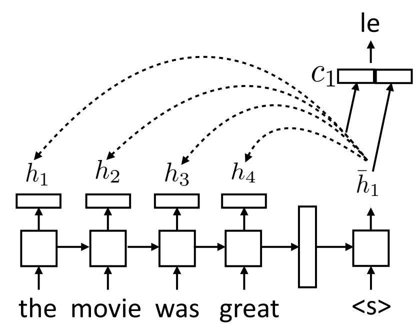
CS388: Natural Language Processing
Lecture 16: Information Extraction

Greg Durrett

Recall: Attention

- For each decoder state, compute weighted sum of input states
- Very helpful for seq2seq models

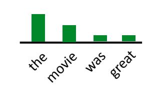


$$P(y_i|\mathbf{x}, y_1, \ldots, y_{i-1}) = \mathrm{softmax}(W[c_i; \bar{h}_i])$$

$$c_i = \sum_j \alpha_{ij} h_j$$

- Weighted sum of input hidden states (vector)

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})}$$

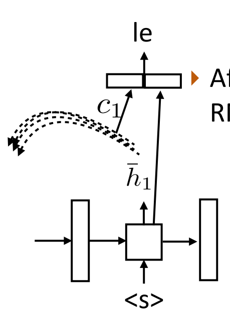


$$e_{ij} = f(\bar{h}_i, h_j)$$

- Unnormalized scalar weight

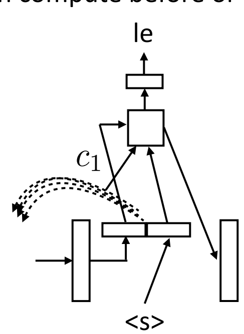
Recall: Alternatives

- When do we compute attention? Can compute before or after RNN cell



- After RNN cell

Luong et al. (2015)



- Before RNN cell; this one is a little more convoluted and less standard

Bahdanau et al. (2015)

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

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 ∃*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

∃*e* stabs(*e*, Brutus, Caesar) ∧ with(*e*, knife) ∧ location(*e*, theater)
∧ time(*e*, Ides of March)

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

some clever driver in America

∃*x* driver(*x*) ∧ clever(*x*) ∧ location(*x*, America)

Example credit: Asad Sayeed

Real Text

which afternoon? 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.

??? which Tuesday?

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

- 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! 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), …
- 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

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)

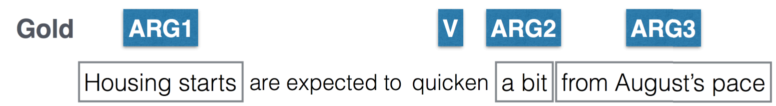
IE: The Big Picture

- 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

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



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

- 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

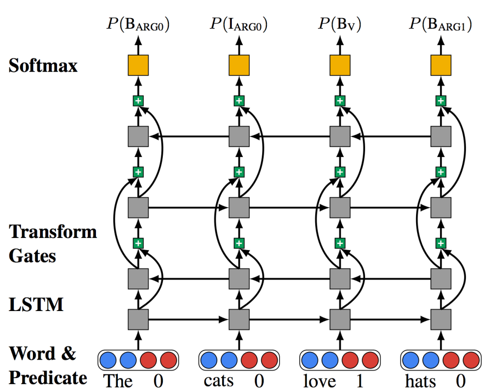


Figure from He et al. (2017)

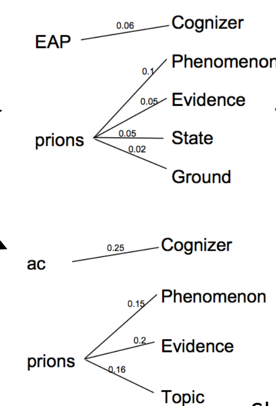
SRL for QA

- 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…*



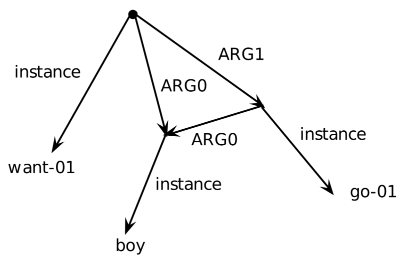
Score by matching expected answer phrase (EAP) against answer candidate (AC)

Shen and Lapata (2007)

Abstract Meaning Representation

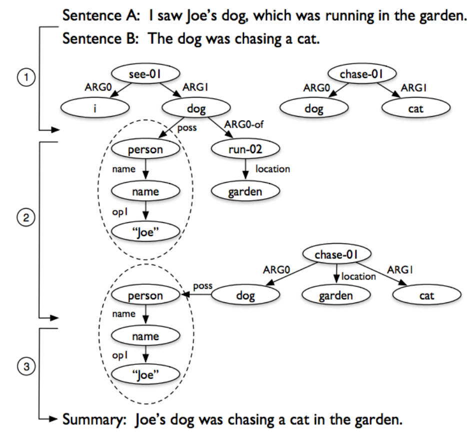
Banarescu et al. (2014)

- 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



- 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

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

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 Exploits i need fud .

(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

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)

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

Director

[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]

Director

Mintz et al. (2009)

Distant Supervision

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

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)

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.

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

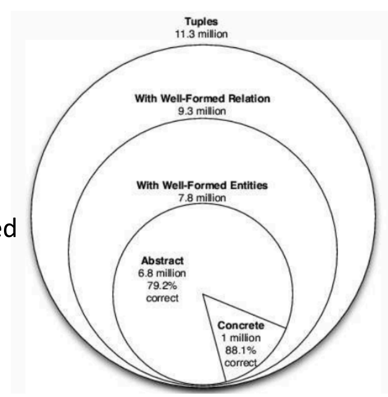
=> Barack_Obama, was born in, Honolulu

- 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

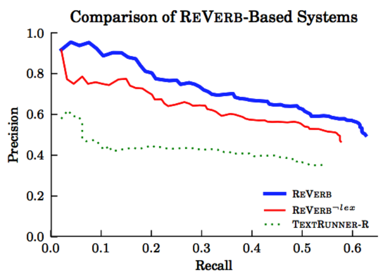
- More constraints: open relations have to begin with verb, end with preposition, be contiguous (e.g., *was born on*)
- Extract more meaningful relations, particularly with light verbs

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)

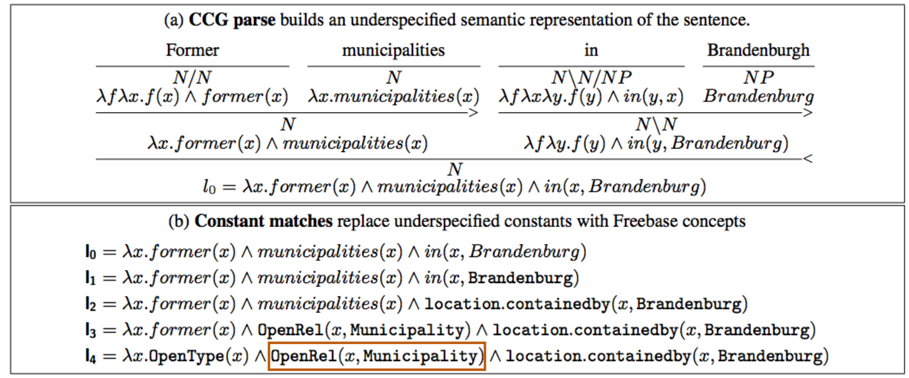
ReVerb

- 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)

QA from Open IE



Choi et al. (2015)

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