Open Information Extraction: the Second Generation

Authors:
Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam
Published In:
International Joint Conference on Artificial Intelligence, 2011

How to Scale IE?

1970s-1980s: heuristic, hand-crafted clues

- Facts from earnings announcements
- Narrow domains; brittle clues

1990s: IE as supervised learning

"Mary was named to the post of CFO, succeeding Joe who retired abruptly."

Does “IE as supervised learning” scale to reading the Web?

No.

Critique of IE=supervised learning

- Relation specific
- Genre specific
- Hand-craft training examples

Does not scale to the Web!

Semi-Supervised Learning

- Few hand-labeled examples per relation!
- → Limit on the number of relations
- → relations are pre-specified

- → Still does not scale to the Web

Machine Reading at Web Scale

- **A "universal schema" is impossible**
- **Global consistency is like world peace**

- **Ontological "glass ceiling"**
  - Limited vocabulary
  - Pre-determined predicates
  - Swamped by reading at scale!

Motivation

- General purpose
  - hundreds of thousands of relations
  - thousands of domains
- Scalable: computationally efficient
  - huge body of text on Web and elsewhere
- Scalable: minimal manual effort
  - large-scale human input impractical
- Knowledge needs not anticipated in advance
  - rapidly retargetable

Open IE Guiding Principles

- Domain independence
  - Training for each domain/fact type not feasible
- Scalability
  - Ability to process large number of documents fast
- Coherence
  - Readability important for human interactions

Open vs. Traditional IE

	Traditional IE	*Open IE*
Input:	***Corpus + Hand-labeled Data***	***Corpus + Existing resources***
Relations:	***Specified in Advance***	***Discovered Automatically***
Complexity:	***O(D * R)*** ***R relations***	***O(D)*** ***D documents***
Output:	***relation-specific***	***Relation-independent***

TextRunner

First Web-scale, Open IE system (Banko, IJCAI '07)

1,000,000,000 distinct extractions

Peak of 0.9 precision (but low recall)

Demo

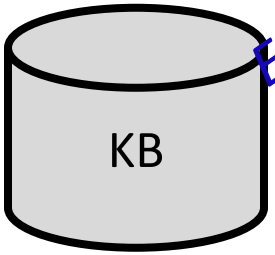
- http://openie.cs.washington.edu

Outline

Inference

Extraction

Fact

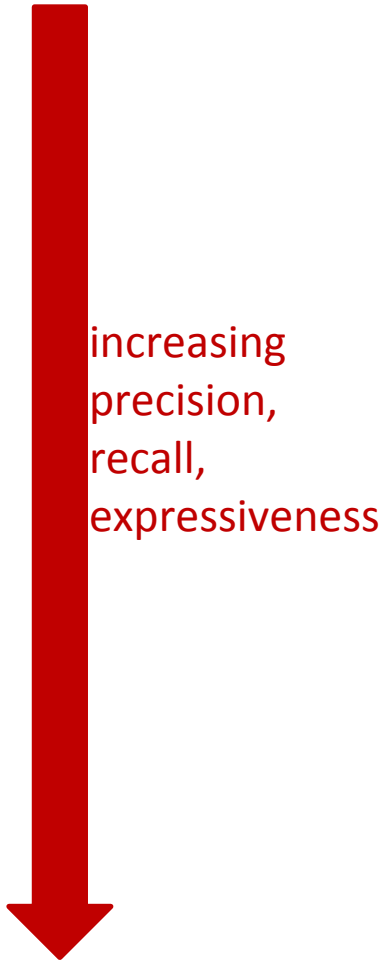


End-user applications

Downstream NLP/AI Tasks

Open Information Extraction

- 2007: Textrunner (~Open IE 1.0)
  - CRF and self-training
- 2010: ReVerb (~Open IE 2.0)
  - POS-based relation pattern
- 2012: OLLIE (~Open IE 3.0)
  - Dep-parse based extraction; nouns; attribution
- 2014: Open IE 4.0
  - SRL-based extraction; temporal, spatial…
- 2016 [@IITD]: Open IE 5.0
  - compound noun phrases, numbers, lists



Fundamental Hypothesis

- ∃ *semantically tractable* subset of English
  - Characterized relations & arguments via POS
  - Characterization is compact, domain independent
  - Covers 85% of binary relations in sample

ReVerb

Identify **Re**lations 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 of ReVerb Relations

invented	*acquired by*	*has a PhD in*
inhibits tumor growth in	*voted in favor of*	*won an Oscar for*
has a maximum speed of	*died from complications of*	*mastered the art of*
gained fame as	*granted political asylum to*	*is the patron saint of*
was the first person to	*identified the cause of*	*wrote the book on*

Lexical Constraint

Problem: "overspecified" relation phrases

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

Solution: must have many distinct args in a large corpus

is offering only modest …

Obama the conference } ≈ 1

is the patron saint of

100s ≈ {
Anne mothers
George England
Hubbins quality footwear
….
}

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 (phrases)	***100,000+***

ReVerb Extraction Algorithm

1. Identify longest relation phrases satisfying constraints

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

arg1 arg2 arg1 arg2

2. Heuristically identify arguments for relation phrase

(**Hudson**, **was born in**, **Hampstead**)

(**Hampstead**, **is a suburb of**, **London**)

ReVerb Strength

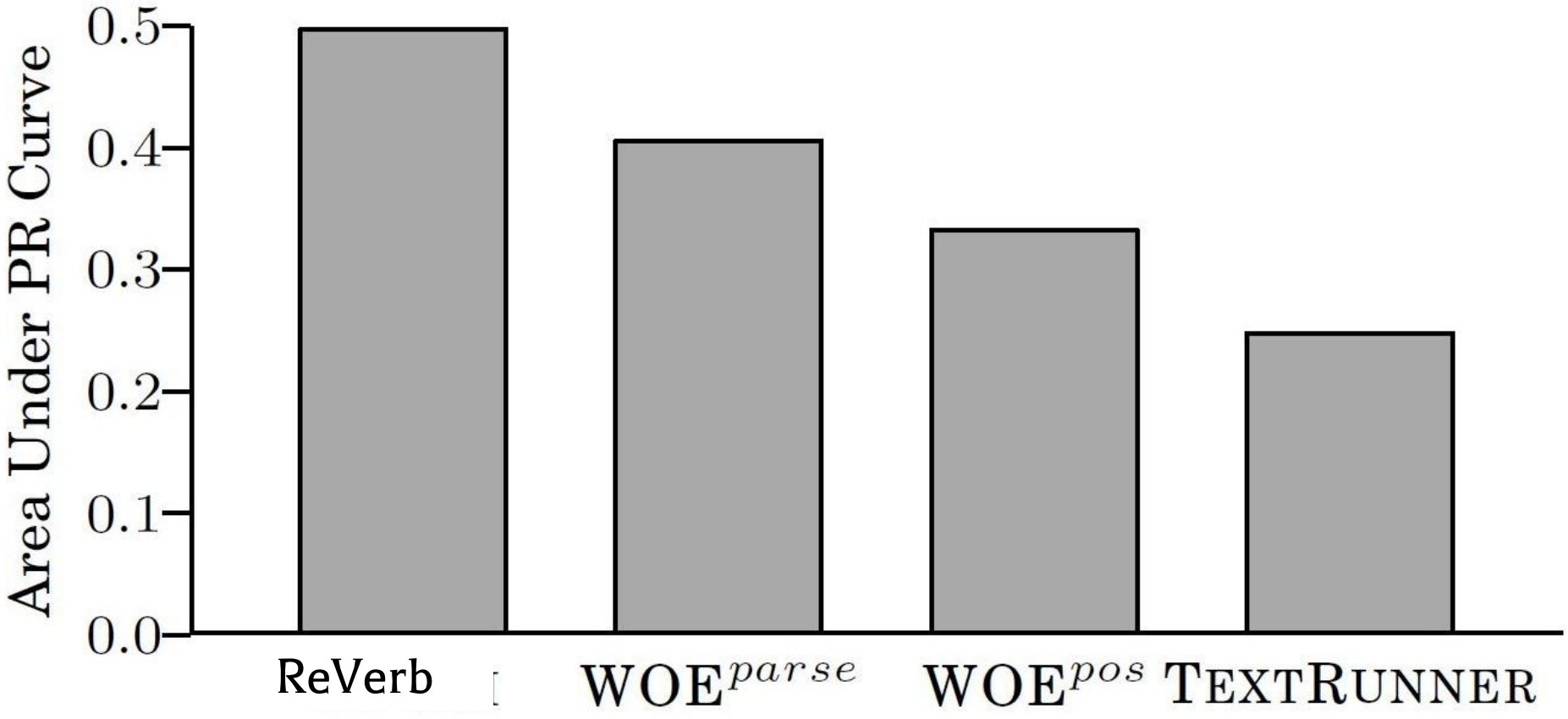
- Outputs more meaningful & informative relations

Homer made a deal with the devil.

TR (**Homer**, **made**, **deal**)

RVerb (**Homer**, **made a deal with**, **devil**)

Experiments: Relation Phrases



ReVerb Error Analysis:

- 65% cases where a relation phrase was correctly identified, but the argument-finding heuristics failed.
- Remaining cases were n-ary relation mistaken as a binary relation.

  For eg. extracting (I, gave, him) from the sentence “I gave him 15 photographs”.
- False negatives (52%) were due to the argument-finding heuristics choosing the wrong arguments, or failing to extract all possible arguments

ArgLearner: Motivating Examples

"The assassination of Franz Ferdinand, improbable as it may seem, began WWI."

(it, began, WWI)

"Republicans in the Senate filibustered an effort to begin debate on the jobs bill."

(the Senate, filibustered, an effort)

"The plan would reduce the number of teenagers who begin smoking."

(The plan, would reduce the number of, teenagers)

Analysis – arg1 substructure

Category	*Pattern*	*Freq*
Basic Noun Phrases ***Chicago was founded in 1833***	***NN, JJ NN, etc***	***65%***
Prepositional Attachments ***The forest in Brazil is threatened by ranching.***	***NP PP NP***	***19%***
List ***Google and Apple are headquartered in Silicon Valley.***	***NP, (NP,)* CC NP***	***15%***
Relative Clause ***Chicago, which is located in Illinois, has three million residents.***	***NP (that|WP|WDT)? NP? VP NP***	***<1%***

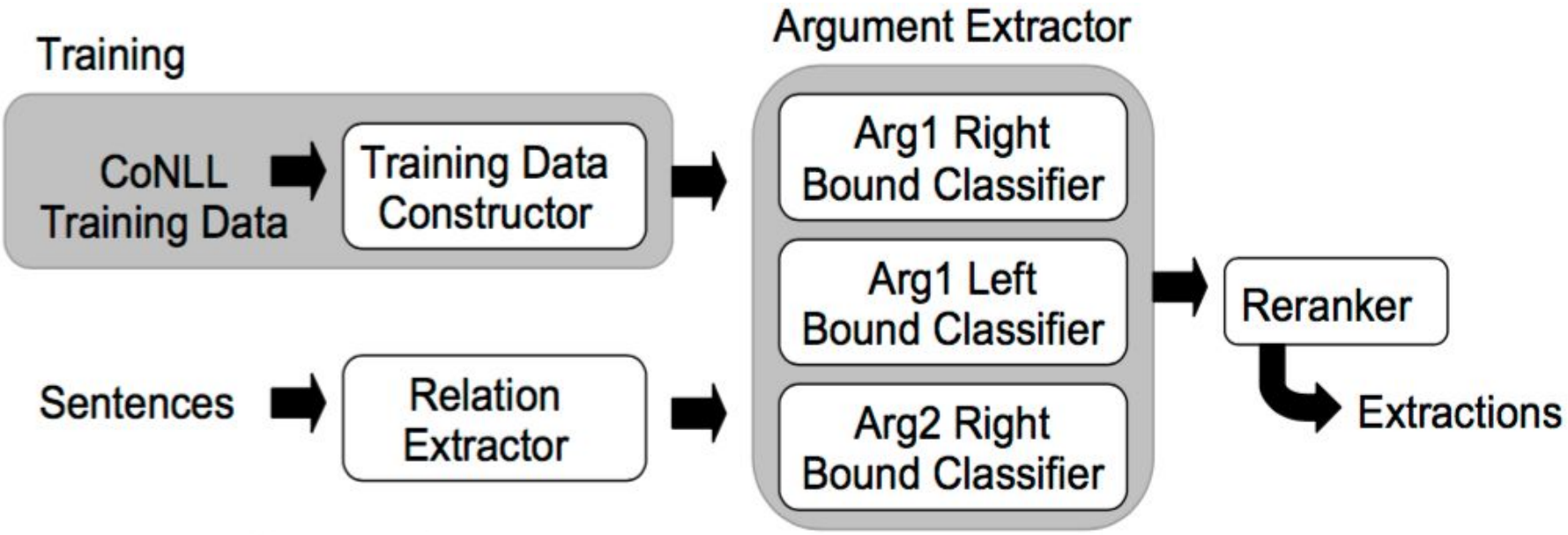
Analysis – arg2 substructure

Category	*Pattern*	*Freq*
Basic Noun Phrases ***Calcium prevents osteoporosis***	***NN, JJ NN, etc***	***60%***
Prepositional Attachments ***Barack Obama is one of the presidents of the United States***	***NP PP NP***	***18%***
List ***A galaxy consists of stars and stellar remnants***	***NP, (NP,)* CC NP***	***15%***
Independent Clause ***Scientists estimate that 80% of oil remains a threat.***	***(that|WP|WDT)? NP? VP NP***	***8%***
Relative Clause ***The shooter killed a woman who was running from the scene.***	***NP (that|WP|WDT)? NP? VP NP***	***6%***

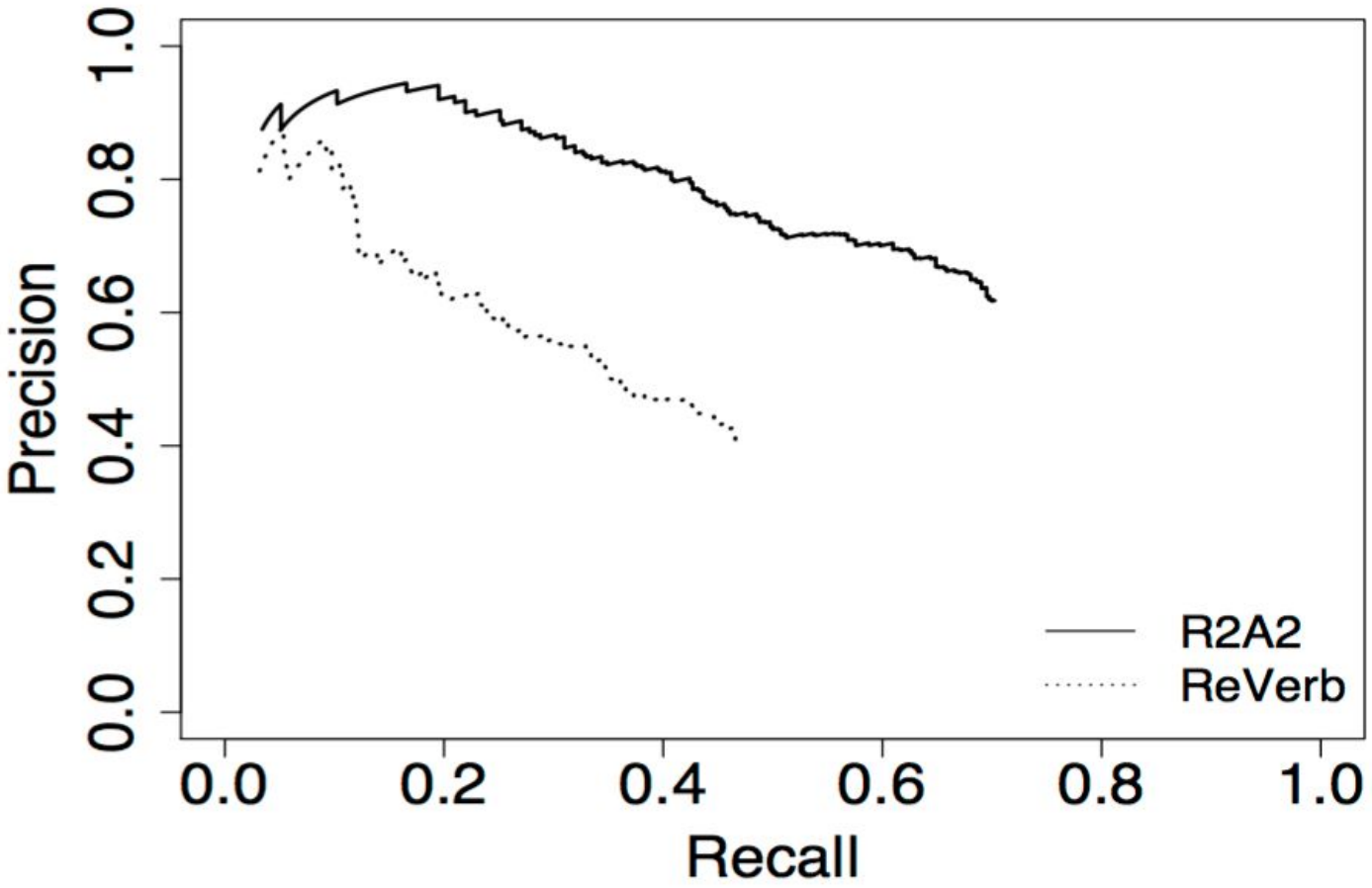
Argument Extraction Methodology

- Break problem into four parts:
  - Identify arg1 right bound
    … TOK TOK TOK TOK TOK rel TOK TOK TOK …
    Classifier (Weka's REPTree)
  - Identify arg1 left bound
    … TOK TOK TOK TOK TOK rel TOK TOK TOK …
    Classifier (CRF Mallet)
  - Identify arg2 left bound
    … TOK TOK TOK TOK TOK rel TOK TOK TOK …
  - Identify arg2 right bound
    … TOK TOK TOK TOK TOK rel TOK TOK TOK …
    Classifier (CRF Mallet)

ArgLearner's System Architecture



Evaluation



R2A2 has substantially higher recall and precision than REVERB.

Possible Extension:

"relation discovery from REVERB can be used as a component in NELL to get a NELL-REVERB hybrid that is better at extending its ontology. In contrast to REVERB, NELL has an aspect of temporality and can extract new/update existing entries from an evolving corpus." - Surag

"Temporality and context not addressed. Ollie incorporates context, but if something was factual at one point but is no longer factual, Ollie will still see it as factual, so temporality needs to be explored." - Akshay

"Ignores dependency parse information which can be used to provide long range context." - Akshay

"Many of the observations are for grammatically correct sentences, something which may not be taken for granted in Social Network platforms like Twitter. Extending this method to work on them might be an interesting task" - Barun

"Confidence for extractions could possibly be based on similarity of their Word2Vec vectors" - Gagan

"n-ary relations and relations not limited to verb. (addressed in OPENIE4)
Using more than POS and other syntactic features (SRL used in openIE4)" - Nupur

Thank You!

Error Analysis