CS395T: Structured Models for NLP
Lecture 25: Information Extraction

Greg Durrett

Administrivia

- Project presentations coming up next week! Come to OHs or email me if you don't feel like you're on track to have something
- Course evaluations: please do these!

Chatbots

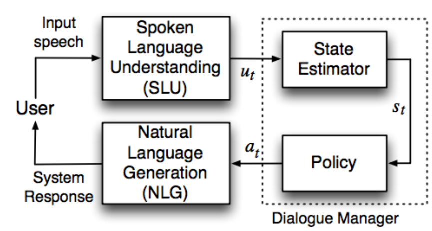
message	SEQ2SEQ	MMI $P(T\|S) - \lambda P(T)$
I did not get the report from an mi-6 agent.	I don't think that's a good idea.	You did the right thing, did you?
You haven't been given an assignment in this case.	I don't know what you are talking about.	I've been looking all over for you.
I'm losing my grip.	I don't know what you are talking about.	I'm the only one in the world.
I am ready to help.	Come on, come on	I have something we need to talk about.
You programmed me to gather intelligence. That's all I've ever done.	You know that, don't you?	You do have fun, don't you?
I mean, we'd have to talk to him.	I mean, I don't know.	I mean, he's a good guy.
This has nothing to do with him.	You don't know anything about him, do you?	I know him as much.
How come you never say it?	I don't know	Because I don't want to hurt you

- Can model as machine translation, but need to endow with diversity, add consistency among answers, …

Li et al. (2016)

Task-oriented dialogue

- Involves both generation and language understanding
- Dialogue state: reflects any information about the conversation (e.g., search history)



- User utterance -> update dialogue state -> take action (e.g., query the restaurant database) -> say something

Young et al. (2013)

Full Dialogue Task

Find me a good sushi restaurant in Chelsea

```
restaurant_type <- sushi
location <- Chelsea
curr_result <- execute_search()
```

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

```
get_value(cost, curr_result)
```

Entrees are around $30 each

This Lecture

- How do we represent information for information extraction?
- Relation extraction
- Slot filling
- Open Information Extraction

Representing Information

Semantic Representations

- "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 $\exists x$ stab(x, Caesar) => true

Example credit: Asad Sayeed

Semantic Representations

Brutus stabs Caesar

stab(Brutus, Caesar)

Brutus stabbed Caesar with a knife

stab(Brutus, Caesar, instrument=knife)

Brutus stabbed Caesar with a knife in the agora

stab(Brutus, Caesar, instrument=knife, location=agora)

Brutus stabbed Caesar with a knife in the agora on the Ides of March

...

Example credit: Asad Sayeed

Neo-Davidsonian Events

Brutus stabbed Caesar with a knife in the agora on the Ides of March

$\exists e$ stabs(e, Brutus, Caesar) $\wedge$ with(e, knife) $\wedge$ location(e, agora) $\wedge$ time(e, Ides of March)

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

some clever driver in America

$\exists x$ driver(x) $\wedge$ clever(x) $\wedge$ location(x, America)

Example credit: Asad Sayeed

Real Text

which afternoon? which Tuesday? who?

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.
- Still unclear how to represent some things precisely or how that information could be leveraged (several prominent Republicans)

Other Challenges

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

$\exists e1 \exists e2$ friends($e1$, Bob, Alice) $\wedge$ moved($e2$, Bob) $\wedge$ end_of($e1$, $e2$)

- 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

(At least) Two Solutions

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

(Barack Obama, presidentOf, United States)

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

Entity-Relation-Entity Pairs

- Represent semantics as relationships between entities; relationships are drawn from a fixed ontology

Table 5: Sample facts of YAGO

Zidane	TYPE+SUBCLASS	football player
Zidane	TYPE	Person from Marseille
Zidane	TYPE	Legion d'honneur recipient
Zidane	BORNINYEAR	1972
"Paris"	FAMILYNAMEOF	Priscilla Paris
"Paris"	GIVENNAMEOF	Paris Hilton
"Paris"	MEANS	Paris, France
"Paris"	MEANS	Paris, Texas
Paris, France	LOCATEDIN	France
Paris, France	TYPE+SUBCLASS	capital
Paris, France	TYPE	Eurovision host city
Paris, France	ESTABLISHEDIN	-300

Suchanek et al. (2007)

Entity-Relation-Entity Pairs

- Can easy query about relations in the knowledge base

when was Barack Obama born? λx. born(Barack_Obama, x)

how many children does Barack Obama have?

sizeof(λx. isParent(x, Barack_Obama))

how old was Barack Obama when he became president?
— no timeOfBecomingPresident relation

how many Wimbledon victories has Serena Williams had?
— Wimbledons are listed, but no isWimbledon predicate

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)

Slot Filling

- Represent information about a particular event like an earthquake

magnitude time

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.

epicenter

IE: The Big Picture

- How do we represent information? What do we extract?
  - Entity-relation-entity triples (fixed ontology or open)
  - Slot fillers
- Where does that information come from? (closed vs. open IE)
  - Closed: limited set of documents, domain-specific
  - Open: try to use lots of information (the whole Internet)

Relation Extraction

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

- Pipelined classifiers looking at surface level, 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 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

- 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]

Mintz et al. (2009)

Distant Supervision

- Learn decently accurate classifiers for ~100 Freebase relations

Relation name	100 instances			1000 instances		
	Syn	Lex	Both	Syn	Lex	Both
/film/director/film	**0.49**	0.43	0.44	**0.49**	0.41	0.46
/film/writer/film	**0.70**	0.60	0.65	**0.71**	0.61	0.69
/geography/river/basin_countries	0.65	0.64	**0.67**	**0.73**	0.71	0.64
/location/country/administrative_divisions	0.68	0.59	**0.70**	**0.72**	0.68	**0.72**
/location/location/contains	0.81	**0.89**	0.84	**0.85**	0.83	0.84
/location/us_county/county_seat	0.51	0.51	**0.53**	0.47	**0.57**	0.42
/music/artist/origin	0.64	0.66	**0.71**	0.61	**0.63**	0.60
/people/deceased_person/place_of_death	0.80	0.79	**0.81**	0.80	**0.81**	0.78
/people/person/nationality	0.61	0.70	**0.72**	0.56	0.61	**0.63**
/people/person/place_of_birth	**0.78**	0.77	**0.78**	0.88	0.85	**0.91**
Average	0.67	0.66	**0.69**	**0.68**	0.67	0.67

Mintz et al. (2009)

Slot Filling

Slot Filling

- Extract a fixed set of roles from a relatively ordered text like a seminar announcement

Speaker: [Alan Clark]$_{Speaker}$
[“Gender Roles in the Holy Roman Empire”]$_{Title}$
[Allagher Center Main Auditorium]$_{Location}$

This talk will discuss…

- Old work: HMMs, later CRFs trained per role

Freitag and McCallum (2000)

Slot Filling: MUC

Template

(a)

SELLER	BUSINESS	ACQUIRED	PURCHASER
CSR Limited	Oil and Gas	Delhi Fund	Esso Inc.

Document

(b) [**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** Dehli].

- 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

TITLE: [buy] Backconnect bot
BODY: Looking for a solid backconnect bot .
If you know of anyone who codes them please let me know

(a) File 0-initiator4856

TITLE: Exploit cleaning ?
BODY: Have some $\boxed{\text{Exploits}}$ i need fud .

(b) File 0-initiator10815

Not a product in this context

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

Open IE + IR

- Can retrieve additional information about specific events
- If we’re uncertain about extractions, fetch another article to improve confidence

current belief

ShooterName	Scott Westerhuis
NumKilled	4
NumWounded	2
City	Platte

latest extraction

ShooterName	Scott Westerhuis
NumKilled	6
NumWounded	0
City	Platte

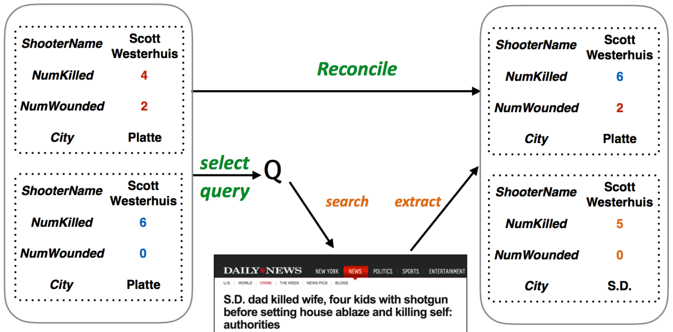
Narasimhan et al. (2016)

Open IE + IR

- Can retrieve additional information about specific events
- If we're uncertain about extractions, fetch another article to improve confidence

current belief

latest extraction



Narasimhan et al. (2016)

Open IE + IR

- Use reinforcement learning to send queries about specific things

⟨*title*⟩
⟨*title*⟩ + (police | identified | arrested | charged)
⟨*title*⟩ + (killed | shooting | injured | dead | people)
⟨*title*⟩ + (injured | wounded | victim)
⟨*title*⟩ + (city | county | area)

System	Shootings			
	ShooterName	**NumKilled**	**NumWounded**	**City**
CRF extractor	9.5	65.4	64.5	47.9
Maxent extractor	45.2	69.7	68.6	53.7
Confidence Agg. (τ)	45.2 (0.6)	70.3 (0.6)	72.3 (0.6)	55.8 (0.6)
RL-Extract	**50.0**	**77.6***	**74.6***	**65.6***
ORACLE	57.1	86.4	83.3	71.8

Narasimhan et al. (2016)

Open IE

Open Information Extraction

- "Open"ness — want to be able to extract all kinds of information from open-domain text
- "Machine reading the web" — acquire commonsense knowledge just from reading about it, but need to process lots of text
- Typically no fixed relation inventory

TextRunner

- Supervised system
  - Extract positive examples of (e, r, e) triples via parsing and heuristics
  - Train a Naive Bayes classifier to filter pairs 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

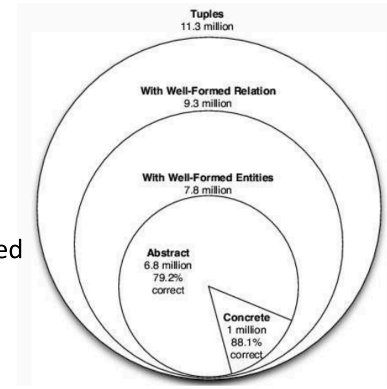
=> Barack_Obama, was born in, Honolulu

- 80x faster than running a parser
- 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

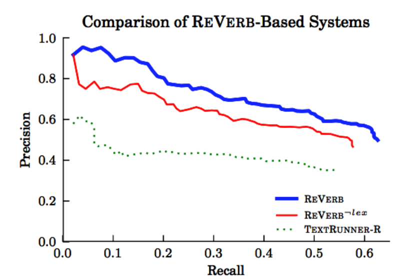
- More constraints: open relations have to begin with verb, end with preposition, be contiguous (e.g., *was born on*)
- Extract more meaningful relations, particularly with light verbs

is	is an album by, is the author of, is a city in
has	has a population of, has a Ph.D. in, has a cameo in
made	made a deal with, made a promise to
took	took place in, took control over, took advantage of
gave	gave birth to, gave a talk at, gave new meaning to
got	got tickets to, got a deal on, got funding from

Fader et al. (2011)

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)

NELL

- Entity typing/resolution + relation classification to read facts about things, combine with logical inference as well
- Coupling constraints: types of arguments to relations must match the relation extracted

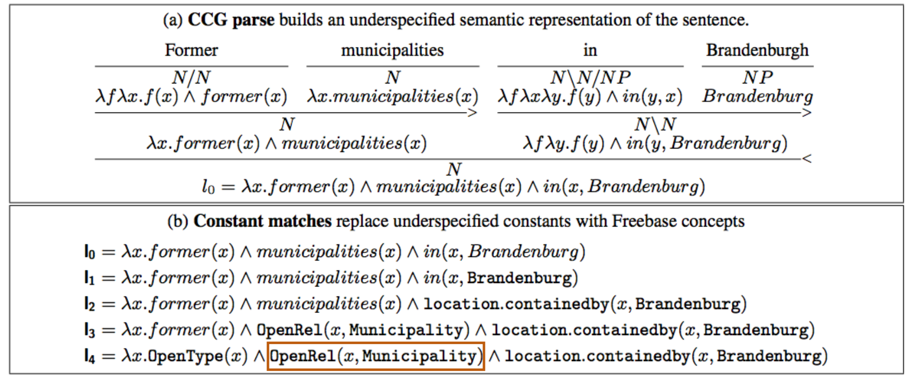
`zooInCity`(*Cincinatti Zoo, Cincinatti*)

The Cincinnati Zoo is located north of downtown Cincinnati

`Zoo` `City`

Mitchell et al. (2015)

QA from Open IE



- Combine open IE with Freebase for question answering

Choi et al. (2015)

Takeaways

- Relation extraction: well-defined task for specific relations, can collect data with distant supervision
- Slot filling: tied to a specific ontology, can be complex and needs annotated data
- 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