

Information extraction

7. Relation extraction II

Simon Razniewski
Winter semester 2019/20

Announcements

- Results assignment 5 online
- Evaluation on cleaned data
→ scores moderately higher than on provided data
- Many great solutions
- Caution
 - Make sure code runs
 - 6 passed assignments
→ All assignments are exam-relevant ☺

Lab 06 Ranking	
2576572	0.581
2550309	0.553
2562559	0.548
2568227	0.532
2579810	0.532
2571663	0.506
2576861	0.506
2572706	0.483
2561347	0.469
2550421	0.447
2576610	0.427
2576796	0.409
2558667	0.405
2565094	0.396
2576611	0.394
2571656	0.364
2571690	0.318
2553344	0.317
2558462	0.303
2576381	0.286
2568101	0.285
2564409	0.264
2570975	0.241
2548617	0.237
2581370	0.197
2572758	0.195
2581455	0.09
2576612	0
2576770	0

Design, implementation, comments:

1. Extracting Date of Birth: function extractDoB

- Design

Given our restricted domain of Wikipedia abstracts, it was surprisingly straightforward to achieve an f1 score of ~80% just by extracting the very first date in the abstract.

- Implementation

The function uses a regex (dateMatcher regex ref: <https://stackoverflow.com/questions/51122413/>) in order to extract the date and returns it in the right format.

- Comments

This method is admittedly crude, and it can be further improved by using either text extracted in parentheses right after the entity mention and/or look for the keyword 'born' followed by the date.

2. Extracting Nationality: function extractNationality

- Design

It was observed that most entities are mentioned with their nationalities such as 'Wayne A. Hendrickson (born April 25, 1941, New York City) is an American biophysicist and University professor at Columbia.' which was matched.

In case that returns no candidate, the verb 'born' is looked for in the abstract and when found, it's prepositional objects are extracted. Those objects that are in fact dates such as 'born in __1955__' are discarded and the rest are returned.

- Implementation

Dependency parsing and ner using spacy.

- Comments

Most nationalities appearing are of demonyms, and the expected nationality (loosely) are country names, a dict of demonym-country has been constructed using data provided in the following link: <https://github.com/knowitall/chunkedextractor/blob/master/src/main/resources/edu/knowitall/chunkedextractor/demonyms.csv>. Credits: Jesujoba ALABI for having discussed it on the IE1920 forum.

3. Extracting alma mater: function `extractAlmaMater`

- Design

The function looks for the following patterns:

studied <something> at <alma_mater>

attended <alma_mater>

[was] obtained/received/awarded/gained/earned/complete/graduated/educated <something> from/at <alma_mater>

and just extracts the alma maters if 'alma_mater' is at least one among 'university', 'school', 'college', 'academy', or 'gynmasium'.

- Implementation

POS tagging, dependency parsing and ner using spacy.

4. Extracting places of work: function `extractWorkPlace`

- Design

This turned out to be quite the challenge with a morass of exceptions. Hence the function takes an overly simplifying approach of extracting all of the organizations mentioned in the abstract apart from alma maters and returns.

5. Extracting awards: `extractAwards`

- Design

Looks for verbs 'won' and 'awarded' and returns the objects.

In order to improve recall, this function makes the assumption that most awards mentioned in the abstract probably belong to the entity in question and hence extracts all of them using a regex that matches 'prize', 'award', 'medal' and returns. The first rule compensates for all those awards that don't get matched by the regex such as 'Spinozapremie'.

- Implementation

Dependency parsing, ner, regex matching

General comments

There seems to be an upper bound on the scores as the ground truth itself is quite noisy.

It is observed that for this restricted domain, given enough time, manual pattern matching can indeed return good enough results, there aren't too many exceptions to warrant a statistical models.

Outline

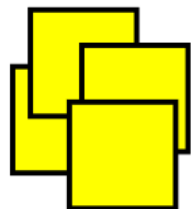
1. Problem
2. Manual patterns
3. Supervised learning
 1. Feature-based
 2. TACRED and BERT
4. Semi- and unsupervised extraction
 1. Iterative pattern learning (DIPRE)
 2. Distant supervision
 - CINEX
5. Evaluation
6. OpenIE
 1. PATTY
 2. Quasimodo
7. Negation

How not to design an IE algorithm



Task: Find Simpson pets

Corpus:



Algorithm: Regex: "Snowball (IIV)*"

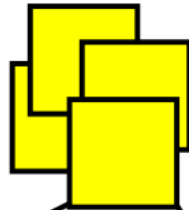
Output: {Snowball I, Snowball II, Snowball IV}

Is this algorithm good?

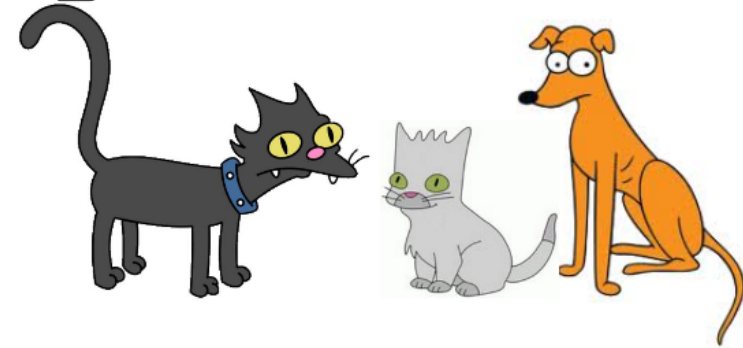
How to design an IE algorithm

Task: Find Simpson pets

Corpus:



Take only a sample
of the corpus

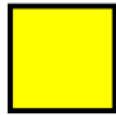


Lisa decides to play music on her saxophone for Coltrane, but the noise frightens him and he commits suicide. As Gil swerves to avoid hitting Snowball V, his car hits a tree and bursts into flames. Since the cat is unhurt, Lisa takes it as a sign of good luck and adopts her. [...]

How to design an IE algorithm

Task: Find Simpson pets

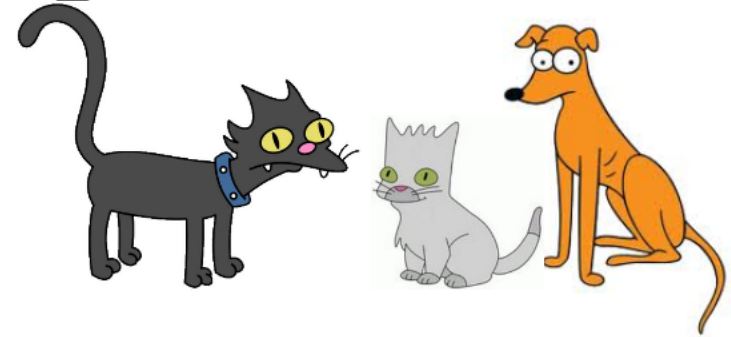
Corpus:



Consider only
the sample corpus.

Gold Standard:
{Coltrane, Snowball I, ...}

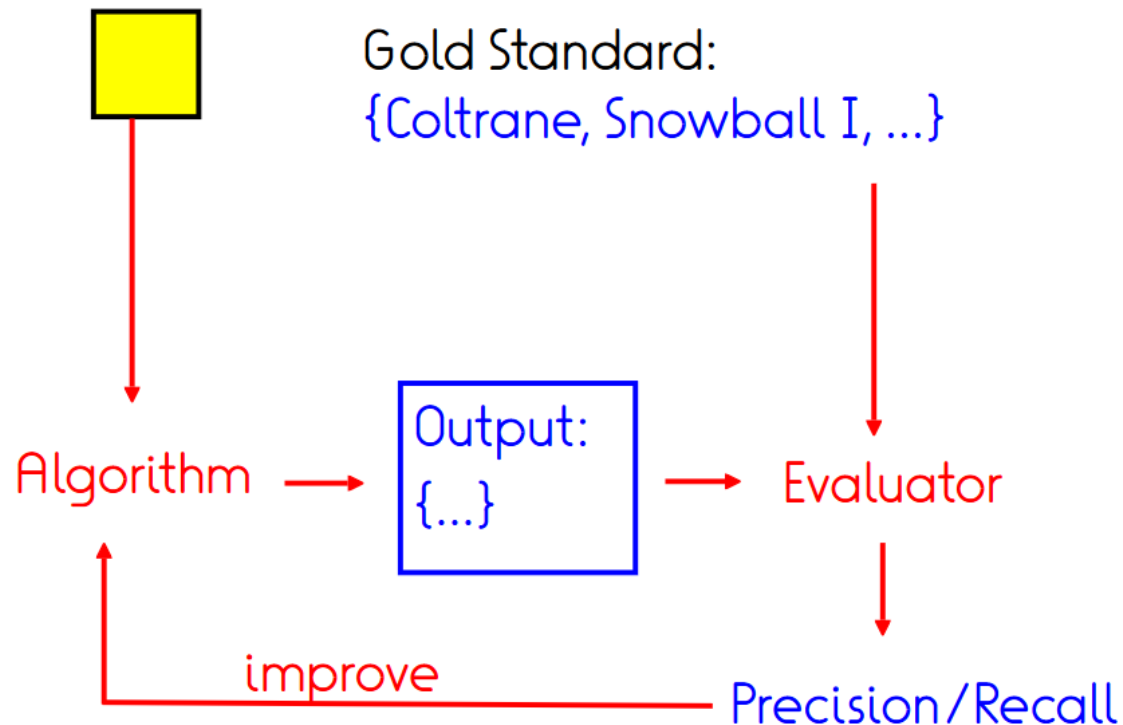
Manually make
a gold standard



How to design an IE algorithm

Task: Find Simpson pets

Corpus:



Def: Problem of imbalanced classes

Population: {Snowball_1,..., Snowball_99, Snowball_100}

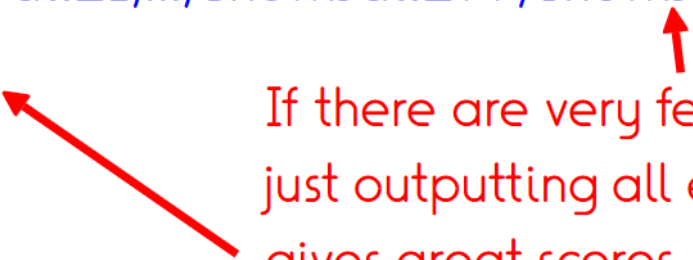
Gold Standard: {Snowball_1,..., Snowball_99}

Output: {Snowball_1,..., Snowball_99, Snowball_100}

Precision: $99/100=99\%$

Recall: $99/99=100\%$

If there are very few negatives,
just outputting all elements
gives great scores.



The problem of **imbalanced classes** appears when only very few of the items of the population are not in the gold standard: An approach that outputs the entire population has a very high precision and a perfect recall. (Example: Citizenship on en-Wikipedia)

The **negatives** are the elements of the population that are not in the gold standard.

Def: Confusion Matrix

Population: {Snowball_1,..., Snowball_99, Snowball_100}

Gold Standard: {Snowball_1,..., Snowball_99}

Output: {Snowball_1,..., Snowball_99, Snowball_100}

The **confusion matrix** for the output of an algorithm looks as follows:

		Gold standard		
		Positive	Negative	Σ
Output	Positive	True Positives	False Positives	Predicted Positives
	Negative	False Negatives	True Negatives	Predicted Negatives
Σ		(Gold) Positives	(Gold) Negatives	

Items of the population that are not in the gold standard

Items of the population that are not output

"Negative" because it was not output, "True" because that was correct.

Def: Confusion Matrix

Population: {Snowball_1,..., Snowball_99, Snowball_100}

Gold Standard: {Snowball_1,..., Snowball_99}

Output: {Snowball_1,..., Snowball_99, Snowball_100}

The **confusion matrix** for the output of an algorithm looks as follows:

		Gold standard		
		Positive	Negative	
Output	Positive	99	1	100
	Negative	0	0	0
		99	1	

1 item was output as positive, but was negative in the gold standard

Precision = true positives / predicted positives = $99/100 = 99\%$

Recall = true positives / gold positives = $99/99 = 100\%$

Confusion with confusion matrixes

A confusion matrix does not always make sense in an information extraction scenario:

Population: {H, Ho, Hom, ..., o, om, ome, ..., r Sim, r Simps, ...}

Gold Standard: {Homer}

Output: {Homer}

		Gold standard	
		Positive	Negative
Output	Positive	1	0
	Negative	0	39462440205

A confusion matrix makes sense only when the population is limited (e.g., in classification tasks)!


Our problem

Population: {Snowball_1,..., Snowball_99, Snowball_100}

Gold Standard: {Snowball_1,..., Snowball_99}

Output: {Snowball_1,..., Snowball_99, Snowball_100}

		Gold standard	
		Positive	Negative
Output	Positive	99	1
	Negative	0	0



The problem is that the algorithm did not catch the negatives, it has a "low recall" on the negatives.

Def: True Negative Rate & FPR

Population: {Snowball_1,..., Snowball_99, Snowball_100}

Gold Standard: {Snowball_1,..., Snowball_99}

Output: {Snowball_1,..., Snowball_99, Snowball_100}

The **true negative rate** (also: TNR, specificity, selectivity) is the ratio of negatives that are output as negatives (= the recall on the negatives):

$$\text{TNR} = \text{true negatives} / \text{gold negatives} = 0 / 1 = 0\%$$

		Positive	Negative
Output	Positive	99	1
	Negative	0	0

The **False Positive Rate** (also: FPR, fall-out) is $1 - \text{TNR}$.

TNR & Precision

Population: {Snowball_1,..., Snowball_99, Snowball_100}

Gold Standard: {Snowball_1,..., Snowball_99}

Output: {Snowball_1,..., Snowball_99, Snowball_100}

Precision: $99/100=99\%$ TNR: $0/1=0\%$

Recall: $99/99=100\%$

TNR and precision both measure the “correctness” of the output.

Precision:

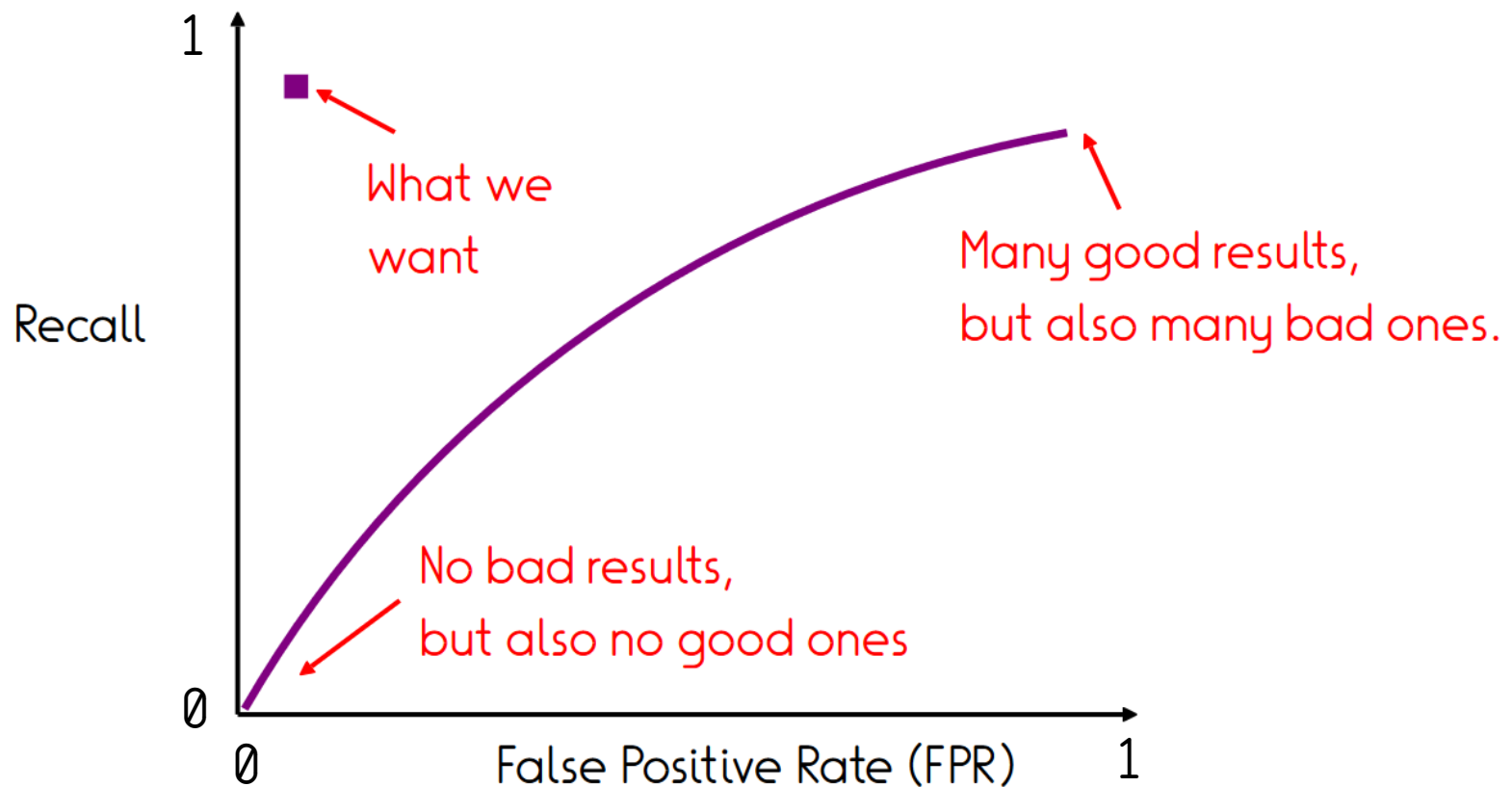
- measures wrt. the output
- suffers from imbalanced classes
- works if population is infinite
(e.g., set of all extractable entities)

TNR:

- measures wrt. the population
- guards against imbalance
- works if population is limited
(e.g., in classification)

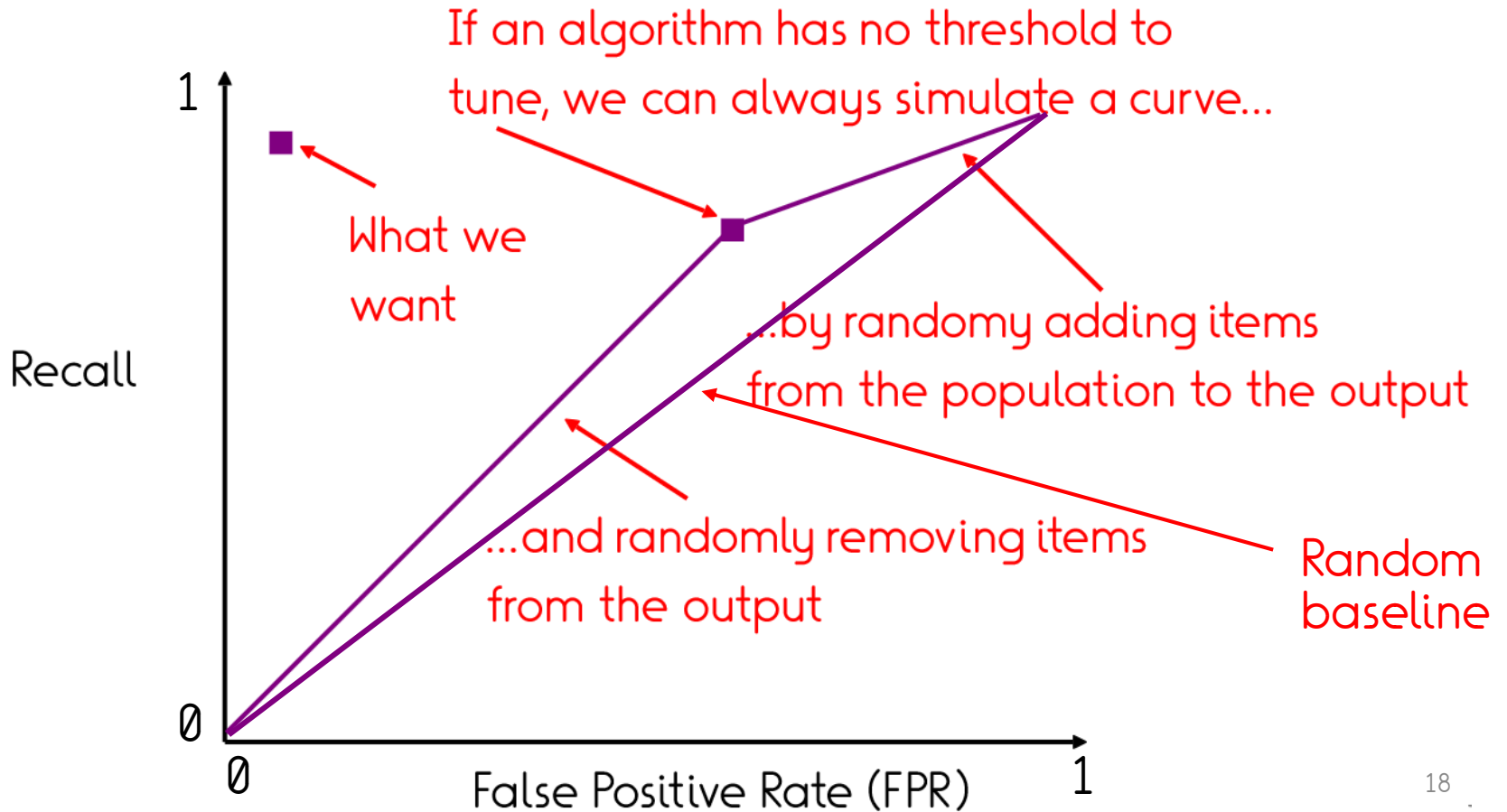
Def: ROC

The **ROC** (receiver operating characteristic) curve plots recall against the FPR for different thresholds of the algorithm. It guards against imbalanced classes, and is applicable when the population is finite.



Def: ROC

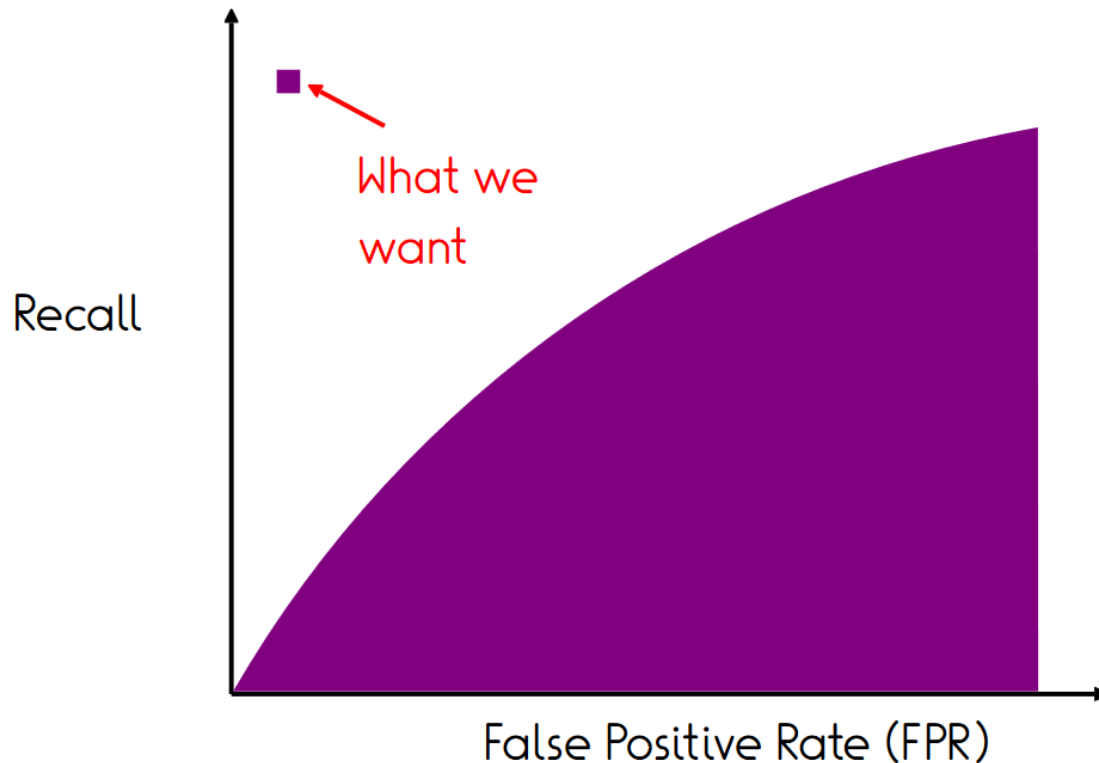
The **ROC** (receiver operating characteristic) curve plots recall against the FPR for different thresholds of the algorithm. It guards against imbalanced classes, and is applicable when the population is finite.



Def: AUC

The **AUC** (area under curve) is the area under the ROC curve.

It corresponds to the probability that the classifier ranks a random positive item over a random negative item. (It's kind of the F1 for a limited population and a varying threshold.)



(AUC measure for PR curves also exists, but has no corresponding probabilistic interpretation)

Def: Micro vs. Macro averaging

- 3 relations (A, B, C)
- Predictions:
 - 10x A (90% correct)
 - 10x B (90% correct)
 - 100x C (10% correct)
- **Micro-avg.** precision: $\frac{10 \times 0.9 + 10 \times 0.9 + 100 \times 0.1}{10 + 10 + 100} = 0.23$
- **Macro-avg.** precision: $\frac{0.9 + 0.9 + 0.1}{3} = 0.63$
- Recall and F1 analogous
- Macro gives tail equal importance

Evaluation of Semi-supervised and Unsupervised Relation Extraction

- Since its intended to extract **totally new statements**
 - Gold set is difficult to prepare
 - Can't compute precision (don't know which ones are correct)
 - Can't compute recall (don't know which ones were missed)
- Instead, we can approximate precision (only)
 - Draw a **random sample of statements from output**, check precision manually

$$\hat{p} = \frac{\text{\# of correctly extracted relations in the sample}}{\text{Total \# of extracted relations in the sample}}$$

- Can also compute **precision at different levels of recall**.
 - Precision for top 1000 new relations, top 10,000 new relations, top 100,000
 - In each case taking a random sample of that set
- But **no realistic way to evaluate recall**

Baselines and yardsticks

- Method: Precision 0.63, recall 0.47, ???
- Baselines
 - Random!
 - Most frequent class!
 - Naive heuristics
 - Trigger word lookup, first noun, 5th word, etc.
- Yardsticks
 - Existing systems
 - Human performance (agreement)
 - (in certain tasks e.g. in vision not a yardstick anymore)

Error analysis (1/3)

- Method: P 0.63 R 0.47
- Baseline: P 0.55 R 0.30
- Humans: P 0.85 R 0.90



- What went wrong?
 - Sample a few errors (false positives and false negatives)
 - Define categories of errors
 - Sample a larger set of errors
 - Count frequencies of error categories
 - Possibly iterate
- Severity of errors?
- Important for
 - Yourself to improve
 - The next one continuing your concrete work
 - Others to understand potential and limits of your approach
- Error meta-categories
 - Limit of effort (effort-performance-derivative/extrapolation?)
 - Limits of methodology
 - Limit of data/metric (next)

Error analysis (2/3) – Question the data

- Data too often with issues
 - Typing assignment: Vocabulary mismatch
 - Relation extraction assignment: Nationalities that are not nationalities
- Semiautomatic data:
 - Systematic errors
- Crowdsourced data:
 - Random noise
- ...



Error analysis (3/3) – Question the rules

- Evaluation metric design not trivial
 - Machine translation and summarization: BLEU
 - Named entity recognition, OpenIE: Partial matches?
 - Typing: Metrics aware of error severity?
 - Disambiguation: Plausible vs. semantically impossible mismatches



(FIFA congress)

How to get gold data?

- Self-annotation
 - Alone or in a team of few researchers, colleagues
 - Confirmation bias
 - Generally discouraged
- Creative reuse of existing data
 - E.g., Wikipedia text links for entity disambiguation
 - Synchronous edits of Wikidata relation and texts
 - Usually still shaky/biased
- Paid annotators
 - Can be known local personnel
 - More often, anonymous online crowdsourcing
 - De-facto standard nowadays

Crowdsourcing

- Prominent platforms: Amazon Mechanical Turk, Prolific
- Typical pay ~10\$/hour
 - In cases total spending 10k+€ for research datasets
- Requires to-the-point instructions
 - Traditional expert annotations guidelines sometimes >100 pages
 - Complex or open-ended annotation tasks difficult
 - Wherever possible, break into smaller tasks
- Quality assurance:
 - Worker education/background
 - Worker reputation
 - Honeypot/test question-based filtering
 - Redundancy (majority opinion on task)
- Creating good crowd tasks takes iterations and effort!

Relation definitions for **has nationality** and **lived in**

Has nationality: The highlighted location must be either a country where the person has citizenship or an adjective for a country such as "American" or "French". If someone holds a national office or plays for a national sports team, this implies **has nationality**. A person's nationality by itself does not imply the **lived in** or **was born in** relations.

Lived in: Means a person spent time in the highlighted location for more than a visit. You can assume a **lived in** relation for the country of national officials. Otherwise, working in a location does not imply that a person has a **lived in** relation. **lived in** does not imply **has nationality** or **was born in**.

Practice sentence 1 of 5 (select all relations that apply):

- "Vice President Joe Biden met today with **Turkish** Prime Minister **Ahmet Davutoglu**"

Yes No

☐ ☐ has nationality

☐ ☐ lived in

Submit

Figure 3: Tutorial page that teaches guidelines for *nationality* and *lived_in*. The worker answers practice sentences with immediate feedback that teach each relation.

Example benchmark dataset: KnowledgeNet

[Mesquita et al., EMNLP 2019
<https://www.aclweb.org/anthology/D19-1069.pdf>]

- Text: Wikipedia abstracts
 - 15 common person relations
 - 9000 exhaustively annotated sentences
 - Interannotator agreement
 - Relation classification: 96%
 - Entity disambiguation: 93%
 - In-house annotators
 - ~2 minutes/annotator/sentence for one property
 - 22% mention detection, 40% relation classification, 28% entity disambiguation
 - 2 annotators, in case of disagreement third annotator
- Total effort ~ 600 annotator hours

Highlight all organization names in the highlighted passage.

Document:	Passage	Passage	Status:
5288	Start:	End:	1239/1277
	164	234	

Butler W. Lampson (born December 23, 1943) is an American computer scientist contributing to the development and implementation of distributed, personal computing. He is a Technical Fellow at Microsoft and an Adjunct Professor at MIT.

Exit Back Clear Submit

(a) Interface to detect mentions of an entity type.

Are the highlighted mentions a person and its employer?

Document:	Passage	Passage	Status:
5288	Start:	End:	245/246
	164	234	

Butler W. Lampson (born December 23, 1943) is an American computer scientist contributing to the development and implementation of distributed, personal computing. He is a Technical Fellow at Microsoft and an Adjunct Professor at MIT.

☒ He works or has worked at Microsoft

☐ He does/did not work at Microsoft

Exit Clear Submit

(b) Interface to classify facts.

Choose the correct Wikidata entry for the highlighted entity.

Document:	Passage	Passage	Status:
5288	Start:	End:	1035/1228
	192	201	

Butler W. Lampson (born December 23, 1943) is an American computer scientist contributing to the development and implementation of distributed, personal computing. He is a Technical Fellow at Microsoft and an Adjunct Professor at MIT.

Link to primary entity?

Search Microsoft

Selected: Microsoft - American multinational technology corporation

Microsoft
American multinational technology corporation

Microsoft Windows
family of operating systems produced for personal computers, servers, smartphones and embedded devices

Microsoft
1118th strip of the webcomic xkcd

Exit Back Clear Submit

(c) Interface to link a mention to a Wikidata entity.

Instructive pipeline implementations

- Mention detection, coreference resolution, relation classification, entity linking
- Human performance as comparison

System	Text evaluation			Link evaluation			
	P	R	F1	P	R	F1	
Stanford TAC KBP + coreference + entity types + ... + BERT	Baseline 1	0.44	0.64	0.52	0.31	0.26	0.28
	Baseline 2	0.49	0.64	0.55	0.37	0.32	0.34
	Baseline 3	0.47	0.66	0.55	0.35	0.37	0.36
	Baseline 4	0.60	0.65	0.62	0.51	0.48	0.49
	Baseline 5	0.68	0.70	0.69	0.53	0.48	0.50
Human	0.88	0.88	0.88	0.81	0.84	0.82	

Text spans of S and O match
vs. KB links match

Outline

1. Problem
2. Manual patterns
3. Supervised learning
 1. Feature-based
 2. TACRED and BERT
4. Semi- and unsupervised extraction
 1. Iterative pattern learning (DIPRE)
 2. Distant supervision
 - CINEX
5. Evaluation
6. OpenIE
 1. PATTY
 2. Quasimodo
7. Negation

Motivation: Open information extraction

- So far assumed a limited set of fixed relations
- Presumably designed by humans (“ontology engineers”)
- Lessons from DB/KR Research
 - Declarative KR is expensive & difficult
 - Formal semantics is at odds with
 - Broad scope
 - Distributed authorship
 - A “universal ontology” is impossible
 - Global consistency is like world peace
 - Micro ontologies--scale? Interconnections?

Open vs. Traditional IE

	Traditional IE	Open IE
Input:	Corpus + $O(R)$ hand-labeled data	Corpus
Relations:	Specified in advance	Discovered automatically
Extractor:	Relation-specific	Relation- independent

How is Open IE Possible?

Semantic Tractability Hypothesis

∃ *easy-to-understand* subset of English

- Characterized relations/arguments syntactically
[Banko et al. ACL '08]
- Characterization is compact, domain independent
- Covers 80-95% of binary relations in sample corpus

Relative Frequency	Category	Simplified Lexico-Syntactic Pattern
37.8	Verb	E ₁ Verb E ₂ <i>X established Y</i>
22.8	Noun+Prep	E ₁ NP Prep E ₂ <i>X settlement with Y</i>
16.0	Verb+Prep	E ₁ Verb Prep E ₂ <i>X moved to Y</i>
9.4	Infinitive	E ₁ to Verb E ₂ <i>X plans to acquire Y</i>
5.2	Modifier	E ₁ Verb E ₂ Noun <i>X is Y winner</i>

(simplified!)

Reverb [Fader et al., 2011]

Identify **Relations** from **Verbs**.

1. Find longest phrase matching a simple syntactic constraint:

$$V \mid VP \mid VW^*P$$

V = verb particle? adv?

W = (noun | adj | adv | pron | det)

P = (prep | particle | inf. marker)

Sample Reverb relations

invented

acquired by

has a PhD in

denied

voted for

**inhibits tumor
growth in**

inherited

born in

mastered the art of

downloaded

aspired to

**is the patron
saint of**

expelled

Arrived from

wrote the book on

Challenges (1)

- Larry Page, the CEO of Google, talks about multi-screen opportunities offered by Google.
- After winning the Superbowl, the Giants are now the top dogs of the NFL.
- Ahmadinejad was **elected** as the new President of Iran.
- Relation arguments can be overly specific

(“The great R. Feynman”; “worked jointly with”; “F. Dyson”)

Challenges (2)

“John refused to visit Vegas.”



(John, refused to visit, Vegas)

“Early astronomers believed that the earth is the center of the universe.”



[(earth, is the center of, universe) Attribution: early astronomers]


“If she wins California, Hillary will be the nominated presidential candidate.”



[(Hillary, will be nominated, presidential candidate) Modifier: if she wins California]

System evolution

- 2007 Texrunner
 - CRF and self-training
- 2010 ReVerb
 - POS-based patterns
- 2012: OLLIE
 - Dependency-parse based
- 2013: ClausIE
 - Sentence restructuring before dependency parsing
- 2014 OpenIE 4.0
 - SRL-based extraction
- 2016 OpenIE 5.0
 - Compound noun phrases, numbers
- 2017 MinIE
 - Minimizing extractions by removal of minor qualifiers etc.



increasing
precision,
recall,
expressiveness

Texrunner

Inference and tuple correction:

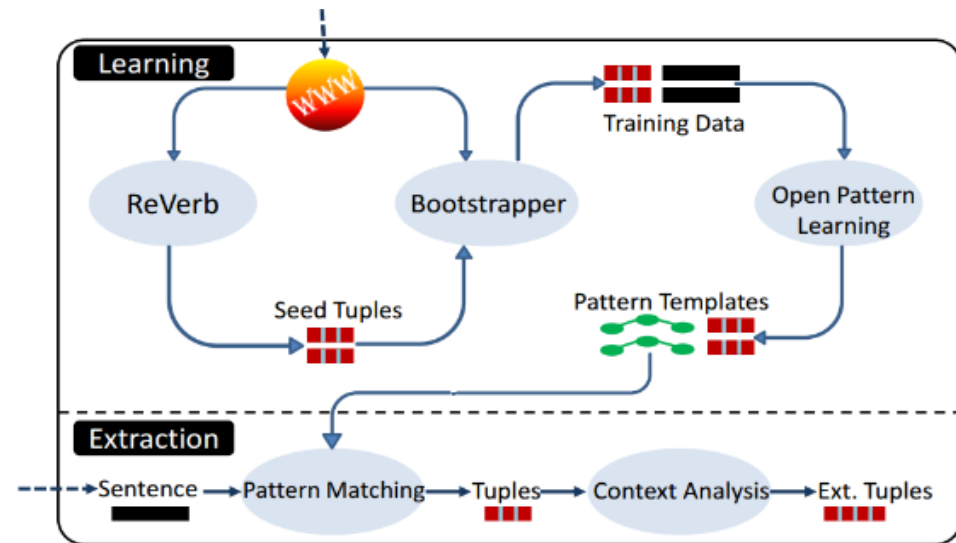
(X, born in, 1941)	(Y, born in, 1941)	
(X, citizen of, US)	(Y, citizen of, US)	→
(X, friend of, Joe)	(Y, friend of, Joe)	P (X = Y) determined by shared relations

(1, R1, 2)	(1, R2, 2)	
(2, R1, 4)	(2, R2, 4)	→
(4, R1, 8)	(4, R2, 8)	P (R1 = R2) determined by shared argument pairs

OLLIE

Learning Open Patterns:

- 1) Extract the high confidence tuples from ReVerb.
- 2) For each tuple, find all sentences in the corpus containing the words in the tuple.
- 3) Using a dependency parser specify the patterns corresponding to each ReVerb tuple selected.



Number of Relations

DARPA MR Domains	<50
NYU, Yago	<100
NELL	~500
DBpedia 3.2	940
PropBank	3,600
VerbNet	5,000
Wikipedia Infoboxes, $f > 10$	~5,000
TextRunner (<i>phrases</i>)	100,000+
ReVerb (<i>phrases</i>)	1,000,000+

<https://openie.allenai.org/>

- Saarland
- Einstein
- Kangaroo
- ...

Semantic role labelling

Can we figure out that these have the **same meaning**?

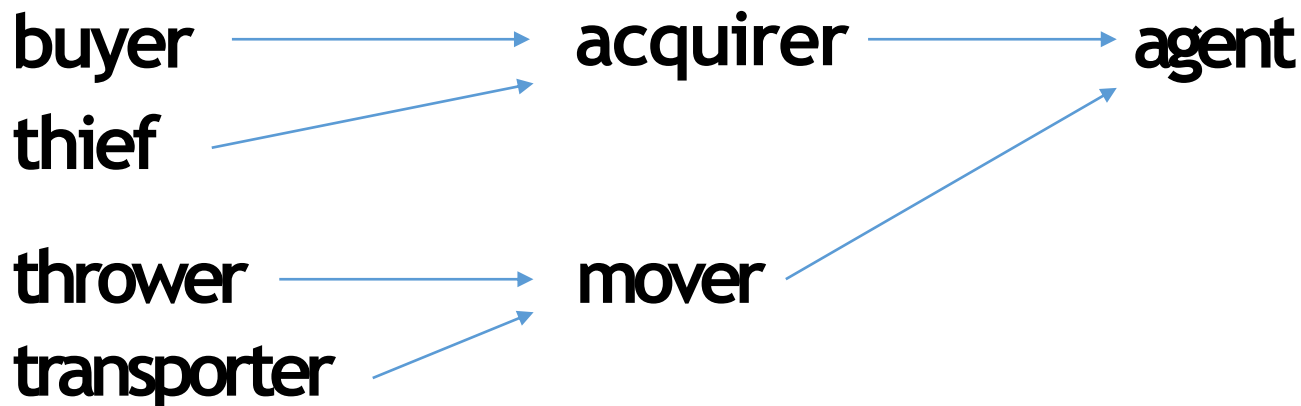
- XYZ corporation bought the stock.
 - They sold the stock to XYZ corporation.
 - The stock was bought by XYZ corporation.
 - The purchase of the stock by XYZ corporation...
 - The stock purchase by XYZ corporation...
-
- How do we **represent** this commonality?

A Shallow Semantic Representation: Semantic Roles

Predicates (bought, sold, purchase) represent an event
semantic roles express the abstract role that arguments of a predicate
can take in the event

More specific

More general



Thematic roles

- Buyer and Thrower have something in common!
 - Volitional actors
 - Often animate
 - Direct causal responsibility for their events
- Thematic roles are a way to capture this semantic commonality between Buyers and Thrower.
- They are both AGENTS.
- The Bought Thing and Thrown Thing, are THEMES.
 - prototypically inanimate objects affected in some way by the action
- One of the oldest linguistic models
 - Indian grammarian Panini between the 7th and 4th centuries BCE

Thematic roles

- A typical set:

Thematic Role	Definition	Example
AGENT	The volitional causer of an event	<i>The waiter</i> spilled the soup.
EXPERIENCER	The experiencer of an event	<i>John</i> has a headache.
FORCE	The non-volitional causer of the event	<i>The wind</i> blows debris from the mall into our yards.
THEME	The participant most directly affected by an event	Only after Benjamin Franklin broke <i>the ice</i> ...
RESULT	The end product of an event	The city built a <i>regulation-size baseball diamond</i> ...
CONTENT	The proposition or content of a propositional event	Mona asked “ <i>You met Mary Ann at a supermarket?</i> ”
INSTRUMENT	An instrument used in an event	He poached catfish, stunning them <i>with a shocking device</i> ...
BENEFICIARY	The beneficiary of an event	Whenever Ann Callahan makes hotel reservations <i>for her boss</i> ...
SOURCE	The origin of the object of a transfer event	I flew in <i>from Boston</i> .
GOAL	The destination of an object of a transfer event	I drove <i>to Portland</i> .

PropBank Frame Files [Palmer et al., 2005]

agree.01

Arg0: Agreeer

Arg1: Proposition

Arg2: Other entity agreeing

Ex1: [Arg0 The group] *agreed* [Arg1 it wouldn't make an offer].

Ex2: [ArgM-TMP Usually] [Arg0 John] *agrees* [Arg2 with Mary]
[Arg1 on everything].

fall.01

Arg1: Logical subject, patient, thing falling

Arg2: Extent, amount fallen

Arg3: start point

Arg4: end point, end state of arg1

Ex1: [Arg1 Sales] *fell* [Arg4 to \$25 million] [Arg3 from \$27 million].

Ex2: [Arg1 The average junk bond] *fell* [Arg2 by 4.2%].

Advantage of a ProbBank Labeling

- **increase.01** “go up incrementally”

Arg0: causer of increase

- Arg1: thing increasing
- Arg2: amount increased by, EXT, or MNR
- Arg3: start point
- Arg4: end point

- This allow to see the **commonalities** in these 3 sentences:

[Arg0 Big Fruit Co.] increased [Arg1 the price of bananas].

[Arg1 The price of bananas] was increased again [Arg0 by Big Fruit Co.]

[Arg1 The price of bananas] increased [Arg2 5%].

QA-SRL [Ido Dagan et al.]

- Formulate roles as natural language questions

UCD **finished** the 2006 championship as Dublin champions ,
by **beating** St Vincents in the final .

finished

Who finished something? - UCD

What did someone finish? - the 2006 championship

What did someone finish something as? - Dublin champions

How did someone finish something? - by beating St Vincents in the final

beating

Who beat someone? - UCD

When did someone beat someone? - in the final

Who did someone beat? - St Vincents

→ Crowd workers write intuitive¹ questions and answers

¹The PropBank annotation guide is 89 pages (Bonial et al., 2010), and the FrameNet guide is 119 pages (Ruppenhofer et al., 2006). Our QA-driven annotation instructions are 5 pages. 51

Supervised OpenIE

[Stanovsky et al., NAACL 2018
<https://www.aclweb.org/anthology/N18-1081>]

- Uses SRL annotations as target and training data
 - ~ Every set of (head, arg0, arg1) corresponds to a triple
- Trains a bi-LSTM to solve OpenIE via sequence labelling

Outline

1. Problem
2. Manual patterns
3. Supervised learning
 1. Feature-based
 2. TACRED and BERT
4. Semi- and unsupervised extraction
 1. Iterative pattern learning (DIPRE)
 2. Distant supervision
 - CINEX
5. Evaluation
6. OpenIE
 1. PATTY
 2. Quasimodo
7. Negation

PATTY

- Resource of 350k synsets of binary relations
- Taxonomical organization
- Key idea: exploit instance overlap/subsumption
- Wikipedia-extractions between two named entities in sentence
- Patterns combine terms, POS tags, types

- Pattern accuracy: 85%
- Subsumption accuracy: 75%

PATTY (2)

ID	Pattern Synset & Support Sets
P_1	$\langle \textit{Politician} \rangle$ was governor of $\langle \textit{State} \rangle$ A,80 B,75 C,70
P_2	$\langle \textit{Politician} \rangle$ politician from $\langle \textit{State} \rangle$ A,80 B,75 C,70 D,66 E,64
P_3	$\langle \textit{Person} \rangle$ daughter of $\langle \textit{Person} \rangle$ F,78 G,75 H,66
P_4	$\langle \textit{Person} \rangle$ child of $\langle \textit{Person} \rangle$ I,88 J,87 F,78 G,75 K,64

A=(Schwarzenegger, California),
80 occurrences

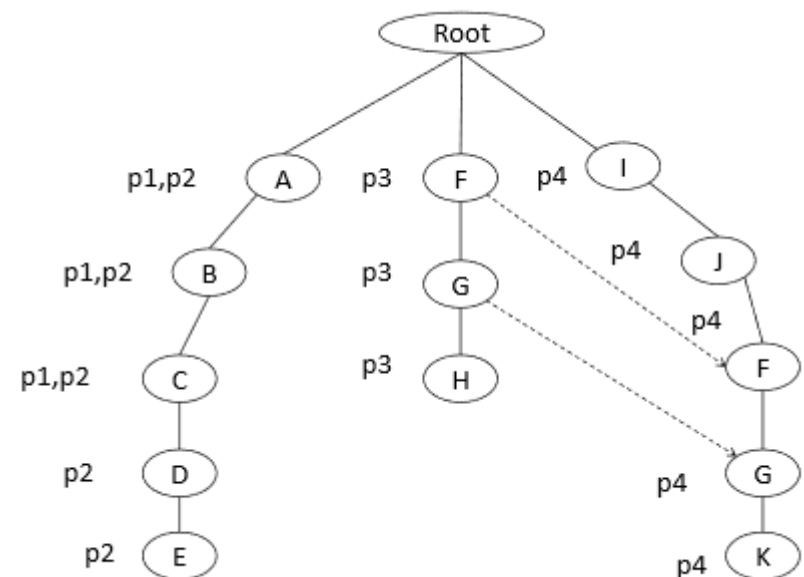
Cluster of relational phrases
$\langle \textit{location} \rangle$ is the heart of $\langle \textit{location} \rangle$
$\langle \textit{location} \rangle$ is situated in $\langle \textit{location} \rangle$
$\langle \textit{location} \rangle$ is enclosed by $\langle \textit{location} \rangle$
$\langle \textit{location} \rangle$ is located amidst $\langle \textit{location} \rangle$
$\langle \textit{location} \rangle$ is surrounded by $\langle \textit{location} \rangle$

$\langle \textit{organization} \rangle$ acquires $\langle \textit{organization} \rangle$
 \uparrow
 $\langle \textit{organization} \rangle$ purchased share $\langle \textit{organization} \rangle$
 \uparrow
 $\langle \textit{organization} \rangle$ bought half of $\langle \textit{company} \rangle$
 \uparrow
 $\langle \textit{company} \rangle$ bought half of $\langle \textit{company} \rangle$
 \uparrow
 $\langle \textit{company} \rangle$ later bought half of $\langle \textit{company} \rangle$

Efficient support set overlap comparison

- n patterns $\rightarrow n^2$ comparisons?

ID	Pattern Synset & Support Sets
P_1	$\langle \text{Politician} \rangle$ was governor of $\langle \text{State} \rangle$ A,80 B,75 C,70
P_2	$\langle \text{Politician} \rangle$ politician from $\langle \text{State} \rangle$ A,80 B,75 C,70 D,66 E,64
P_3	$\langle \text{Person} \rangle$ daughter of $\langle \text{Person} \rangle$ F,78 G,75 H,66
P_4	$\langle \text{Person} \rangle$ child of $\langle \text{Person} \rangle$ I,88 J,87 F,78 G,75 K,64



Prefix tree allows quick retrieval of subsumed patterns

Outline

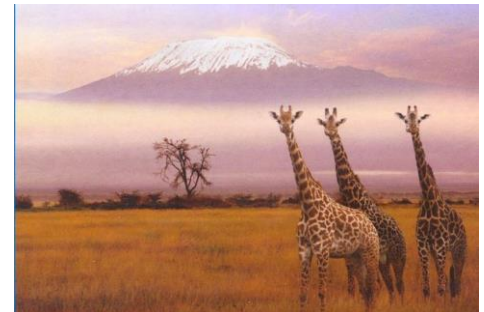
1. Problem
2. Manual patterns
3. Supervised learning
 1. Feature-based
 2. TACRED and BERT
4. Semi- and unsupervised extraction
 1. Iterative pattern learning (DIPRE)
 2. Distant supervision
 - CINEX
5. Evaluation
6. OpenIE
 1. PATTY
 2. Quasimodo
7. Negation

Quasimodo - Goal

- Mine Commonsense Knowledge (CSK) about :
 - Object properties
 - Human behavior
 - General concepts
- Focus on salient properties like
 - (bananas, are, edible)
 - (children, like, bananas)
- Avoid non salient properties like (from ConceptNet)
 - (elephant, CapableOf, visit the grocery store)
 - (dog, HasProperty, one among many animals)

Applications

- Chatbot
 - Me: Hi Pandora, what do you suggest for breakfast?
 - Her: What about bouillabaisse for a starter?
- (Visual) Question Answering
 - Q: What's taller, the giraffe or the mountain?
 - A: The giraffe
- Visual content understanding
- Queries Interpretation
 - Jordan weather next week



Challenges

- Seldom expressed in assertions
- Non-encyclopedic (no Wikipedia)
- Noise and high bias on online content
- No way to prescribe limited fixed set of relations

Banana

From Wikipedia, the free encyclopedia

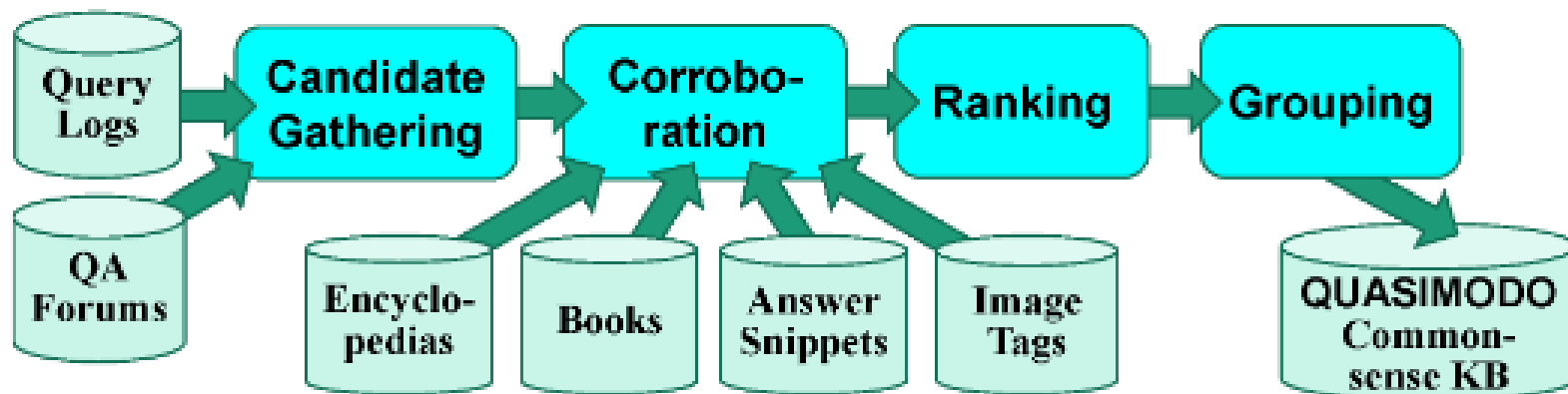
This article is about bananas generally. For the genus to which banana p starchier bananas used in cooking, see [Cooking banana](#). For other uses,

A **banana** is an edible [fruit](#) – botanically a [berry](#)^{[1][2]} – produced by several kinds of large [herbaceous flowering plants](#) in the [genus *Musa*](#).^[3] In some countries, [bananas used for cooking](#) may be called "plantains", distinguishing them from **dessert bananas**. The fruit is variable in size, color, and firmness, but is usually elongated and curved, with soft flesh rich in [starch](#) covered with a rind, which may be [green, yellow, red, purple,](#) or brown when ripe. The fruits grow in clusters hanging from the top of the plant. Almost all modern edible seedless ([parthenocarp](#)) bananas come from two wild species – *[Musa acuminata](#)* and *[Musa balbisiana](#)*. The [scientific names](#) of most cultivated bananas are *Musa acuminata*, *Musa*

Previous Work

- Traditional Knowledge Bases
 - No commonsense
- ConceptNet
 - ~20 meta-relations
("is capable of", "can be used for", ...)
 - Manual, does not scale
- Webchild
 - ~20 relations, inspired by ConceptNet
 - Focus on possible properties, not salient ones
- TupleKB
 - OpenIE predicates
 - Still limited domain, science knowledge only

Quasimodo Pipeline



Candidate Gathering

- Main idea : Extract facts from questions
 - Asking certain questions conveys knowledge

Why are bananas yellow?  Bananas are yellow!

- Harvest human curiosity, « wisdom of the crowds »

Candidate Gathering – Query Logs

- Indirect access to the query logs through autocomplete

why do cats

why do cats **purr**

why do cats **like boxes**

why do cats **meow**

why do cats **knead**

why do cats **sleep so much**

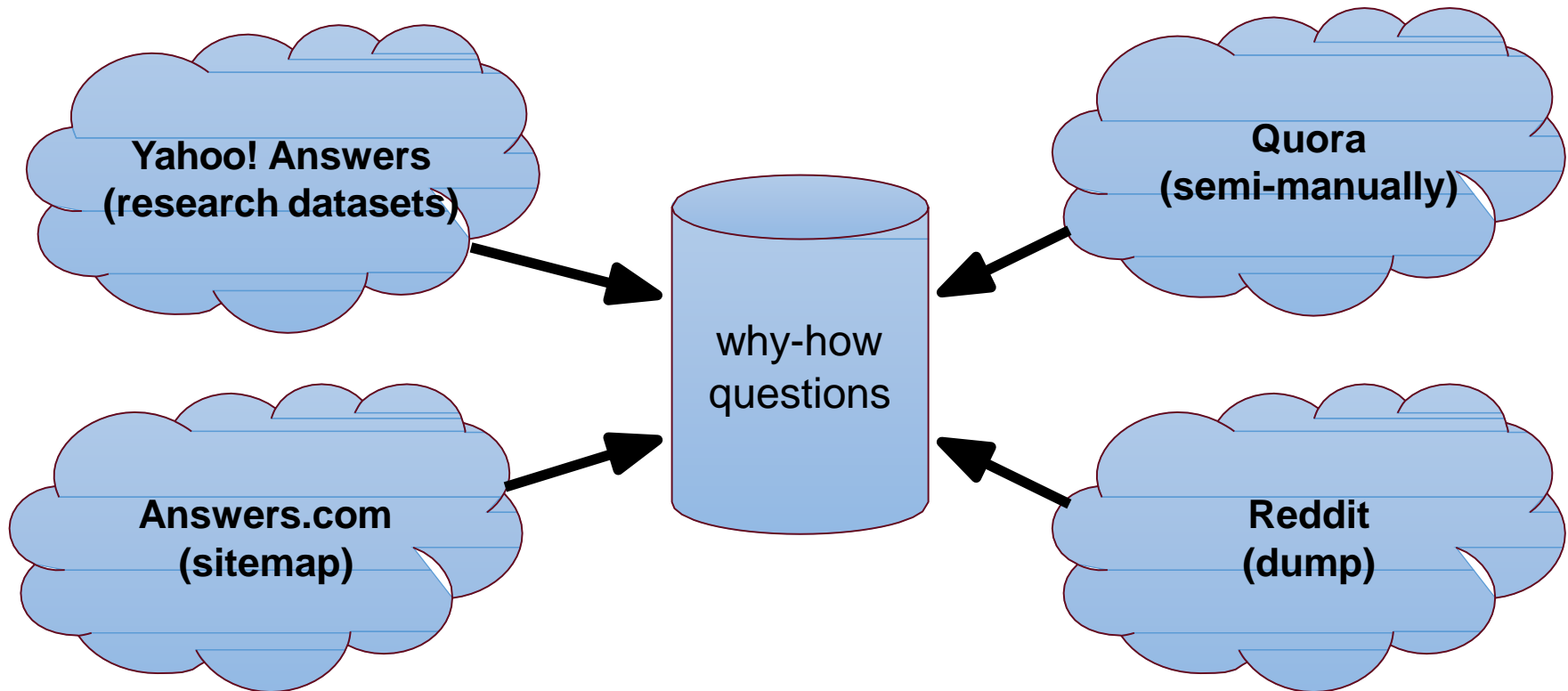
why do cats **hate water**

why do cats **like catnip**

why do cats **lick you**

why do cats **have whiskers**

Candidate Gathering – QA Forums

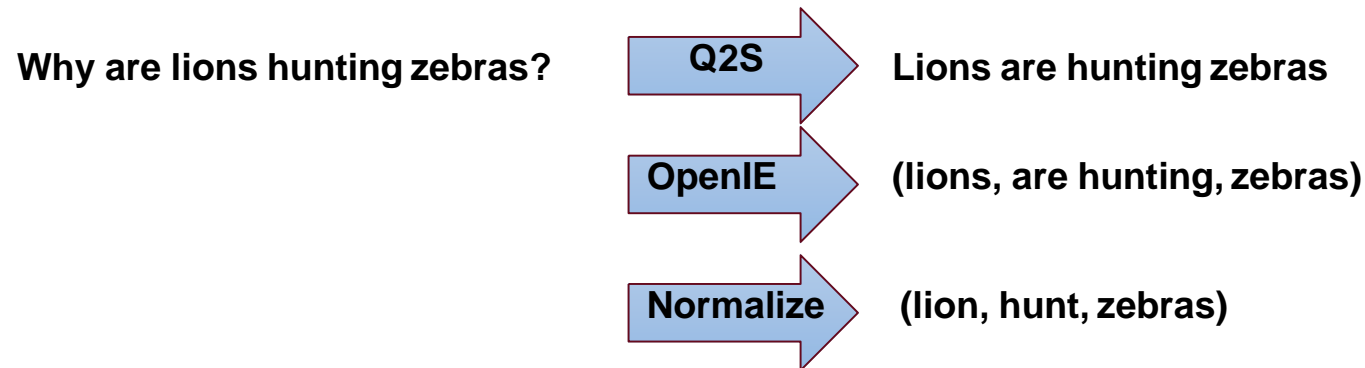


Candidate Gathering – Statistics

Pattern	In Query Logs	In QA Forums
how does	19.4%	7.5%
why is	15.8%	10.4%
how do	14.9%	38.07%
why do	10.6%	9.21%
how is	10.1 %	4.31%
why does	8.97%	5.46%
why are	8.68%	5.12%
how are	5.51%	1.8%
how can	3.53%	10.95%
why can't	1.77%	1.40%
why can	0.81%	0.36%

Candidate Gathering – Results

- Questions to statements to tuples using OpenIE



Corroboration

- Reduce noise with cooccurrence signals from :
 - Wikipedia and Simple Wikipedia
 - Answer snippets from search engines
 - Google Books
 - Image Tags from OpenImages and Flickr
 - Google's Conceptual Captions dataset



Wildlife Photographer of the Year award goes to Yongqing Bao for image of Tibetan fox attacking marmot

- Train classifier from all signals on 700 manually annotated triples

Grouping

- Reduce redundancy
- Co-clustering method based on tri-factorization
- Compute clusters for SO pairs and clusters for P phrases and align them with each other when meaningful
- Number of (soft) clusters for SO pairs and for P phrases can be different

P cluster	SO cluster
make noise at, be loud at, croak in	fox-night, frog-night, donkey-night
sleep in, be bored in, talk in	student-class, student-lectures

Statistics

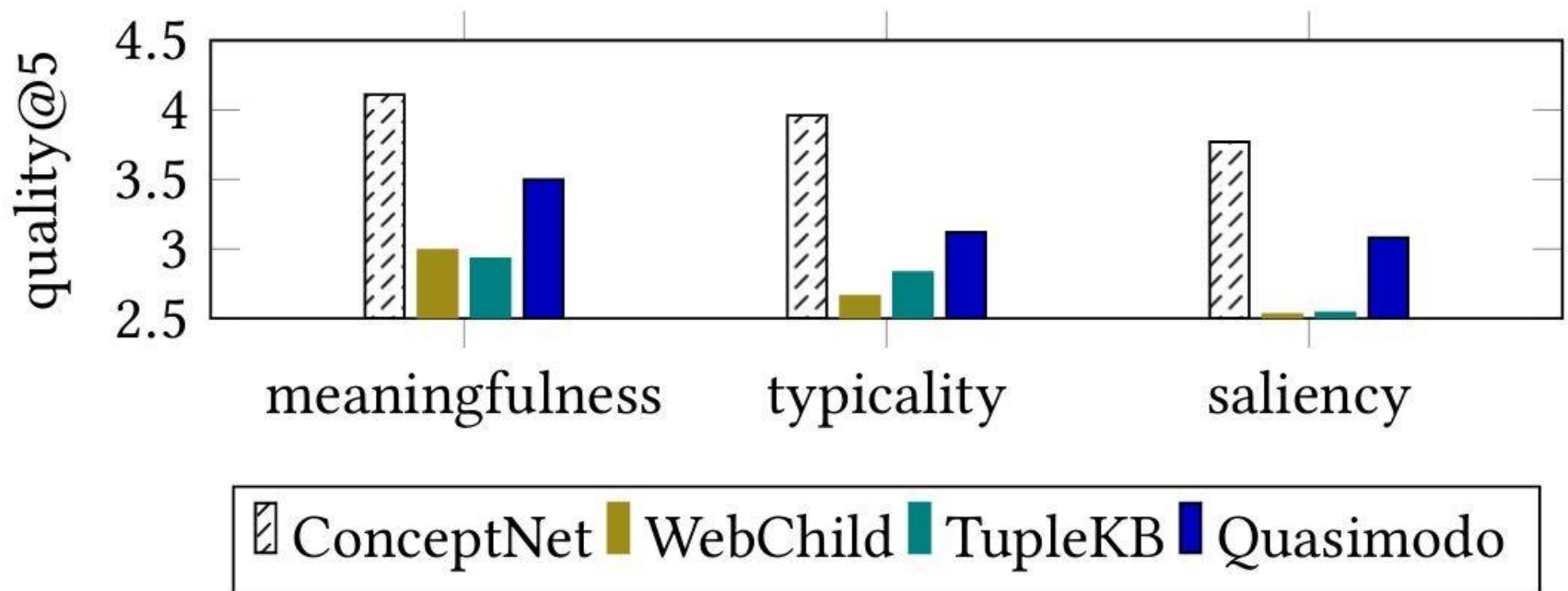
	Full KB								
	#S	#P	#P \geq 10	#SPO	#SPO/S	animals		occupations	
	#S	#SPO	#S	#SPO		#S	#SPO	#S	#SPO
ConceptNet-full@en	842,532	39	39	1,334,425	1.6	50	2,678	50	1,906
ConceptNet-CSK@en	41,331	19	19	214,606	5.2	50	1,841	50	1,495
TupleKB	28,078	1,605	1,009	282,594	10.1	49	16,052	38	5,321
WebChild	55,036	20	20	13,323,132	242.1	50	27,223	50	26,257
Quasimodo	80,145	78,636	6084	2,262,109	28.2	50	39,710	50	18,212

Anecdotal Examples

Practical knowledge from human	(car, slip on, ice)
Problems linked to a subject	(pen, can, leak)
Emotions linked to events	(divorce, can, hurt)
Human behaviors	(ghost, scare, people)
Visual assertions	(road, has_color, black)
Cultural knowledge (here U.S.)	(school, have, locker)
Comparative knowledge	(light, faster than, sound)

Precision

Sample from a list of common subjects (most popular animals and occupations)



Overview

We are evaluating the quality of computer-generated general knowledge. Your task is to evaluate the quality of the generated knowledge along three aspects: 1) Meaningfulness, 2) Correctness, 3) Importance.

Examples:

- **Lion, hunts, zebras:** Meaningful, correct, important
- **Lion, has, shinbone:** Meaningful, correct, not so important
- **Lion, is, vegetarian:** Meaningful, **incorrect**, not important
- **Equity, causes, solution:** **Not meaningful**, incorrect, not important

Fact: muscle - is used for - flexing arm

Is this statement meaningful? (required)

	1	2	3	4	5	
Meaningful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Gibberish

Is this true for most muscle(s)? (required)

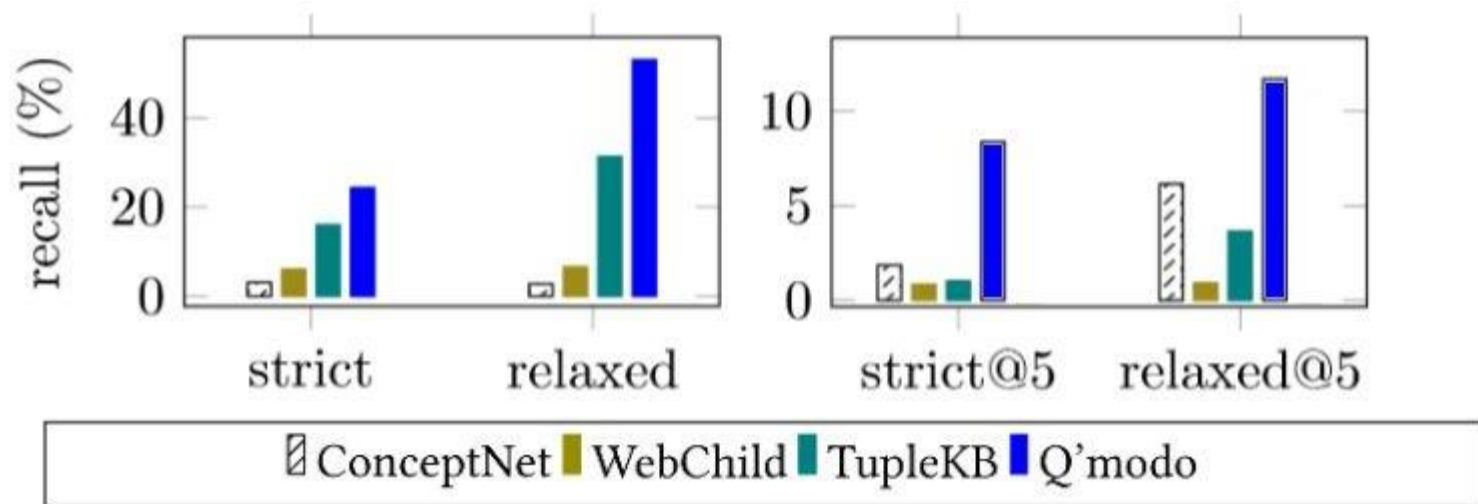
	1	2	3	4	5	
True	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	False

Is this an important fact? (required)

	1	2	3	4	5	
Important	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Boring/obscure

Recall

- Given a subject, ask crowd workers to give a statement starting with "<Subjects> ...", like "Elephants ... are grey"
- Strict = exact match, Relaxed = partial match



Multiple Choice Question Answering

Where would I not want a fox?

👍 hen house, 👎 england, 👎 mountains,
👎 english hunt, 👎 california

KB	All
#Questions (Train/Test)	10974/3659
Random	22.0
word2vec	27.2
Quasimodo	31.3
ConceptNet	27.5
TupleKB	27.5
WebChild	24.1

Outline

1. Problem
2. Manual patterns
3. Supervised learning
 1. Feature-based
 2. TACRED and BERT
4. Semi- and unsupervised extraction
 1. Iterative pattern learning (DIPRE)
 2. Distant supervision
 - CINEX
5. Evaluation
6. OpenIE
 1. PATTY
 2. Quasimodo
7. Negation

References

- Papers:
 - Stanovsky and Dagan, Creating a Large Benchmark for Open Information Extraction, EMNLP 2016
 - Nakashole et al., PATTY: A Taxonomy of Relational Patterns with Semantic Types, EMNLP 2012
 - Romero et al., Salient Commonsense Properties from Query Logs and Question Answering Forums, CIKM 2019
- Slides
 - Adopted from Fabian Suchanek, Julien Romero and Oren Etzioni
- Code/APIs
 - OpenIE
 - <https://www.textrazor.com/demo>
 - <https://gate.d5.mpi-inf.mpg.de/ClausIEGate/ClausIEGate/>
 - <https://github.com/dair-iitd/OpenIE-standalone>
- Link collection on OpenIE
 - <https://github.com/gkiril/oie-resources>

Assignment 7

- Code your own open information extraction
- Evaluation on benchmark data from [Stanovsky and Dagan, EMNLP 2017]
- F1 on extractions (head word match for predicate)

Take home

- Fixed relations

- Supervised learning data bottleneck, but performant
- Iterative pattern learning and distant supervision as alternatives
- BERT allows to bypass feature engineering

- Evaluation

- Right metric for right problem
- Evaluation of novel discoveries nontrivial
- Error analysis

- Much effort in data preparation, labelling

- Open information extraction

- Alternative requiring no decision on schema upfront