Beyond TREC-QA II

Ling573
NLP Systems and Applications
May 29, 2014

Roadmap

- Beyond TREC Question Answering
  - Distant supervision for web-scale relation extraction
  - Machine reading: Question Generation

New Strategy

- Distant Supervision:
  - Supervision (examples) via large semantic database
- Key intuition:
  - If a sentence has two entities from a Freebase relation,
  - they should express that relation in the sentence
- Secondary intuition:
  - Many witness sentences expressing relation
  - Can jointly contribute to features in relation classifier
- Advantages: Avoids overfitting, uses named relations

Freebase

- Freely available DB of structured semantic data
  - Compiled from online sources
    - E.g. Wikipedia infoboxes, NNDB, SEC, manual entry
- Unit: Relation
  - Binary relations between ordered entities
    - E.g. person-nationality: <John Steinbeck, US>
- Full DB: 116M instances, 7.3K rels, 9M entities*
- Largest relations: 1.8M inst., 102 rels, 940K entities*

As of paper publication (2009): currenty 43M topics; 2.5B 'facts'
Largest 14M

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin

Basic Method

- Training:
  - Identify entities in sentences, using NER
  - If find two entities participating in Freebase relation,
    - Extract features, add to relation vector
  - Combine features by rel'n across sent. in multiclass LR
- Testing:
  - Identify entities with NER
  - If find two entities in sentence together
    - Add features to vector
  - Predict based on features from all sents
    - Pair appears 10x, 3 features ➔ 30 features

Examples

- Exploiting strong info: Location-contains:
  - Freebase: <Virginia,Richmond>,<France,Nantes>
  - Training sentences: 'Richmond, the capital of Virginia'
    - 'Edict of Nantes helped the Protestants of France'
  - Testing: 'Vienna, the capital of Austria'
- Combining evidence: <Spielberg, Saving Private Ryan>
  - [Spielberg]'s film, [Saving Private Ryan] is loosely based...
    - Director? Writer? Producer?
  - Award winning [Saving Private Ryan] , directed by [Spielberg]
    - CEO? (Film-)Director?
  - If see both ➔ Film-director

Feature Extraction

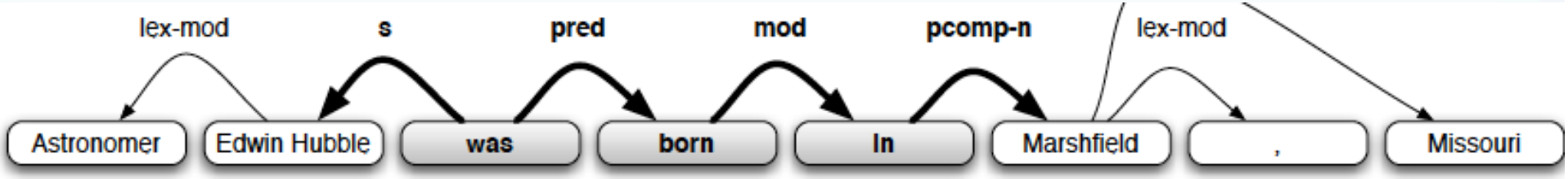
- Lexical features: Conjuncts of
  - Sequence of words between entities
  - POS tags of sequence between entities
  - Flag for entity order
  - k words+POS before 1st entity
  - k words+POS after 2nd entity

  - Astronomer Edwin Hubble was born in Marshfield,MO

Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]

Feature Extraction II

- Syntactic features: Conjuncts of:
  - Dependency path between entities, parsed by Minipar
    - Chunks, dependencies, and directions
  - Window node not on dependency path



Syntactic	[]	PER	$[\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}]$	LOC	[]
Syntactic	[Edwin Hubble $\downarrow_{lex-mod}$]	PER	$[\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}]$	LOC	[]
Syntactic	[Astronomer $\downarrow_{lex-mod}$]	PER	$[\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}]$	LOC	[]
Syntactic	[]	PER	$[\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}]$	LOC	[$\downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\downarrow_{lex-mod}$]	PER	$[\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}]$	LOC	[$\downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\downarrow_{lex-mod}$]	PER	$[\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}]$	LOC	[$\downarrow_{lex-mod}$,]

High Weight Features

- Features highly specific: Problem?
  - Not really, attested in large text corpus

'author_editor	LEX↶	ORG	s novel	PER
	SYN	PER	$\Uparrow_{nn}$ series $\Downarrow_{gen}$	PER
founders	LEX	ORG	co - founder	PER
	SYN	ORG	$\Uparrow_{nn}$ owner $\Downarrow_{person}$	PER
place_founded	LEX↶	ORG	- based	LOC
	SYN	ORG	$\Uparrow_{s}$ founded $\Downarrow_{mod}$ in $\Downarrow_{pcn}$	LOC

Evaluation Paradigm

- Train on subset of data, test on held-out portion
- Train on all relations, using part of corpus
  - Test on new relations extracted from Wikipedia text
  - How evaluate newly extracted relations?
    - Send to human assessors
    - Issue: 100s or 1000s of each type of relation
      - Crowdsource: Send to Amazon Mechanical Turk

Results

- Overall: on held-out set
  - Best precision combines lexical, syntactic
  - Significant skew in identified relations
    - @100,000: 60% *location-contains,* 13% *person-birthplace*
  - Syntactic features helpful in ambiguous, long-distance
  - E.g.
    - Back Street is a 1932 film made by Universal Pictures, directed by John M. Stahl,...

Human-Scored Results

- @ Recall 100: Combined lexical, syntactic best
  - @1000: mixed

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

Distant Supervision

- Uses large database as source of true relations
- Exploits co-occurring entities in large text collection
- Scale of corpus, richer syntactic features
  - Overcome limitations of earlier bootstrap approaches
- Yields reasonably good precision
  - Drops somewhat with recall
  - Skewed coverage of categories

Roadmap

- Beyond TREC Question Answering
  - Distant supervision for web-scale relation extraction
  - Machine reading: Question Generation

Question Generation

- Mind the Gap: Learning to Choose Gaps for Question Generation
  - Becker, Basu, Vanderwende '12
  - Other side of question-answering
    - Related to "machine reading"
    - Generate questions based on arbitrary text

Motivation

- Why generate questions?
  - Aside from playing Jeopardy!, of course
- Educational (self-)assessment
  - Testing aids in retention of studied concepts
    - Reading, review relatively passive
    - Active retrieval, recall of concepts more effective
  - Shifts to less-structured learning settings
    - Online study, reading, MOOCs
  - Assessment difficult ➔ automatically generate

Generating Questions

- Prior work:
  - Shared task on question generation
  - Focused on:
    - Grammatical question generation
    - Creating distractors for multiple choice questions
- Here: Generating *Good* Questions
  - Given a text, decompose into two steps:
    - What sentences form the basis of good questions?
    - What aspects of these sentences should we focus on?

Overview

- Goal: Generate good gap-fill questions
  - Allow focus on content vs form
  - Later extend to wh-questions or multiple choice
- Approach
  - Summarization-based sentence selection
  - Machine learning-based gap selection
    - Training data creation on Wikipedia data
  - Preview: 83% true pos vs 19% false positives

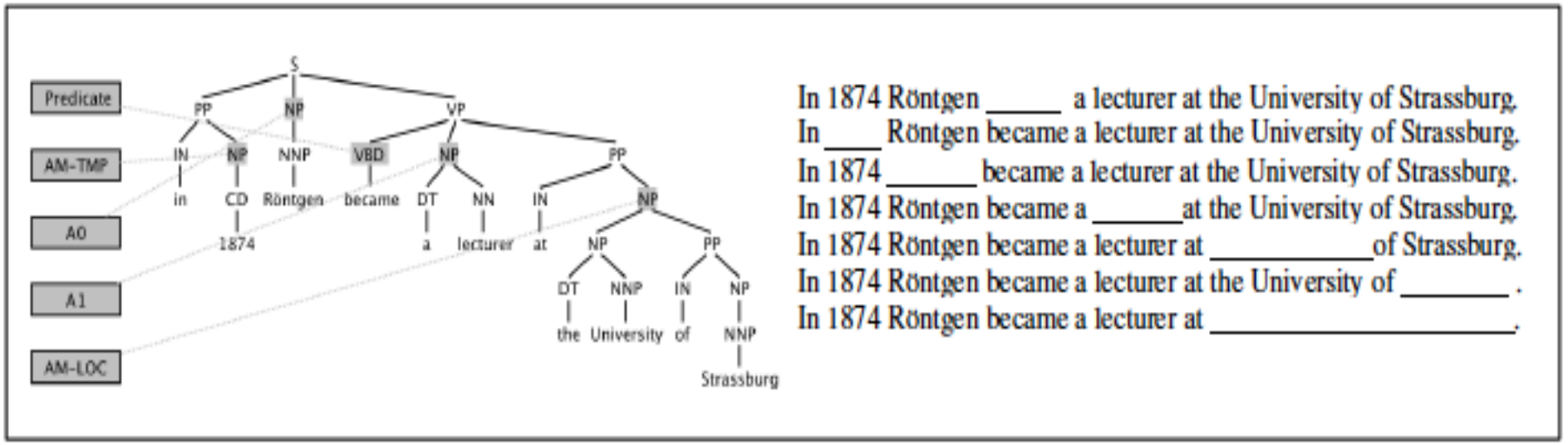
Sentence Selection

- Intuition:
  - When studying a new subject, focus on main concepts
    - Nitty-gritty details later
- Insight: Similar to existing NLP task ➔
  - Summarization: Pick out key content first
- Exploit simple summarization approach
  - Based on SumBasic
  - Best sentences: most representative of article
    - ~ Contain most frequent (non-stop) words in article

Generate

- What's a good gap?
  - Any word position?
    - No – stopwords: only good for a grammar test
    - No – bad for training: too many negative examples
  - Who did what to whom…
    - Items filling main semantic roles:
      - Verb predicate, child NP, AP
    - Based on syntactic parser, semantic role labeling

Processing Example



Classification

- Identify positive/negative examples
  - Create single score
  - > Threshold ➔ positive, o.w. negative
- Extract feature vector
- Train logistic regression learner

Data

- Documents:
  - 105 Wikipedia articles listed as vital/popular
    - Across a range of domains
- Sentences:
  - 10 selected by summarization measure
  - 10 random: why? ➔ avoid tuning
- Labels:?
  - Human judgments via crowdsourcing
  - Rate questions (Good/Okay/Bad)– alone? not reliable
    - In context of original sentence and other alternatives

HIT Set-up

- Good: key concepts, fair to answer
- Okay: key concepts, odd/ambiguous/long to answer
- Bad: unimportant, uninteresting
- Good = 1; o.w. 0

Source Sentence:
The large scale production of chemicals was an important development during the Industrial Revolution.

Question	Answer	Ratings
The _ _ _ _ _ _ _ _ _ _ _ _ of chemicals was an important development during the Industrial Revolution.	large scale production	○Good ◉ Okay
The large scale production of _ _ _ _ _ _was an important development during the Industrial Revolution.	chemicals	○Good ◉ Okay
The large scale production of chemicals was an important development during the _ _ _ _ _ _ _ _ _ _ _ _ .	Industrial Revolution	◉ Good ○Okay

Another QA

- Maintaining quality in crowdsourcing
- Basic crowdsourcing issue:
  - Malicious or sloppy workers
- Validation:
  - Compare across annotators:
    - If average > 2 stdev from median ➔ reject
  - Compare questions:
    - Exclude those with high variance (>0.3)
- Yield: 1821 questions, 700 “Good”: alpha = 0.51

Features

- Diverse features:
  - Count: 5
  - Lexical: 11
  - Syntactic: 112
  - Semantic role: 40
  - Named entity: 11
  - Link: 3
- Characterizing:
  - Answer, source sentence, relation b/t

Features

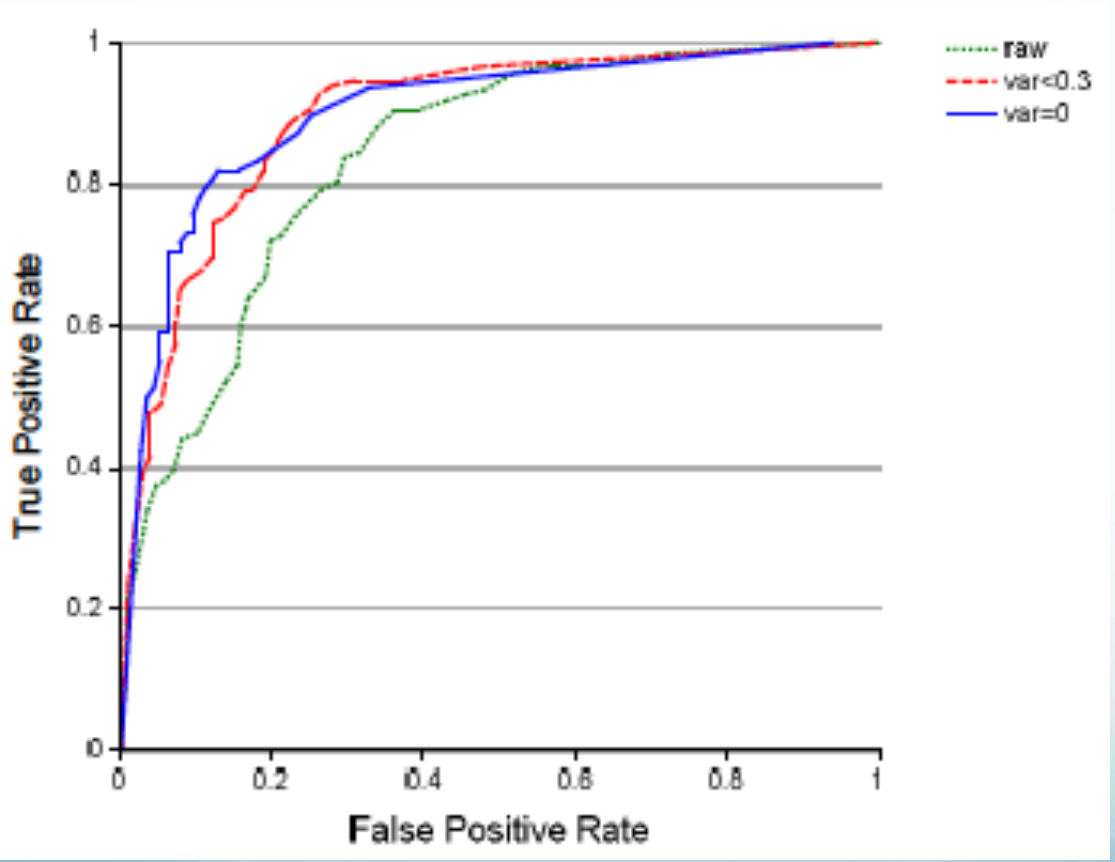
- Token count features:
  - Questions shouldn't be too short
  - Answer gaps shouldn't be too long
    - Lengths, overlaps
- Lexical features:
  - Specific words? ➔ Too specific
  - Word class densities: Good and bad
    - Capitalized words in answer; pronoun, stopwords in ans.

Features

- Syntactic features:
  - Syntactic structure: e.g. answer depth, relation to verb
  - POS: composition, context of answer
- Semantic role label features:
  - SRL spans relations to answer
    - Primarily subject, object roles
    - Verb predicate: strong association, but BAD
- Named entity features:
  - NE density, type frequency in answer; frequency in sent
  - Likely in question, but across types, majority w/o
- Link features: cue to importance of linked spans
  - Density and ratio of links

Results

- Equal error rate: 83% TP, 19% FP



Observations

- Performs well, can tune to balance errors
- Is it Wikipedia centric?
  - No, little change w/o Wikipedia features
- How much data does it need?
  - Learning curve levels out around 1200 samples
- Which features matter?
  - All types, feature weights distributed across types

False Positives

	Question	Answer	Confidence
1	*In 1821 the Greeks declared ___ on the sultan.*	*war*	*0.732*
2	*He also introduced much of the modern mathematical terminology and notation, particularly for mathematical analysis, such as _____ of a mathematical function.*	*the notion*	*0.527*
3	*Not only is there much ice atop _____, the volcano is also being weakened by hydrothermal activity.*	*the volcano*	*0.790*

False Negatives

	Question	Answer	Confidence
1	*Caesar then pursued Pompey to Egypt, where Pompey was soon ____.*	*murdered*	0.471
2	*Over the course of decades, individual wells draw down local temperatures and water levels until _____ is reached with natural flows.*	*a new equilibrium*	0.306
3	*About 7.5% of world sea trade is carried via the canal ____.*	*today*	0.119
4	*Asante and Dahomey concentrated on the development of "legitimate commerce" in _____, forming the bedrock of West Africa's modern export trade,*	*the form of palm oil, cocoa, timber, and gold.*	0.029

Error Analysis

- False positives:
  - Need some notion of predictability, repetition
- False negatives:

Look Back

- TREC QA vs
  - Jeopardy!, web-scale relation extraction, question gen.
- Obvious differences:

Looking Back

- TREC QA vs
  - Jeopardy!, web-scale relation extraction, question gen.
- Obvious differences:
  - Scale, web vs fixed documents, question-answer direction
  - Task-specific constraints: betting, timing, evidence,…
- Core similarities:
  - Similar wide range of features applied
  - Deep + shallow approaches; learning&rules
  - Relations b/t question/answer (pattern/relation)