntax, joint predicate-argument finding

Figure from He et al. (2017)
SRL for QA

- Question and several answer candidates

Q: Who discovered prions?

AC1: In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...

AC2: Prions were researched by...
SRL for QA

- Question and several answer candidates

Q: Who discovered prions?

AC1: In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...

AC2: Prions were researched by...

Shen and Lapata (2007)
SRL for QA

- Question and several answer candidates

Q: *Who discovered prions?*

AC1: *In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...*

AC2: *Prions were researched by...*

Shen and Lapata (2007)
SRL for QA

- Question and several answer candidates

Q: Who discovered prions?

AC1: In 1997, Stanley B. Prusiner, a scientist in the United States, discovered prions...

AC2: Prions were researched by...

Score by matching expected answer phrase (EAP) against answer candidate (AC)

Shen and Lapata (2007)
Abstract Meaning Representation

- Graph-structured annotation

The boy wants to go

Banarescu et al. (2014)
Abstract Meaning Representation

- Graph-structured annotation
- Superset of SRL: full sentence analyses, contains coreference and multi-word expressions as well

The boy wants to go

Banarescu et al. (2014)
Abstract Meaning Representation

- Graph-structured annotation
- Superset of SRL: full sentence analyses, contains coreference and multi-word expressions as well
- F1 scores in the 60s: hard!
Abstract Meaning Representation

- Graph-structured annotation
- Superset of SRL: full sentence analyses, contains coreference and multi-word expressions as well
- F1 scores in the 60s: hard!
- So comprehensive that it’s hard to predict, but still doesn’t handle tense or some other things…

The boy wants to go
Summarization with AMR

Sentence A: I saw Joe’s dog, which was running in the garden.
Sentence B: The dog was chasing a cat.

Summary: Joe’s dog was chasing a cat in the garden.

Liu et al. (2015)
Summarization with AMR

Sentence A: I saw Joe’s dog, which was running in the garden.
Sentence B: The dog was chasing a cat.

- Merge AMRs across multiple sentences

Summary: Joe’s dog was chasing a cat in the garden.
Summarization with AMR

- Merge AMRs across multiple sentences
- Summarization = subgraph extraction

Liu et al. (2015)
Summarization with AMR

- Merge AMRs across multiple sentences
- Summarization = subgraph extraction
- No real systems actually work this way (more when we talk about summarization)

Liu et al. (2015)
Slot Filling
Slot Filling

- Most conservative, narrow form of IE

Freitag and McCallum (2000)
Most conservative, narrow form of IE

Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

Freitag and McCallum (2000)
Most conservative, narrow form of IE

\[ \text{magnitude} \quad \text{time} \]

Indian Express — A massive earthquake of \text{magnitude 7.3} struck Iraq on \text{Sunday}, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

\text{Speaker: Alan Clark} \quad \text{title}

“Gender Roles in the Holy Roman Empire”

\text{Allagher Center Main Auditorium} \quad \text{location}

This talk will discuss...

Freitag and McCallum (2000)
Most conservative, narrow form of IE

Indian Express — A massive earthquake of magnitude 7.3 struck Iraq on Sunday, 103 kms (64 miles) southeast of the city of As-Sulaymaniyah, the US Geological Survey said, reports Reuters. US Geological Survey initially said the quake was of a magnitude 7.2, before revising it to 7.3.

Speaker: Alan Clark

“Gender Roles in the Holy Roman Empire”

Old work: HMMs, later CRFs trained per role

Freitag and McCallum (2000)
Slot Filling: MUC

Key aspect: need to combine information across multiple mentions of an entity using coreference

Haghighi and Klein (2010)
Slot Filling: Forums

- Extract product occurrences in cybercrime forums, but not everything that looks like a product is a product.

<table>
<thead>
<tr>
<th>TITLE: [ buy ] Backconnect bot</th>
</tr>
</thead>
<tbody>
<tr>
<td>BODY: Looking for a solid backconnect bot.</td>
</tr>
<tr>
<td>If you know of anyone who codes them please let me know</td>
</tr>
</tbody>
</table>

(a) File 0-initiator4856

<table>
<thead>
<tr>
<th>TITLE: Exploit cleaning ?</th>
</tr>
</thead>
<tbody>
<tr>
<td>BODY: Have some Exploits i need fud.</td>
</tr>
</tbody>
</table>

(b) File 0-initiator10815

Not a product in this context

Portnoff et al. (2017), Durrett et al. (2017)
Relation Extraction
Relation Extraction

- Extract entity-relation-entity triples from a fixed inventory

ACE (2003-2005)
During the war in Iraq, American journalists were sometimes caught in the line of fire
Relation Extraction

- Extract entity-relation-entity triples from a fixed inventory

*During the war in* [Iraq](#), *[American journalists](#)* were sometimes caught in the line of fire

ACE (2003-2005)
During the war in **Iraq**, **American journalists** were sometimes caught in the line of fire.
During the war in **Iraq**, **American journalists** were sometimes caught in the line of fire

- Extract entity-relation-entity triples from a fixed inventory

- Use NER-like system to identify entity spans, classify relations between entity pairs with a classifier

ACE (2003-2005)
Relation Extraction

- Extract entity-relation-entity triples from a fixed inventory
  
  Located_In

  Nationality

  During the war in **Iraq**, **American journalists** were sometimes caught in the line of fire

- Use NER-like system to identify entity spans, classify relations between entity pairs with a classifier

- Systems can be feature-based or neural, look at surface words, syntactic features (dependency paths), semantic roles

ACE (2003-2005)
Relation Extraction

- Extract entity-relation-entity triples from a fixed inventory

Located_In

Nationality

During the war in **Iraq**, **American journalists** were sometimes caught in the line of fire

- Use NER-like system to identify entity spans, classify relations between entity pairs with a classifier

- Systems can be feature-based or neural, look at surface words, syntactic features (dependency paths), semantic roles

- Problem: limited data for scaling to big ontologies

ACE (2003-2005)
Hearst Patterns

- Syntactic patterns especially for finding hypernym-hyponym pairs ("is a" relations)
Hearst Patterns

- Syntactic patterns especially for finding hypernym-hyponym pairs ("is a" relations)
Hearst Patterns

- Syntactic patterns especially for finding hypernym-hyponym pairs ("is a" relations)

\[ Y \text{ is a } X \quad \text{Berlin is a city} \]
Hearst Patterns

- Syntactic patterns especially for finding hypernym-hyponym pairs ("is a" relations)

- $Y$ is a $X$
  
  \begin{center}
  \textit{Berlin is a city}
  \end{center}

- $X$ such as [list]
  
  \begin{center}
  \textit{cities such as Berlin, Paris, and London.}
  \end{center}
Hearst Patterns

- Syntactic patterns especially for finding hypernym-hyponym pairs ("is a" relations)

  \[
  Y \text{ is a } X \quad \text{Berlin is a city}
  \]
  \[
  X \text{ such as [list]} \quad \text{cities such as Berlin, Paris, and London.}
  \]
  \[
  \text{other } X \text{ including } Y \quad \text{other cities including Berlin}
  \]
Hearst Patterns

- Syntactic patterns especially for finding hypernym-hyponym pairs ("is a" relations)

  - $Y$ is a $X$
    - Berlin is a city
  
  - $X$ such as [list]
    - cities such as Berlin, Paris, and London.

  - other $X$ including $Y$
    - other cities including Berlin

- Totally unsupervised way of harvesting world knowledge for tasks like parsing and coreference (Bansal and Klein, 2011-2012)

Hearst (1992)
Distant Supervision

Mintz et al. (2009)
Distant Supervision

- Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data
Distant Supervision

- Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data
- If two entities in a relation appear in the same sentence, assume the sentence expresses the relation

Mintz et al. (2009)
Distant Supervision

- Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data

- If two entities in a relation appear in the same sentence, assume the sentence expresses the relation

  Director

  [Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers’ story
Lots of relations in our knowledge base already (e.g., 23,000 film-director relations); use these to bootstrap more training data

If two entities in a relation appear in the same sentence, assume the sentence expresses the relation

- [Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brothers’ story
- Allison co-produced the Academy Award-winning [Saving Private Ryan], directed by [Steven Spielberg]
Distant Supervision

- Learn decently accurate classifiers for ~100 Freebase relations
- Could be used to crawl the web and expand our knowledge base

<table>
<thead>
<tr>
<th>Relation name</th>
<th>100 instances</th>
<th>1000 instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Syn</td>
<td>Lex</td>
</tr>
<tr>
<td>/film/director/film</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>/film/writer/film</td>
<td>0.70</td>
<td>0.60</td>
</tr>
<tr>
<td>/geography/river/basin_countries</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>/location/country/administrative.divisions</td>
<td>0.68</td>
<td>0.59</td>
</tr>
<tr>
<td>/location/location/contains</td>
<td>0.81</td>
<td>0.89</td>
</tr>
<tr>
<td>/location/us_county/country_seat</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>/music/artist/origin</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td>/people/deceased_person/place_of_death</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>/people/person/nationality</td>
<td>0.61</td>
<td>0.70</td>
</tr>
<tr>
<td>/people/person/place_of_birth</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td>Average</td>
<td>0.67</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Mintz et al. (2009)
Open IE
Open Information Extraction

- “Open”ness — want to be able to extract all kinds of information from open-domain text

- Acquire commonsense knowledge just from “reading” about it, but need to process lots of text (“machine reading”)

- Typically no fixed relation inventory
TextRunner

- Extract positive examples of (e, r, e) triples via parsing and heuristics
- Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

Banko et al. (2007)
Extract positive examples of (e, r, e) triples via parsing and heuristics

Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

Barack Obama, 44th president of the United States, was born on August 4, 1961 in Honolulu

=> Barack Obama, was born in, Honolulu

Banko et al. (2007)
Extract positive examples of (e, r, e) triples via parsing and heuristics

Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

Barack Obama, 44th president of the United States, was born on August 4, 1961 in Honolulu

=> Barack Obama, was born in, Honolulu

80x faster than running a parser (which was slow in 2007...)

Banko et al. (2007)
Extract positive examples of (e, r, e) triples via parsing and heuristics

Train a Naive Bayes classifier to filter triples from raw text: uses features on POS tags, lexical features, stopwords, etc.

Barack Obama, 44th president of the United States, was born on August 4, 1961 in Honolulu

=> Barack Obama, was born in, Honolulu

80x faster than running a parser (which was slow in 2007...)

Use multiple instances of extractions to assign probability to a relation

Banko et al. (2007)
Exploiting Redundancy

Banko et al. (2007)
Exploiting Redundancy

- 9M web pages / 133M sentences

Banko et al. (2007)
Exploiting Redundancy

- 9M web pages / 133M sentences
- 2.2 tuples extracted per sentence, filter based on probabilities

Banko et al. (2007)
Exploiting Redundancy

- 9M web pages / 133M sentences
- 2.2 tuples extracted per sentence, filter based on probabilities
- Concrete: definitely true
  Abstract: possibly true but underspecified

Banko et al. (2007)
Exploiting Redundancy

- 9M web pages / 133M sentences
- 2.2 tuples extracted per sentence, filter based on probabilities
- Concrete: definitely true
  Abstract: possibly true but underspecified
- Hard to evaluate: can assess precision of extracted facts, but how do we know recall?

Banko et al. (2007)
More constraints: open relations have to begin with verb, end with preposition, be contiguous (e.g., *was born on*)
ReVerb

- More constraints: open relations have to begin with verb, end with preposition, be contiguous (e.g., *was born on*)

- Extract more meaningful relations, particularly with light verbs

<table>
<thead>
<tr>
<th>is</th>
<th>is an album by, is the author of, is a city in</th>
</tr>
</thead>
<tbody>
<tr>
<td>has</td>
<td>has a population of, has a Ph.D. in, has a cameo in</td>
</tr>
<tr>
<td>made</td>
<td>made a deal with, made a promise to</td>
</tr>
<tr>
<td>took</td>
<td>took place in, took control over, took advantage of</td>
</tr>
<tr>
<td>gave</td>
<td>gave birth to, gave a talk at, gave new meaning to</td>
</tr>
<tr>
<td>got</td>
<td>got tickets to, got a deal on, got funding from</td>
</tr>
</tbody>
</table>
For each verb, identify the longest sequence of words following the verb that satisfy a POS regex (V .* P) and which satisfy heuristic lexical constraints on specificity

Find the nearest arguments on either side of the relation
For each verb, identify the longest sequence of words following the verb that satisfy a POS regex (V.*P) and which satisfy heuristic lexical constraints on specificity.

Find the nearest arguments on either side of the relation.

Annotators labeled relations in 500 documents to assess recall.
(a) **CCG parse** builds an underspecified semantic representation of the sentence.

\[
\begin{align*}
\text{Former} & \quad \text{municipalities} & \quad \text{in} & \quad \text{Brandenburg} \\
\frac{N/N}{\lambda f \lambda x. f(x) \land \text{former}(x)} & \quad \frac{N}{\lambda x. \text{municipalities}(x)} & \quad \frac{N \backslash N/NP}{\lambda f \lambda x \lambda y. f(y) \land \text{in}(y, x)} & \quad \frac{NP}{\text{Brandenburg}} \\
\frac{N}{\lambda x. \text{former}(x) \land \text{municipalities}(x)} & \quad \frac{N \backslash N}{\lambda f \lambda y. f(y) \land \text{in}(y, \text{Brandenburg})} & \quad < \\
\_0 & = \lambda x. \text{former}(x) \land \text{municipalities}(x) \land \text{in}(x, \text{Brandenburg})
\end{align*}
\]

(b) **Constant matches** replace underspecified constants with Freebase concepts

\[
\begin{align*}
\_0 & = \lambda x. \text{former}(x) \land \text{municipalities}(x) \land \text{in}(x, \text{Brandenburg}) \\
\_1 & = \lambda x. \text{former}(x) \land \text{municipalities}(x) \land \text{in}(x, \text{Brandenburg}) \\
\_2 & = \lambda x. \text{former}(x) \land \text{municipalities}(x) \land \text{location.containedby}(x, \text{Brandenburg}) \\
\_3 & = \lambda x. \text{former}(x) \land \text{OpenRel}(x, \text{Municipality}) \land \text{location.containedby}(x, \text{Brandenburg}) \\
\_4 & = \lambda x. \text{OpenType}(x) \land \text{OpenRel}(x, \text{Municipality}) \land \text{location.containedby}(x, \text{Brandenburg})
\end{align*}
\]
Takeaways

- **SRL/AMR**: handle a bunch of phenomena, but more or less like syntax++ in terms of what they represent.
- **Relation extraction**: can collect data with distant supervision, use this to expand knowledge bases.
- **Slot filling**: tied to a specific ontology, but gives fine-grained information.
- **Open IE**: extracts lots of things, but hard to know how good or useful they are.
  - Can combine with standard question answering.
  - Add new facts to knowledge bases.
- Many, many applications and techniques.