Open Information Extraction

Prof. Dr. Alexander Fraser
Slides by Dr. Matthias Huck

Center for Information and Language Processing
LMU Munich

Winter Semester 2020/2021

Schedule

Today: last lecture
Next week: **review**
10.02: no class
17.02: **exam**, time & location: Feb 17th, 16:15, online
(**you need a working webcam!**)

You have to register
for the exam in LSF!

Wednesday Seminar

The seminar on February 3rd will have two topics because of the Software Technik exam on 10.02

Both topics will have the usual length, so we will certainly go well over on time

Reading

You should (have) read Sarawagi Chapter 4 on relation extraction (see the lecture website)

Bachelorarbeit Topics

The list of offered bachelors (and masters) topics will be available on Monday 15.02 on the CIS web page.

Introduction

Reminder: We know how to:

- Recognize named entities in natural language text
- Extract binary relations between named entities

We have seen an application scenario:

- Relations can be stored in a knowledge base
- And be used in question answering or spoken dialogue systems

But so far, there are limitations, most notably:

- We have dealt with narrow domains (such as geographical location, food, plant seed development)
- The sets of entity types and relations were closed and manually defined

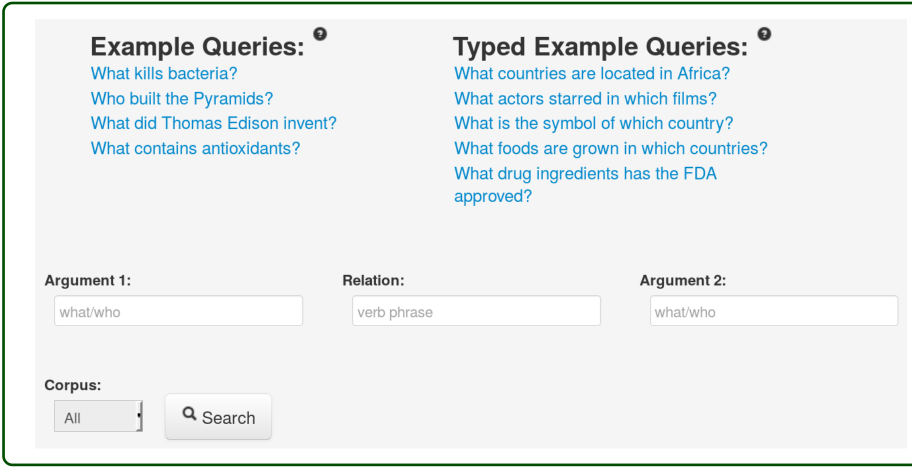
Open IE aims at:

- Not being limited to any single domain
- Not being limited to pre-defined entity types and relations

Outline

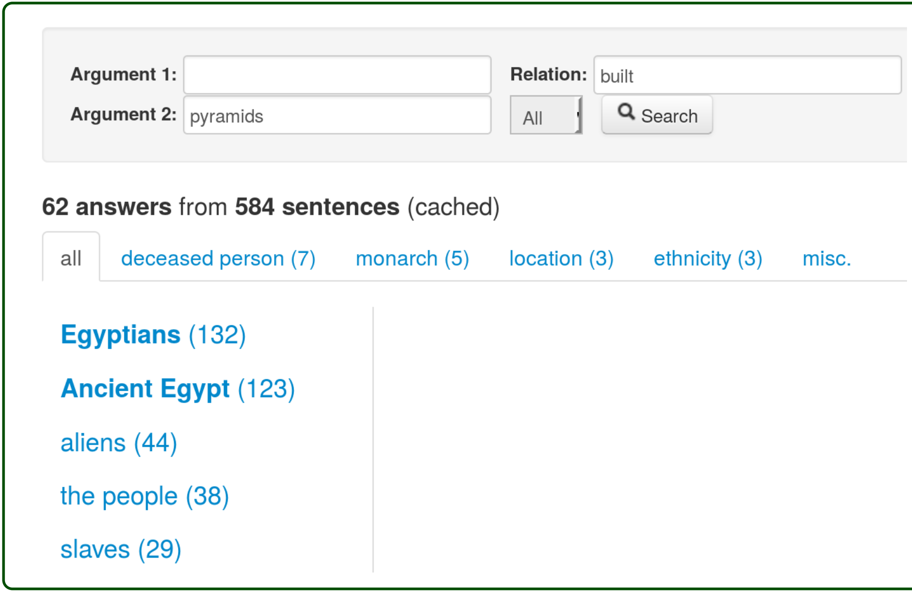
OPEN IE:
MOTIVATION & TASK DEFINITION

Open IE: Motivation (1)



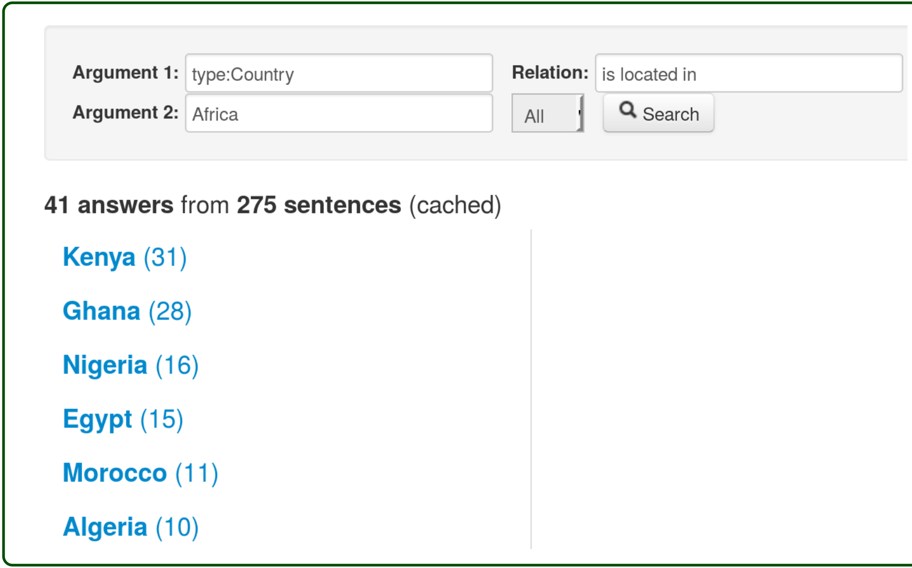
[openie.allenai.org query, 16 Jan. 2017]

Open IE: Motivation (2)



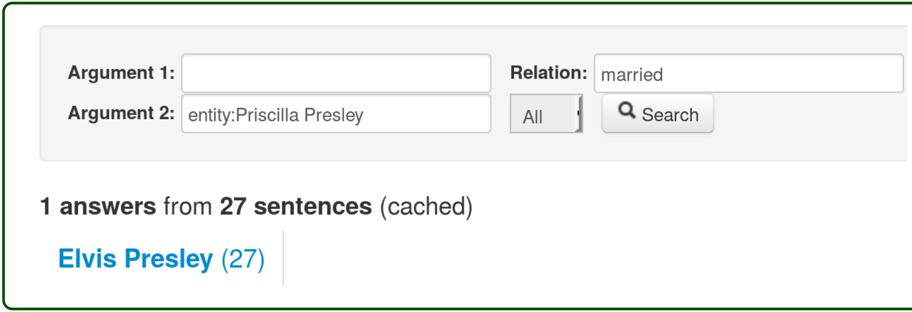
[openie.allenai.org query, 16 Jan. 2017]

Open IE: Motivation (3)



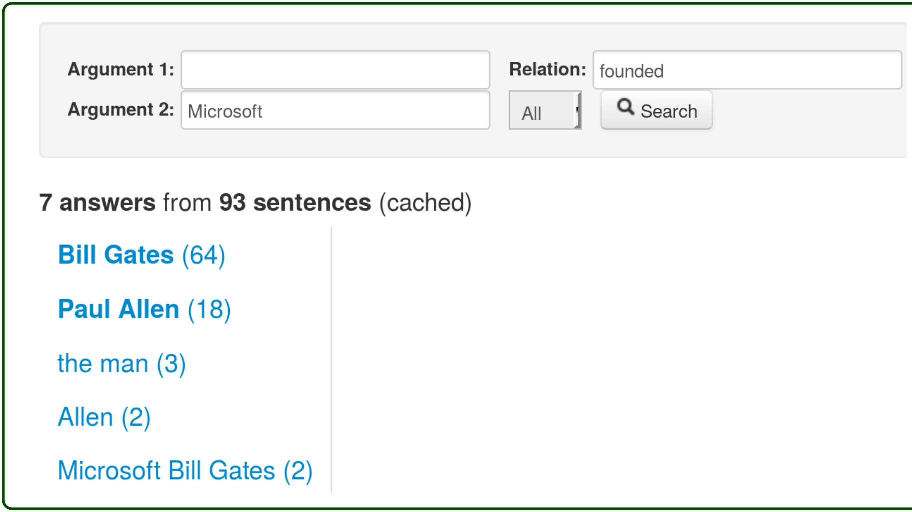
[openie.allenai.org query, 16 Jan. 2017]

Open IE: Motivation (4)



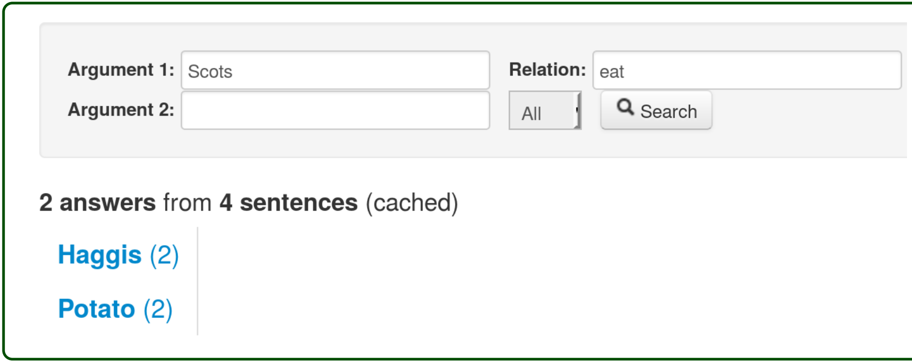
[openie.allenai.org query, 16 Jan. 2017]

Open IE: Motivation (5)



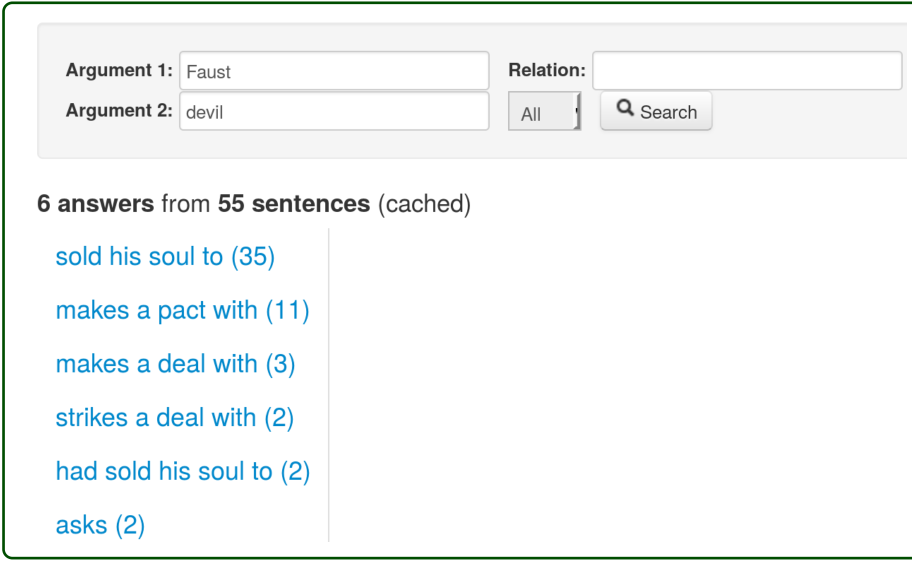
[openie.allenai.org query, 16 Jan. 2017]

Open IE: Motivation (6)



[openie.allenai.org query, 16 Jan. 2017]

Open IE: Motivation (7)



[openie.allenai.org query, 16 Jan. 2017]

Open IE: Motivation (8)

Argument 1: students

Relation: sleep in

Argument 2:

All

Search

6 answers from **35 sentences** (cached)

- **Tent** (20)
- **Dormitory** (7)
- classrooms (2)
- car (2)
- thatched-roof bandas (2)
- class (2)

[openie.allenai.org query, 16 Jan. 2017]

Open IE: Task Definition

We want to find any relation that is expressed in large data.

- Cannot resort to specialized domain knowledge
- Cannot think of all possible relation types beforehand
- Should rather not force all possible arguments into a rigid set of entity types

We want to scale to billions of documents that are heterogeneous wrt. domains, quality, credibility.

- Which relations are correct?
- Which are uninformative or incoherent?
- Which are redundant?

OPEN RELATION EXTRACTION: TEXTRUNNER & REVERB

Open Relation Extraction: Example

`Hudson was born in Hampstead, which is a suburb of London.`

(Hudson, was born in, Hampstead)
(Hampstead, is a suburb of, London)

[Fader *et al*.. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

Open Relation Extraction: Basic Approach

Learn a general model of how (arbitrary) relations are expressed in a particular language.

- Neither relation names nor argument types known in advance
- Bootstrap with heuristics or distant supervision
- Train a (sequence) classifier (often with unlexicalized features)

Rel. Freq.	Category	Simplified Lexico-Syntactic Pattern	Example
37.8	Verb	E_1 Verb E_2	X established Y
22.8	Noun+Prep	E_1 NP Prep E_2	X settlement with Y
16.0	Verb+Prep	E_1 Verb Prep E_2	X moved to Y
9.4	Infinitive	E_1 to Verb E_2	X plans to acquire Y
5.2	Modifier	E_1 Verb E2 Noun	X is Y winner
1.8	Coordinate_n	E_1 (and|,|-|:) E_2 NP	X-Y deal
1.0	Coordinate_v	E_1 (and|,) E_2 Verb	X , Y merge
0.8	Appositive	E_1 NP (:|,)? E_2	X hometown : Y

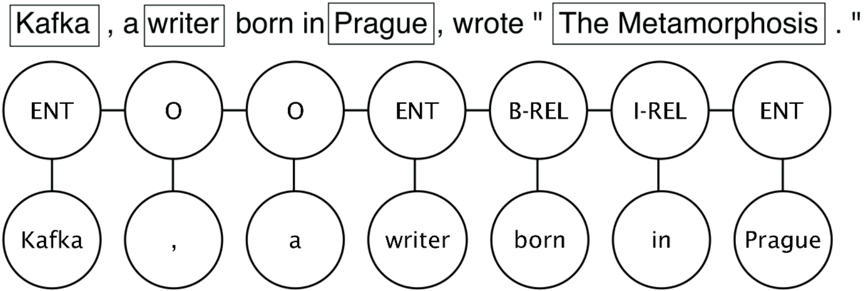
[Etzioni *et al.*. Open Information Extraction from the Web. Communications of the ACM, vol. 51 no. 12, Dec. 2008.]
[Banko and Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. Proc. of the ACL, Columbus, OH, USA, June 2008.]

Open Relation Extraction: "Three-Step Method"

1. Label: Sentences are automatically labeled with extractions using heuristics or distant supervision.
2. Learn: A relation phrase extractor is learned, e.g. using a sequence-labeling graphical model (CRF).
3. Extract: The system takes a sentence as input, identifies a candidate pair of NP arguments (arg1, arg2) from the sentence, and then uses the learned extractor to label each word between the two arguments as part of the relation phrase or not.

[Fader *et al.*. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]
Describes ReVerb. See also: TextRunner (Banko et al. 2007) and WOE^{pos}/WOE^{parse} (Wu and Weld, 2010)

Open Relation Extraction as Sequence Labeling



[Banko and Etzioni. The Tradeoffs Between Open and Traditional Relation Extraction. Proc. of the ACL, Columbus, OH, USA, June 2008.]

Uninformative Relations

Problem 1: The sequence classifier may come up with an uninformative relation name.

`Faust made a deal with the devil.`

(Faust, made, deal)
(Faust, made deal with, devil)

Uninformative	Completion
is	is an album by, is the author of, is a city in
has	has a population of, has a Ph.D. 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*.. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

Incoherent Relations

Problem 2: The sequence classifier may come up with an incoherent relation name.

Sentence	Incoherent Relation
The guide contains dead links and omits sites.	contains omits
The Mark 14 was central to the torpedo scandal of the fleet.	was central torpedo
They recalled that Nungesser began his career as a precinct leader.	recalled began

[Fader *et al*.. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

POS-based Constraints to Avoid Incoherence & Uninformativeness

```
Extendicare agreed to buy Arbor Health Care
for about US $432 million in cash and assumed
debt.
```

(Arbor Health Care, for assumed, debt)

- POS-based regular expressions help avoid extraction of uninformative or incoherent relation phrases
- Manually written; e.g. the relation phrase must match:

V | V P | V W* P
V = verb particle? adv? W = (noun | adj | adv | pron | det) P = (prep | particle | inf. marker)

- Choose longest possible match
- Require the relation phrase to appear between its arguments

[Fader *et al.*. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

Overspecific Relations & How to Avoid Them

Problem 3: Some relations are specific to an argument pair, or have only a few possible instances.

`The Obama administration is offering only modest greenhouse gas reduction targets at the conference.`

(Obama administration, is offering only modest greenhouse gas reduction targets at, conference)

- Intuition: a valid relation phrase should take many distinct arguments in a large corpus
- Lexical constraint: relation phrases are required to match at least k distinct argument pairs in the data (e.g., $k = 20$)

[Fader *et al.*. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

Relation Phrase Normalization

`Shakespeare (` *has written* `|` *wrote* `|` *was writing* `)`
`Hamlet.`

Allow for minor variations in relation phrases.

- Remove inflection
- Remove auxiliary verbs, adjectives, adverbs

[Fader *et al.*. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

Confidence Function

- Train classifier to assign a confidence score to each extraction
- Trade recall for precision by tuning a confidence threshold

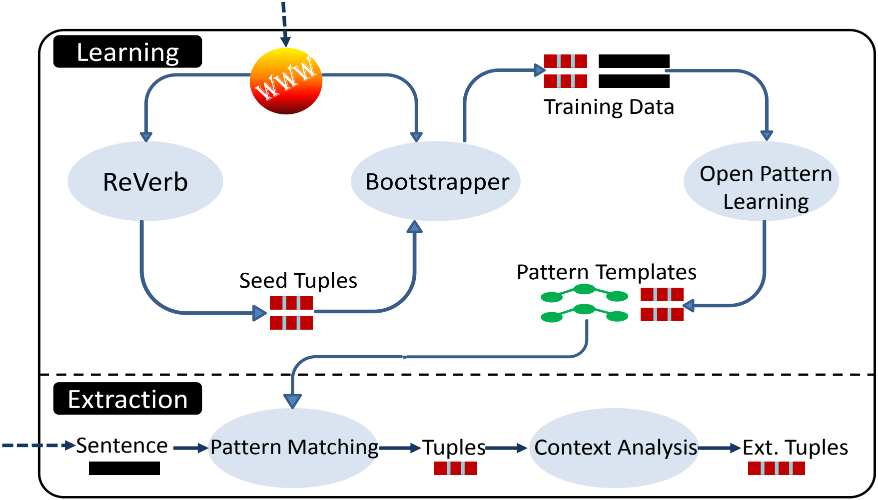
Weight	Feature
1.16	(x, r, y) covers all words in s
0.50	The last preposition in r is *for*
0.49	The last preposition in r is *on*
0.46	The last preposition in r is *of*
0.43	len(s) $\leq$ 10 words
0.43	There is a *WH*-word to the left of r
0.42	r matches VW* P
0.39	The last preposition in r is *to*
0.25	The last preposition in r is *in*
0.23	10 words $<$ len(s) $\leq$ 20 words
0.21	s begins with x
0.16	y is a proper noun
0.01	x is a proper noun
-0.30	There is an NP to the left of x in s
-0.43	20 words $<$ len(s)
-0.61	r matches V
-0.65	There is a preposition to the left of x in s
-0.81	There is an NP to the right of y in s
-0.93	Coord. conjunction to the left of r in s

[Fader *et al.*. Identifying Relations for Open Information Extraction. Proc. of EMNLP, Edinburgh, Scotland, UK, July 2011.]

Open Relation Extraction: Ollie

OLLIE (Open Language Learning for Information Extraction)

Bootstrapping with high precision seed tuples from existing system (REVERB, cf. previous part)

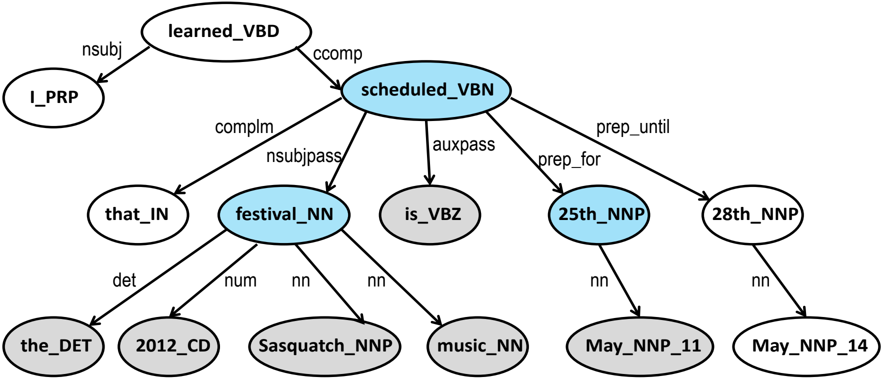


[Mausam *et al.*. Open Language Learning for Information Extraction. Proc. of EMNLP, Jeju Island, Korea, July 2012.]

Employing Dependency Parses

`I learned that the 2012 Sasquatch music festival is scheduled for May 25th until May 28th.`

(the 2012 Sasquatch Music Festival, is scheduled for, May 25th)



[Mausam *et al.*. Open Language Learning for Information Extraction. Proc. of EMNLP, Jeju Island, Korea, July 2012.]

Open Pattern Templates

Open pattern templates encode the ways in which a relation may be expressed in a sentence.

- Based on a dependency parse path
- with lexical constraint
- and POS constraint

Extraction Template	**Open Pattern**
1. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓{prep_∗}↓ {arg2}
2. (arg1; {rel}; arg2)	{arg1} ↑nsubj↑ {rel:postag=VBD} ↓dobj↓ {arg2}
3. (arg1; be {rel} by; arg2)	{arg1} ↑nsubjpass↑ {rel:postag=VBN} ↓agent↓ {arg2}
4. (arg1; be {rel} of; arg2)	{rel:postag=NN;**type=Person**} ↑nn↑ {arg1} ↓nn↓ {arg2}
5. (arg1; be {rel} {prep}; arg2)	{arg1} ↑nsubjpass↑ {slot:postag=VBN;**lex** ∈**announce**|**name**|**choose...**} ↓dobj↓ {rel:postag=NN} ↓{prep_∗}↓ {arg2}

[Mausam *et al.*. Open Language Learning for Information Extraction. Proc. of EMNLP, Jeju Island, Korea, July 2012.]

OLLIE: Advantages

Previously (in REVERB), we required the relation phrase to appear between its arguments:

`Elvis married Priscilla.`

Open pattern templates may help with:

`Elvis and Priscilla are married.`

Other systems are designed to have verb-mediated relation phrases:

`Bill Gates founded Microsoft.`

OLLIE can deal with noun-mediated relations:

`Bill Gates is founder of Microsoft.`

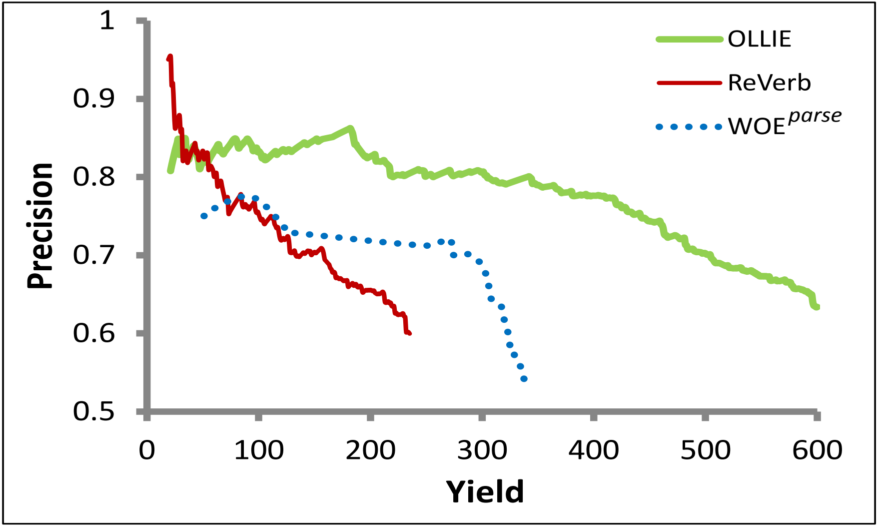
Many relationships are most naturally expressed via noun phrases:

`is capital of`, `is president of`, `is professor at`, ...

Dependency parse is useful; parsers not deemed too slow any more.

[Mausam *et al.*. Open Language Learning for Information Extraction. Proc. of EMNLP, Jeju Island, Korea, July 2012.]

OLLIE: Evaluation



[Mausam *et al.*. Open Language Learning for Information Extraction. Proc. of EMNLP, Jeju Island, Korea, July 2012.]

Open Relation Extraction: Stanford OpenIE

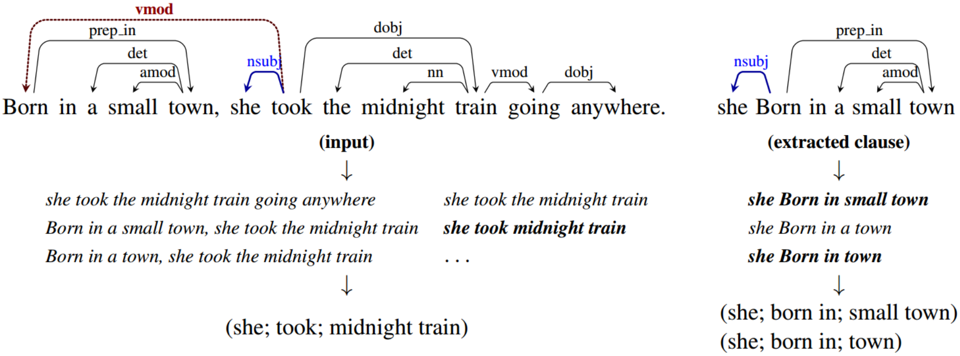
Stanford OpenIE

Heavily based on dependency parses.

1. Each dependency-parsed sentence is first split into a set of entailed clauses
2. Clauses are then maximally shortened, producing a set of entailed shorter sentence fragments
3. The fragments are segmented into relation triples, and output by the system

[Angeli *et al.*. Leveraging Linguistic Structure For Open Domain Information Extraction. Proc. of the ACL, Beijing, China, July 2015.]

STANFORD OPENIE: Illustration



[`http://nlp.stanford.edu/software/openie.html`, 17 Jan. 2017]
[Angeli *et al.*. Leveraging Linguistic Structure For Open Domain Information Extraction. Proc. of the ACL, Beijing, China, July 2015.]

Clause Splitting as a Classification Task

- Inspect the dependency structure
- Decide whether to split on a dependency arc
- Classifier using a set of dependency-based features
- Distant supervision for training:
  sequence which recovers a known relation is correct

Feature Class	**Feature Templates**
Edge taken	$\{l, \text{short\_name}(l)\}$
Last edge taken	$\{\text{incoming\_edge}(p)\}$
Neighbors of parent	$\{\text{nbr}(p), (p, \text{nbr}(p))\}$
Grandchild edges	$\{\text{out\_edge}(c),$ $(e, \text{out\_edge}(c))\}$
Grandchild count	$\{\text{count}\,(\text{nbr}(e_{\text{child}}))$ $(e, \text{count}\,(\text{nbr}(e_{\text{child}})))\}$
Has subject/object	$\forall_{e \in \{e, e_{\text{child}}\}} \forall_{l \in \{\mathit{subj}, \mathit{obj}\}}$ $\mathbb{1}(l \in \text{nbr}(e))$
POS tag signature	$\{\text{pos}(p), \text{pos}(c),$ $(\text{pos}(p), \text{pos}(c))\}$
Features at root	$\{\mathbb{1}(p = \mathit{root}), \text{POS}(p)\}$

[Angeli *et al.*. Leveraging Linguistic Structure For Open Domain Information Extraction. Proc. of the ACL, Beijing, China, July 2015.]

Atomic Patterns over Short Entailed Sentences

Verb-mediated:

Input	Extraction
cats play with yarn	(cats; play with; yarn)
fish like to swim	(fish; like to; swim)
cats have tails	(cats; have; tails)
cats are cute	(cats; are; cute)
Tom and Jerry are fighting	(Tom; fighting; Jerry)
There are cats with tails	(cats; have; tails)

Noun-mediated:

Input	Extraction
Durin, son of Thorin	(Durin; is son of; Thorin)
Thorin's son, Durin	(Thorin; 's son; Durin)
IBM CEO Rometty	(Rometty; is CEO of; IBM)
President Obama	(Obama; is; President)
Fischer of Austria	(Fischer; is of; Austria)
IBM's research group	(IBM; 's; research group)
US president Obama	(Obama; president of; US)
Our president, Obama,	(Our president; be; Obama)

[Angeli *et al.*. Leveraging Linguistic Structure For Open Domain Information Extraction. Proc. of the ACL, Beijing, China, July 2015.]

Validating Deletions with Natural Logic

Scopes of operators *all*, *no*, *many*, ...

- `all rabbits eat fresh vegetables`
  yields (rabbits, eat, vegetables)
- `all young rabbits drink milk`
  does not yield (rabbits, drink, milk)

Non-subsective adjectives

- a *fake gun* is not a gun

Prepositional attachment

- `Alice played baseball on Sunday`
  entails `Alice played on Sunday`
- `Obama signed the bill on Sunday`
  should not entail `Obama signed on Sunday`

[Angeli *et al.*. Leveraging Linguistic Structure For Open Domain Information Extraction. Proc. of the ACL, Beijing, China, July 2015.]

STANFORD OPENIE: Example Extractions

Born in Honolulu, Hawaii, Obama is a US Citizen.

Our System	**Ollie**
(Obama; is; US citizen)	(Obama; is; a US citizen)
(Obama; born in; Honolulu, Hawaii)	(Obama; be born in; Honolulu)
	(Honolulu; be born in; Hawaii)
	(Obama; is citizen of; US)

Friends give true praise.

Enemies give fake praise.

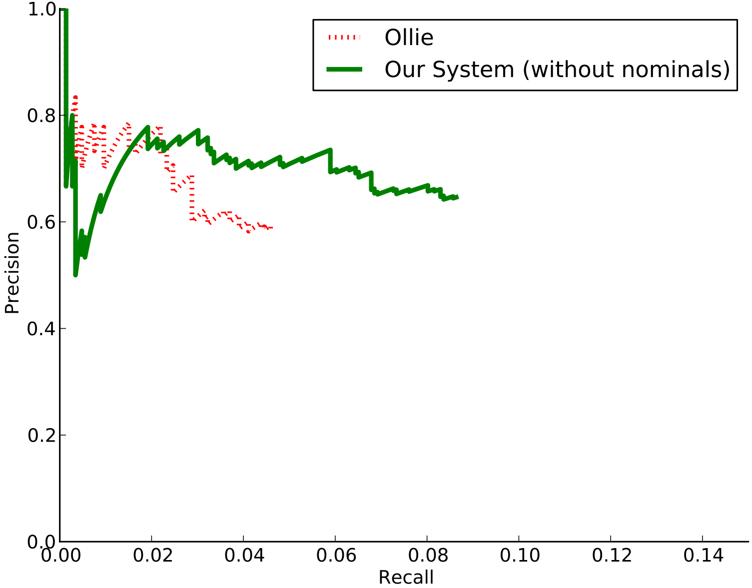
Our System	**Ollie**
(friends; give; true praise)	(friends; give; true praise)
(friends; give; praise)	
(enemies; give; fake praise)	(enemies; give; fake praise)

Heinz Fischer of Austria visits the US

Our System	**Ollie**
(Heinz Fischer; visits; US)	(Heinz Fischer of Austria; visits; the US)

[Angeli *et al*.. Leveraging Linguistic Structure For Open Domain Information Extraction. Proc. of the ACL, Beijing, China, July 2015.]

STANFORD OPENIE: Evaluation



[Angeli *et al.*. Leveraging Linguistic Structure For Open Domain Information Extraction. Proc. of the ACL, Beijing, China, July 2015.]

Discussion:
Further Challenges

Synonym Resolution

The same entity may be referred to by a variety of names.

- *Michael Jackson*; *Jacko*; *The King of Pop*; ...

The same fact may be expressed in a variety of ways.

- *IBM built Watson*
  IBM created Watson
  IBM invented Watson
  ...
- *Dookie is a record by Green Day*
  Dookie is an album by Green Day
  ...

RESOLVER identifies synonymous relations and objects

[Yates and Etzioni. Unsupervised Resolution of Objects and Relations on the Web. Proc. of the NAACL, Rochester, NY, USA, April 2007.]

[Etzioni *et al.*. Open Information Extraction from the Web. Communications of the ACM, vol. 51 no. 12, Dec. 2008.]

Disambiguation

The same string may refer to different entities (especially across different domains).

- *Watson*, the founder of IBM; *Watson*, the computer system
- *mouse*, the animal; *mouse*, the input device
- *1984*, the year; *1984*, the book
- *Paris*, France; *Paris*, Texas

[Etzioni *et al.*. Open Information Extraction from the Web. Communications of the ACM, vol. 51 no. 12, Dec. 2008.]

Vagaries of Natural Language

- pronoun resolution
- metaphor
- anaphora
- complex or ungrammatical sentences
- irony, sarcasm
- …

[Etzioni *et al.*. Open Information Extraction from the Web. Communications of the ACM, vol. 51 no. 12, Dec. 2008.]

Incorrect Information

Nowadays referred to as “fake news”.

- *Elvis killed JFK*

Rate the reliability of an extracted relation.

- The relation extractor may have made an error: cf. the previously discussed confidence function
- Occurrence frequencies over the whole corpus can give an indication
- Credibility of the source of a document

YAGO-NAGA ranks facts f via:

$$\text{confidence}(f) = \max\left\{\text{accuracy}(f, s) \times \text{trust}(s) \;\middle|\; s \in \text{witnesses}(f)\right\}$$

[Kasneci *et al.*. The YAGO-NAGA Approach to Knowledge Discovery. ACM SIGMOD Record Volume 37 Issue 4, Dec. 2008. `https://suchanek.name/work/publications/sigmodrec2008.pdf`]

[Etzioni *et al.*. Open Information Extraction from the Web. Communications of the ACM, vol. 51 no. 12, Dec. 2008.]

Temporal and Spatial Aspects

Time.

- The capital city of the Federal Republic of Germany?
  Bonn in 1981. Berlin in 2016.
- Plato has not met with Tsipras

Space.

- An elephant does not fit into a coffee mug
- Trees don't travel
- Somebody who pays in GBP is probably located in Britain
- Plato has never seen a kangaroo

Fact Consistency Checks

Avoid contradictory facts within the knowledge base.

- `Elvis died in 460 AD` cannot refer to *Elvis Presley* if we already knows that Elvis Presley was born in 1935
- born(X,Y) $\wedge$ died(X,Z) $\Rightarrow$ Y $<$ Z
- appears(A,P,B) $\wedge$ R(A,B) $\Rightarrow$ expresses(P,R)
  appears(A,P,B) $\wedge$ expresses(P,R) $\Rightarrow$ R(A,B)
- means("Elvis",Elvis_Presley, 0.8)
  means("Elvis",Elvis_Costello, 0.2)

Implemented in the SOFIE IE system, which aims to extend the YAGO knowledge base

[Suchanek. Information Extraction for Ontology Learning. Book chapter in Völker and Lehman: Perspectives on Ontology Learning, 2014. `https://suchanek.name/work/publications/ontologybookchapter.pdf`]

Conclusion

Summary: Open IE

- Discovering relations without a closed set of pre-defined relation types
- Open-domain
- Learning from the whole Web
- Distant supervision / bootstrapping to get started
- Attention to detail required to avoid pitfalls
- The system should benefit from the sheer size of the data
- It should learn more by itself when being run perpetually, and become more reliable

The End!

Thank you for your attention

Thanks to Matthias Huck for the slides