Lecture 20:
Information Extraction, SRL, etc.

Greg Durrett

CS388: Natural Language Processing

Administrivia

- Final project presentation slots announced
- Project 2 due today

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
Semantics Representations

- “World” is a set of entities and predicates
  
<table>
<thead>
<tr>
<th>person</th>
<th>president</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brutus</td>
<td>Obama</td>
</tr>
<tr>
<td>Caesar</td>
<td>Bush</td>
</tr>
</tbody>
</table>
  
- Statements are logical expressions that evaluate to true or false
  
  Brutus stabs Caesar: \( \text{stab}(\text{Brutus}, \text{Caesar}) \Rightarrow \text{true} \)
  
  Caesar was stabbed: \( \exists x \, \text{stab}(x, \text{Caesar}) \Rightarrow \text{true} \)

Example credit: Asad Sayeed

Neo-Davidsonian Events

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

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

- Lets us describe events as having properties
- Unified representation of events and entities:
  
  Some clever driver in America
  
  \[ \exists x \, \text{driver}(x) \land \text{clever}(x) \land \text{location}(x, \text{America}) \]

Example credit: Asad Sayeed

Real Text

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. Several prominent Republicans were quick to denounce the new law.

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

- Need to impute missing information, resolve coreference, etc.
- Still unclear how to represent some things precisely or how that information could be leveraged (several prominent Republicans)

Other Challenges

Example credit: Asad Sayeed
(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 entities is pretty broad: can include events like World War II, etc.)
  
  (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

  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

- Performing event
- Subject: Lady Gaga
- Object: a concert
- Audience: students

Same event described but the representation looks different
Verb (predicate) associated with several arguments (roles): “Agent”, “Theme”, and “Beneficiary”

VerbNet

- Defines the semantics of verbs, arguments for every verb in English

Semantic Roles

- “Postprocessing” layer on top of dependency parsing that exposes useful information, canonicalizes across grammatical constructions
- Related to theta roles in linguistics
- Agent (~ subject), patient/theme (~ object), goal (~ indirect object)

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

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

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...*

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

Shen and Lapata (2007)

Abstract Meaning Representation

- Graph-structured annotation
  
  Banarescu et al. (2014)

- 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

Flanigan et al. (2016), Lyu et al. (2018)
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

- 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”

Allagher Center Main Auditorium

This talk will discuss...

Freitag and McCallum (2000)

Slot Filling: MUC

Template

<table>
<thead>
<tr>
<th>Template</th>
<th>SELLER</th>
<th>BUSINESS</th>
<th>ACQUIRED</th>
<th>PURCHASER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSR Ltd.</td>
<td>Oil and Gas</td>
<td>Delhi Fund</td>
<td>Esso Inc.</td>
<td></td>
</tr>
</tbody>
</table>

Document

[S CSR] has said that [S it] has sold [S its] [B oil interests] held in [A Delhi Fund]. [P Esso Inc.] did not disclose how much [P they] paid for [A Dehl].

Freitag and McCallum (2000)
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:</th>
<th>Buy</th>
<th>Backconnect bot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body:</td>
<td>Looking for a solid backconnect bot. If you know of anyone who codes them please let me know</td>
<td></td>
</tr>
</tbody>
</table>

(a) File 0-initiator4856

<table>
<thead>
<tr>
<th>Title:</th>
<th>Exploit cleaning?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body:</td>
<td>Have some exploits I need fixed.</td>
</tr>
</tbody>
</table>

(b) File 0-initiator10815

Not a product in this context

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

Relation Extraction

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

Relation Extraction

- 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).

\[ Y \text{ is a } X \]

\[ X \text{ such as [list]} \]

\[ \text{other } X \text{ including } Y \]

\[ \text{other cities including Berlin} \]

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

Hearst (1992)
**Distant Supervision**

- 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]

**FewRel**

- Treats relation classification as a few-shot classification problem
- 100 classes x 700 instances, goal is to generalize to each class with just a few instances
- BERT can handle this fairly well (Soares et al., 2019)
- “FewRel 2.0”: new dataset with “none of the above” type, which makes things much harder

**Entity Tracking / Procedural Text**
**Entity Tracking**

- Information extraction for “procedural text”: text describing some kind of process
- For a recipe: what ingredients are involved at each timestep?
- Involves global constraints and being able to model complex entity interactions

**Recipes Dataset**

<table>
<thead>
<tr>
<th>Seq. of Steps</th>
<th>sugar</th>
<th>eggs</th>
<th>flour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combine sugar, oil, and vinegar</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Add eggs one at a time</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>In a separate bowl, combine flour, soda, and salt</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Add to the sugar mixture alternately with milk</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Stir remaining ingredients one at a time.</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Kiddon et al. (2016), Bosselut et al. (2018)

**ProPara Dataset**

- Process paragraphs: predict when objects are created, moved, or destroyed in a scientific process
- Structured prediction problem, tied to the particular information conveyed in these paragraphs
- Use a neural CRF to make a coherent prediction for each entity

**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.

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

- 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)

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

ReVerb

- 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

Fader et al. (2011)
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