Finding Better Argument Spans
Formulation, Crowdsourcing, and Prediction

Gabriel Stanovsky

Intro

Obama, the U.S president, was <u>born</u> in Hawaii

- Arguments are perceived as answering **role questions**
  - Who *was <u>born</u>* somewhere?
  - Where *was* someone *<u>born</u>*?

- Various predicate-argument annotations

  - PropBank
  - FrameNet
  - Recently - QA-SRL

- Open IE
  - ReVerb
  - OLLIE
  - Stanford Open IE

Background: QA-SRL

- Recently, He et al. (2015) suggested pred-arg annotation by explicitly asking and answering **argument role questions**

Obama, the U.S president, was **born in** Hawaii

- *Who was* ***born*** *somewhere?* Obama
- *Where was someone* ***born****?* Hawaii

Intro

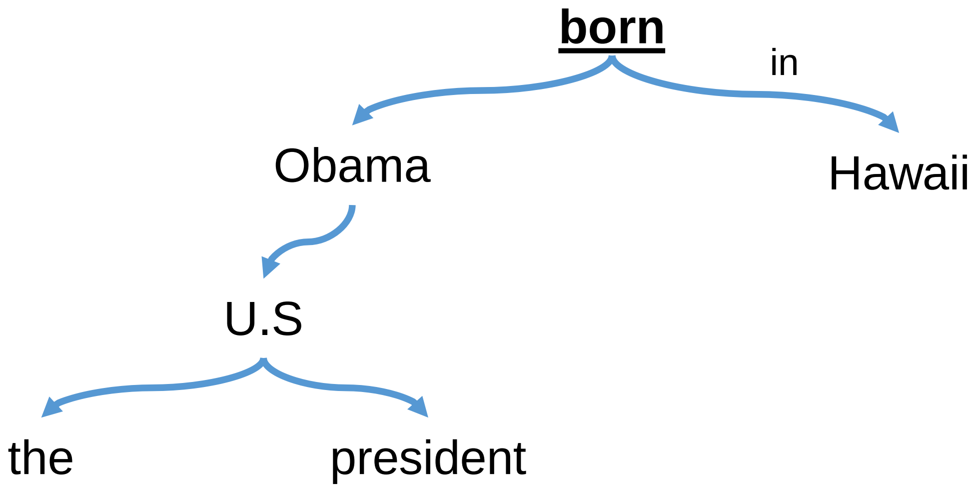
Obama, the U.S president, was <u>born</u> in Hawaii

- Given a predicate in a sentence –

  What is the **“best choice”** for the span of its arguments?

“Inclusive” Approach

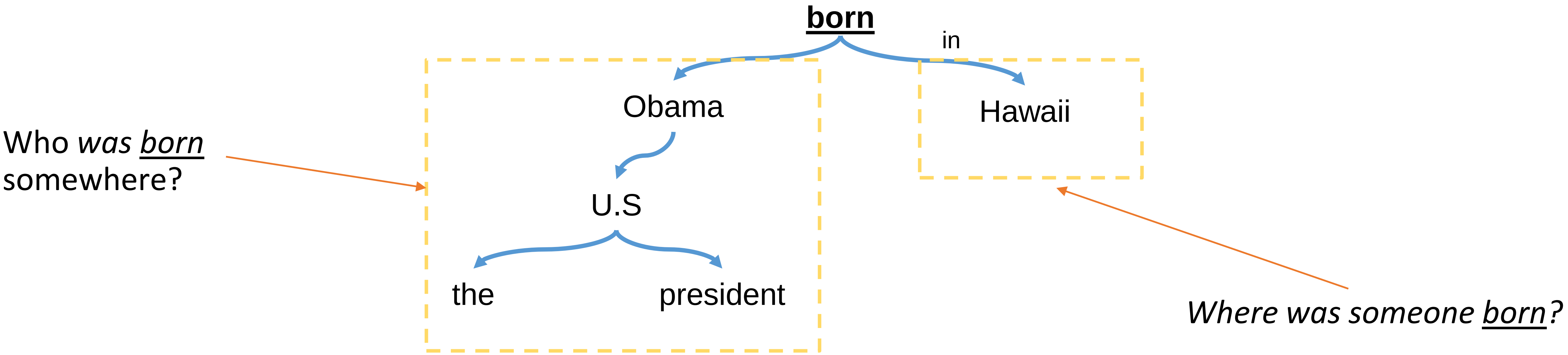
- Arguments are full syntactic constituents



- PropBank
- FrameNet
- AMR

“Inclusive” Approach

- Arguments are full syntactic constituents



- PropBank
- FrameNet
- AMR

“Minimalist” Approach

- Arguments are the shortest spans from which the entity is identifiable

Obama, the U.S president, was **born in** Hawaii
→ (Obama, **born in**, Hawaii)

- Open IE
  - ReVerb
  - OLLIE
  - Stanford Open IE

Motivation

- **Question answering**
  - Matching entities between questions and answers which might have different modifications
- **Abstractive summarization**
  - Remove non-integral modifications to shorten the sentence
- **Knowledge representation**
  - Minimally scoped arguments yields salient and recurring entities

Motivation

- **Shorter arguments** are beneficial for a wide variety of applications
  - Corro et al. (2013) Open-IE system which focused on shorter arguments
  - Angeli et al. (2015) State of the art TAC-KBP Slot Filling task
  - Stanovsky et al. (2015) Open-IE 4 in state of the art in lexical similarity

Previous Work

- No accepted Open IE guidelines
- No formal definition for a desired argument scope
- **No gold standard**

In this talk

- Formulation of an **argument reduction criterion**
  - Intuitive enough to be **crowdsourced**
- Automatic classification of **non-restrictive modification**
- Creating a large scale **gold standard for Open IE**

Annotating Reduced Argument Scope Using QA-SRL

Stanovsky, Dagan and Adler, ACL 2016

Formal Definitions

- Given:
  - p - predicate in a sentence
    - *Obama, the newly elected president, <u>flew</u> to Russia*
  - $a = \{w_1, \ldots w_n\}$ - non-reduced argument
    - *Barack Obama, the newly elected president*
  - $Q(p, a)$ - argument role question
    - *Who flew somewhere?*

Argument Reduction Criterion

$M(p, a)$- **a set of <u>minimally scoped</u> arguments, jointly answering $\boldsymbol{Q}$**

- Barack Obama, the 44th president, <u>congratulated</u> the boy who won the spelling bee
  - Q_1: Who <u>congratulated</u> someone? $M(Q_1)$: **Barack Obama**
  - Q_2: Who <u>was</u> congratulated? $M(Q_2)$: **the boy who won the spelling bee**

Expert Annotation Experiment

- Using **questions** annotated in QA-SRL
  - **Re-answer** according to the formal definition
  - Annotated 260 arguments in 100 predicates

Annotation	Argument	Word
Expert - IAA	94.6%	97.1%

Expert Annotation Experiment

- Using **questions** annotated in QA-SRL
  - **Re-answer** according to the formal definition
  - Annotated 260 arguments in 100 predicates

Annotation	Argument	Word
Expert - IAA	94.6%	97.1%

Our criterion can be consistently annotated by expert annotators

Reduction Operations

1. **Removal** of tokens from a
   => Omission of *non-restrictive modification*

2. **Splitting** a
   => Decoupling *distributive coordinations*

Restrictive vs. Non-Restrictive

- Restrictive
  - *She wore the necklace* ***that her mother gave her***
- Non – Restrictive
  - ***Obama ~~, the newly elected president,~~*** *flew to Russia*

Distributive vs. Non-Distributive

- Distributive
  - ***Obama*** *and* ***Clinton*** *were born in America*
- Non-Distributive
  - ***John and Mary*** *met at the university*

Distributive vs. Non-Distributive

- Distributive
  - ***Obama*** *and* ***Clinton*** *were born in America*
- Non-Distributive
  - ***John and Mary*** *met at the university*

V ***Obama was born in America***

V ***Clinton was born in America***

X ***John met at the university***

X ***Mary met at the university***

Comparison with PropBank

Arguments reduced	24%
Non-Restrictive	19%
Distributive	5%

The average reduced argument shrunk by 58%

Our annotation **significantly reduces** PropBank argument spans

Does QA-SRL Captures Minimality?

- QA-SRL guidelines do not specifically aim to minimize arguments
- Does the paradigm itself solicits shorter arguments?

Annotation	Argument	Word
Expert - IAA	94.6%	97.1%
QA-SRL - Expert	80%	88.5%

Does QA-SRL Captures Minimality?

- QA-SRL guidelines do not specifically aim to minimize arguments
- Does the paradigm itself solicits shorter arguments?

Annotation	Argument	Word
Expert - IAA	94.6%	97.1%
QA-SRL - Expert	80%	88.5%

Our criterion is **captured to a good extent in QA-SRL**

Can We Do Better?

- Using **turkers** to repeat the re-answering experiment
  - Asked annotators to specify the shortest possible answer from which the entity is identifiable

Annotation	Argument	Word
Expert - IAA	94.6%	97.1%
QA-SRL - Expert	80%	88.5%
Our Crowdsourcing - Expert	**89.1%**	**93.5%**

Can We Do Better?

- Annotators are asked to specify the **shortest possible answer from which the entity is identifiable**

Annotation	Argument	Word
Expert - IAA	94.6%	97.1%
QA-SRL - Expert	80%	88.5%
Our Crowdsourcing - Expert	**89.1%**	**93.5%**

Focused guidelines can get **more consistent** argument spans

To Conclude this Part…

- We formulated an **argument reduction criterion**
- Shown to be:
  - Consistent enough for **expert annotation**
  - Intuitive enough to be annotated by **crowdsourcing**
  - Captured in the QA-SRL paradigm

Annotating and Predicting Non-Restrictive Modification

Stanovsky and Dagan, ACL 2016

Different types of NP modifications

(from Huddleston et.al)

- Restrictive modification
  - The content of the modifier is an **integral part** of the meaning of the containing clause
  - AKA: integrated (Huddleston)
- Non-restrictive modification
  - The modifier presents an **separate or additional unit of information**
  - AKA: supplementary (Huddleston), appositive, parenthetical

	Restrictive	Non-Restrictive
Relative Clause	She took the necklace **that her mother gave her**	The speaker thanked president Obama **who just came back from Russia**
Infinitives	People **living near the site** will have to be evacuated	Assistant Chief Constable Robin Searle, **sitting across from the defendant**, said that the police had suspected his involvement since 1997.
Appositives		Keeping the Japanese happy will be one of the most important tasks facing **conservative leader** Ernesto Ruffo
Prepositional modifiers	the kid **from New York** rose to fame	Franz Ferdinand **from Austria** was assassinated om Sarajevo
Postpositive adjectives	George Bush’s **younger brother** lost the primary	Pierre Vinken, **61 years old**, was elected vice president
Prenominal adjectives	The **bad** boys won again	The water rose a **good** 12 inches

Goals

- Create a **large corpus** annotated with non-restrictive NP modification
  - Consistent with gold dependency parses
- **Automatic prediction** of non-restrictive modifiers
  - Using lexical-syntactic features

Previous work

- Rebanking CCGbank for improved NP interpretation
  (Honnibal, Curran and Bos, ACL ‘10)
  - Added automatic non-restrictive annotations to the CCGbank
  - Simple punctuation implementation
    - Non restrictive modification ←→ The modifier is preceded by a comma
  - No intrinsic evaluation

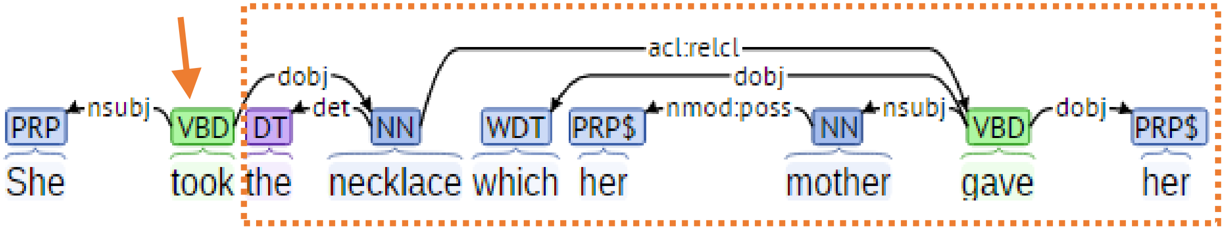
Previous work

- Relative clause extraction for syntactic simplification
  (Dornescu et al., COLING ‘14)
  - Trained annotators marked spans as restrictive or non-restrictive
  - Conflated argument span with non-restrictive annotation
    - This led to low inter-annotator-agreement
      - **Pairwise F1 score of 54.9%**
    - Develop rule based and ML baselines (CRF with chunking feat.)
      - **Both performing around ~47% F1**

Our Approach

Consistent corpus with QA based classification

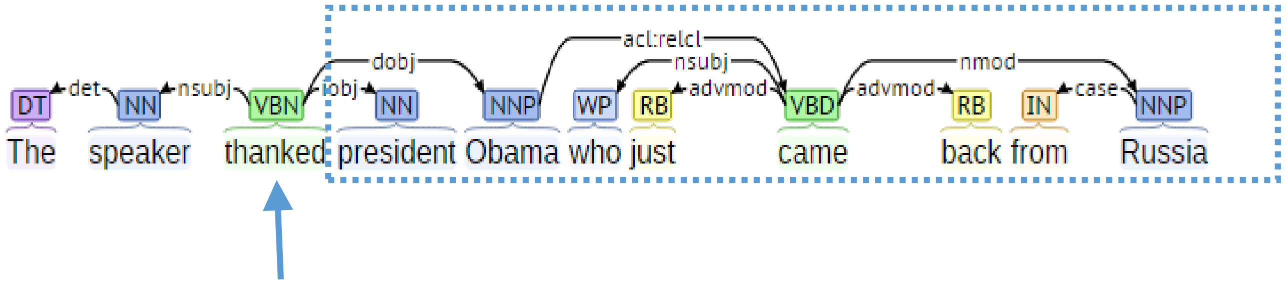
1. Traverse the syntactic tree from predicate to NP arguments
2. Phrase an argument role question, which is answered by the NP (*what? who? to whom? Etc.*)
3. For each candidate modifier (= syntactic arc) - check whether when omitting it the NP still provides the same answer to the argument role question



What did someone take?

X The necklace ~~which her mother gave her~~

Restrictive



Who was thanked by someone?

V President Obama ~~who just came back from Russia~~

Non-restrictive

Crowdsourcing

- This seems fit for crowdsourcing:
  - **Intuitive** - Question answering doesn’t require linguistic training
  - **Binary decision** – Each decision directly annotates a modifier

Corpus

- CoNLL 2009 dependency corpus
  - Recently annotated by QA-SRL -- we can borrow most of their role questions
- Each NP is annotated on Mechanical Turk
  - Five annotators for 5c each
- Final annotation by majority vote

Expert annotation

- Reusing our previous expert anntoation, we can assess if crowdsourcing captures non-restrictiveness

- Agreement
  - Kappa = 73.79 (substantial agreement)
  - F1 =85.6

Candidate Type Distribution

	#instances	%Non-Restrictive	Agreement (K)
Prepositive adjectival modifiers	677	41%	74.7
Prepositions	693	36%	61.65
Appositions	342	73%	60.29
Non-Finite modifiers	279	68%	71.04
Prepositive verbal modifiers	150	69%	100
Relative Clauses	43	79%	100
Postpositive adjectival modifiers	7	100%	100
Total	**2191**	**51.12%**	**73.79**

- The annotation covered 1930 NPs in 1241 sentences

Candidate Type Distribution

	#instances	%Non-Restrictive	Agreement (K)
Prepositive adjectival modifiers	677	41%	74.7
Prepositions	693	36%	61.65
Appositions	342	73%	60.29
Non-Finite modifiers	279	68%	71.04
Prepositive verbal modifiers	150	69%	100
Relative Clauses	43	79%	100
Postpositive adjectival modifiers	7	100%	100
Total	**2191**	**51.12%**	**73.79**

- Prepositions and appositions are harder to annotate

Candidate Type Distribution

	#instances	%Non-Restrictive	Agreement (K)
Prepositive adjectival modifiers	677	41%	74.7
Prepositions	693	36%	61.65
Appositions	342	73%	60.29
Non-Finite modifiers	279	68%	71.04
Prepositive verbal modifiers	150	69%	100
Relative Clauses	43	79%	100
Postpositive adjectival modifiers	7	100%	100
Total	**2191**	**51.12%**	**73.79**

- The corpus is balanced between the two classes

Predicting non-restrictive modification

- CRF features:
  - Dependency relation
  - NER
    - Modification of named entity tend to be non-restrictive
  - Word embeddings
    - Contextually similar words will have similar restricteness value
  - Linguistically motivated features
    - The word introducing the modifier,
      - “that” indicates restrictive, while a wh-pronoun as indicates non-restrictive (Huddleston)

Results

Modifier Type	#	Precision			Recall			F1		
		Honnibal	Dornescu	Our	Honnibal	Dornescu	Our	Honnibal	Dornescu	Our
Prepositional	135	.83	.67	.69	.1	.16	.41	.18	.26	.51
Adjectival	111	.33	.38	.59	.06	.06	.21	.11	.11	.31
Appositive	78	.77	.81	.82	.34	.93	.98	.47	.87	.89
Non-Finite	55	.77	.63	.64	.29	.97	.97	.42	.76	.77
Verbal	20	0	.75	.75	0	1	1	0	.86	.86
Relative clause	13	1	.85	.85	.27	1	1	.43	.92	.92
Total	412	.72	.72	**.73**	.19	.58	**.68**	.3	.64	**.72**

Results

Modifier Type	#	Precision			Recall			F1		
		Honnibal	Dornescu	Our	Honnibal	Dornescu	Our	Honnibal	Dornescu	Our
Prepositional	135	.83	.67	.69	.1	.16	.41	.18	.26	.51
Adjectival	111	.33	.38	.59	.06	.06	.21	.11	.11	.31
Appositive	78	.77	.81	.82	.34	.93	.98	.47	.87	.89
Non-Finite	55	.77	.63	.64	.29	.97	.97	.42	.76	.77
Verbal	20	0	.75	.75	0	1	1	0	.86	.86
Relative clause	13	1	.85	.85	.27	1	1	.43	.92	.92
Total	412	.72	.72	**.73**	.19	.58	**.68**	.3	.64	**.72**

Prepositions and adjectives are harder to predict

Results

Modifier Type	#	Precision			Recall			F1		
		Honnibal	Dornescu	Our	Honnibal	Dornescu	Our	Honnibal	Dornescu	Our
Prepositional	135	.83	.67	.69	.1	.16	.41	.18	.26	.51
Adjectival	111	.33	.38	.59	.06	.06	.21	.11	.11	.31
Appositive	78	.77	.81	.82	.34	.93	.98	.47	.87	.89
Non-Finite	55	.77	.63	.64	.29	.97	.97	.42	.76	.77
Verbal	20	0	.75	.75	0	1	1	0	.86	.86
Relative clause	13	1	.85	.85	.27	1	1	.43	.92	.92
Total	412	.72	.72	**.73**	.19	.58	**.68**	.3	.64	**.72**

Commas are good in precision but poor for recall

Results

Modifier Type	#	Precision			Recall			F1		
		Honnibal	Dornescu	Our	Honnibal	Dornescu	Our	Honnibal	Dornescu	Our
Prepositional	135	.83	.67	.69	.1	.16	.41	.18	.26	.51
Adjectival	111	.33	.38	.59	.06	.06	.21	.11	.11	.31
Appositive	78	.77	.81	.82	.34	.93	.98	.47	.87	.89
Non-Finite	55	.77	.63	.64	.29	.97	.97	.42	.76	.77
Verbal	20	0	.75	.75	0	1	1	0	.86	.86
Relative clause	13	1	.85	.85	.27	1	1	.43	.92	.92
Total	412	.72	.72	**.73**	.19	.58	**.68**	.3	.64	**.72**

Dornescu et al. performs better on our dataset

Results

Modifier Type	#	Precision			Recall			F1		
		Honnibal	Dornescu	Our	Honnibal	Dornescu	Our	Honnibal	Dornescu	Our
Prepositional	135	.83	.67	.69	.1	.16	.41	.18	.26	.51
Adjectival	111	.33	.38	.59	.06	.06	.21	.11	.11	.31
Appositive	78	.77	.81	.82	.34	.93	.98	.47	.87	.89
Non-Finite	55	.77	.63	.64	.29	.97	.97	.42	.76	.77
Verbal	20	0	.75	.75	0	1	1	0	.86	.86
Relative clause	13	1	.85	.85	.27	1	1	.43	.92	.92
Total	412	.72	.72	**.73**	.19	.58	**.68**	.3	.64	**.72**

Our system highly improves recall

To Conclude this part…

- A large non-restrictive gold standard
  - Directly augments dependency trees
- Automatic classifier
  - Improves over state of the art results

Creating a Gold Benchmark for Open IE

Stanovsky and Dagan, EMLP 2016 (hopefully!)

Teaser!

Open Information Extraction

- Extracts SVO tuples from texts
  - Barack Obama, the U.S president, was **born in** Hawaii
    → (Barack Obama, **born in**, Hawaii)
  - Clinton and Bush were born in America
    → (Clinton , born in, America), (Bush , born in, America)

- Used in various applications for populating large databases from raw open domain texts
  - A scalable and open variant of the Information Extraction task

Open IE Evaluation

- Open IE task formulation has been lacking formal rigor
  - No common guidelines → **No large corpus for evaluation**

- Annotators examine *a small sample* of their system's output and judge it according to *some* guidelines

→ **Precision oriented** metrics

→ Numbers are **not comparable**

→ Experiments are **hard to reproduce**

Goal

- In this work we -
  - Analyze common evaluation principles in prominent recent work

  - Create a **large gold standard corpus** which follows these principles
    - Uses previous annotation efforts
    - Provides both precision **and recall** metrics

  - Automatically evaluate the performance of the most prominent OIE systems on our corpus
    - First automatic & comparable OIE evaluation
    - Future systems can easily compare themselves

Converting QA-SRL to Open IE

- Intuition:
  - All of the QA pairs over a single predicate in QA-SRL correspond to a single Open IE extraction
- Example:
  - "**Barack Obama**, the newly elected president, **flew** **to Moscow** **on Tuesday**"
  - QA-SRL:
    - Who **flew** somewhere? **Barack Obama**
    - Where did someone **fly**? **to Moscow**
    - When did someone **fly**? **on Tuesday**

    → (Barack Obama, **flew**, to Moscow, on Tuesday)

Example

- John Bryce, Microsoft's head of marketing **refused** to **greet** Arthur Black
  - Who refused something? John Bryce
  - Who refused something? Microsoft's head of marketing
  - What did someone refuse to do? **greet** Arthur Black
  - Who was not greeted? Arthur Black
  - Who did not greet someone? John Bryce

→

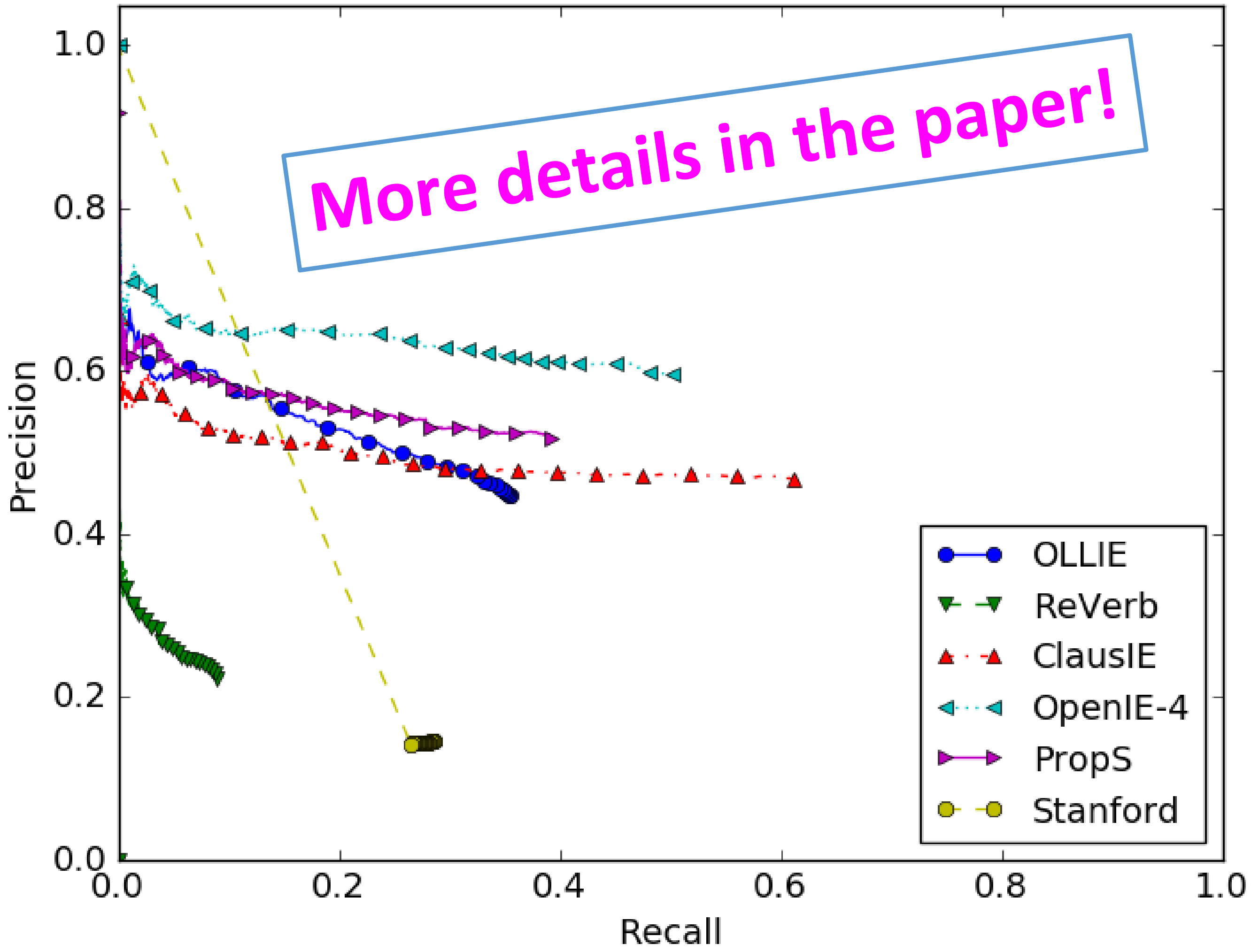
(John Bryce, **refused to greet**, Arthur Black),
(Microsoft's head of Marketing , **refused to greet**, Arthur Black)

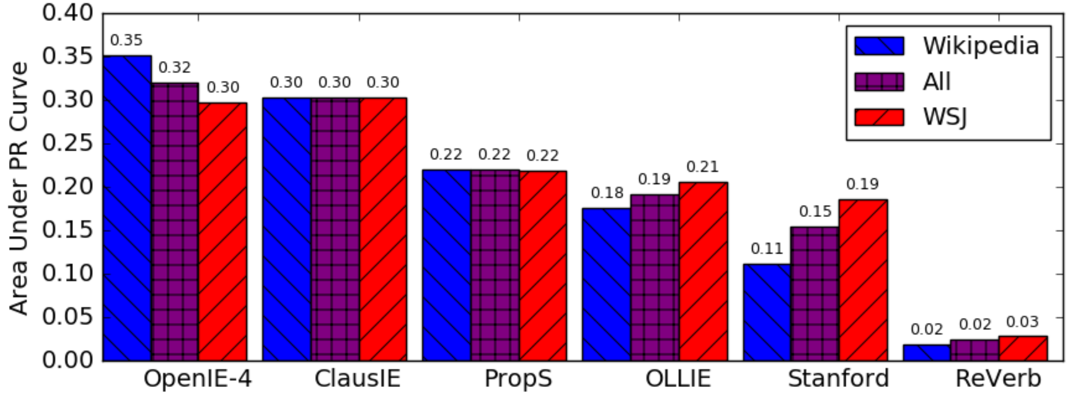
Resulting Corpus

Corpus	**WSJ**	**WIKI**	**All**
#Sentences	1241	1959	3200
#Predicates	2020	5690	7710
#Questions	8112	10798	18910
#Extractions	**4481**	**5878**	**10359**

- 13 times bigger than largest previous corpus (ReVerb)

Evaluations: PR-Curve





- Stanford – Assigns a probability of 1 to most of its extractions (94%)
- Low Recall
  - Most missed extractions seem to come from questions with multiple answers (usually **long range dependencies**)
- Low Precision
  - Allowing for softer matching functions (lowering threshold), raises precision and **keeps the same trends**

Conclusions

- We discussed a framework for argument annotation:
  - Formal Definition
  - Expert and crowdsource annotation
  - Automatic prediction
  - Automatic conversion from quality annotations

Conclusions

- We discussed a framework for argument annotation:
  - Formal Definition
  - Expert and crowdsource annotation
  - Automatic prediction
  - Automatic conversion from quality annotations

Thanks For Listening!