

# Numerical Relation Extraction from the Web

## MTP Stage 1 Presentation

Aman Madaan    Ashish Mittal

Indian Institute of Technology Bombay, Mumbai

22nd October, 2014

# Outline

1 Motivation

2 Problem Statement

3 Relation Extraction as a Machine Learning Problem

4 Distant Supervision

5 Distant Supervision Techniques

6 TAC Submission

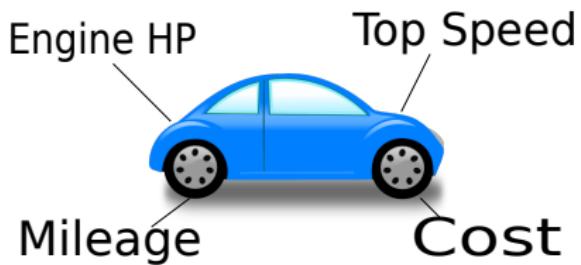
7 Numerical Relation Extraction

8 Units in Numerical Relation Extraction

# Entities have Numerical Attributes



# Entities have Numerical Attributes



# Entities and Numerical Attributes

- For popular entities, finding complete knowledge bases is possible.

# Entities and Numerical Attributes

- For popular entities, finding complete knowledge bases is possible.
- data.worldbank.org, Wikipedia infoboxes, freebase ...

# Entities and Numerical Attributes

- For popular entities, finding complete knowledge bases is possible.
- data.worldbank.org, Wikipedia infoboxes, freebase ...
- What about less popular entities?

# Entities and Numerical Attributes

- For popular entities, finding complete knowledge bases is possible.
- data.worldbank.org, Wikipedia infoboxes, freebase ...
- What about less popular entities?
  - What is the population of Arbit Apartments, Powai?

# Entities and Numerical Attributes

- For popular entities, finding complete knowledge bases is possible.
- data.worldbank.org, Wikipedia infoboxes, freebase ...
- What about less popular entities?
  - What is the population of Arbit Apartments, Powai?
  - What is the GDP of Sugarcane Industry of India?

# Entities and Numerical Attributes

- For popular entities, finding complete knowledge bases is possible.
- data.worldbank.org, Wikipedia infoboxes, freebase ...
- What about less popular entities?
  - What is the population of Arbit Apartments, Powai?
  - What is the GDP of Sugarcane Industry of India?
  - Percent of Internet users in Mumbai?

# Motivation

Activities Google Chrome ▾ Mon Oct 20, 7:42 PM enz ▾

Internet users in India x https://www.google.co.in/search?q=internet+users+in+india&oq=internet+use&aqs=chrome.0.69i59l2j69i57.1830j0j1&sourceid=chrome&es\_sm=93&ie=UTF-8#q=ir Q ⌂

Maths ML dev awesome smtp pblog Ganglia: YH Grid R YH cs215 Nucleus tml stats ML/DS Reading NLP shortcuts Other bookmarks

internet users in india

Google Sign in

Web News Images Maps Videos More Search tools

About 2,49,00,000 results (0.23 seconds)

**15.1% of the population (2013)**  
India, Internet users

Year	Internet Users (%)
1990	~0.1
1995	~0.2
2000	~0.5
2005	~3.0
2010	~6.0
2013	15.1

Explore more

Sources Include: World Bank

Feedback

**India**  
Country  
India, officially the Republic of India, is a country in South Asia. It is the seventh-largest country by area, the second-most populous country with over 1.2 billion people, and the most populous democracy in the world.  
Wikipedia

**Related statistics**

Statistic	Value
Population	1.252 billion (2013)
Life expectancy	66.21 years (2012)

# Motivation

A screenshot of a Google Chrome browser window. The address bar shows the URL [https://www.google.co.in/search?q=internet+users+in+india&oq=internet+use&aqs=chrome.0.69i59l2j69i57.1830j0j1&sourceid=chrome&es\\_sm=93&ie=UTF-8#q=ir](https://www.google.co.in/search?q=internet+users+in+india&oq=internet+use&aqs=chrome.0.69i59l2j69i57.1830j0j1&sourceid=chrome&es_sm=93&ie=UTF-8#q=ir). The search query "internet users in mumbai" is entered in the search bar. The results page shows several news articles and reports about Mumbai's internet user statistics. The top result is from Internet Society, stating that Mumbai has more internet users than any other city in India. Other results mention IMAI data, Indian Express, and FirstPost, all reporting that Mumbai leads in internet penetration.

Activities Google Chrome ▾ enz Mon Oct 20, 7:37 PM

Internet users in mun x

Maths ML dev awesome sgmt pblog Ganglia: YH Grid R YH cs215 Nucleus tml stats ML/DS Reading NLP shortcuts Other bookmarks

internet users in mumbai

Sign in

Google

Web Images News Maps Videos More Search tools

About 26,70,000 results (0.41 seconds)

**Global Internet Users - Internet Penetration by Country**

Ad [www.internetsociety.org/Internet-Users](http://www.internetsociety.org/Internet-Users)

View the Map & Download the Report  
Download the Report - Introducing the Report - Executive Summary

**Mumbai has most number of internet users: IMAI data - Th...**

timesofindia.indiatimes.com › Tech

Nov 7, 2013 - At 12 million, Mumbai has more internet users than any other city in the country, according to data released by the Internet And Mobile ...

**Internet penetration in India: Mumbai tops Internet users list ...**

archive.indianexpress.com/.../internet...mumbai...internet-users.../119189... Nov 7, 2013 - With 12 million active internet users, Mumbai has emerged as the city with the highest number of internet users in the country, followed by Delhi ...

**With 12 million Internet users, Mumbai tops list of connected...**

tech.firstpost.com/.../with-12-million-internet-users-mumbai-tops-list-of... Nov 8, 2013 - India's economic capital, Mumbai, has emerged to be the city with the highest number of Internet users in India, ahead of cities like New Delhi ...

**Mumbai has the most number of Internet users in India ...**

tech.firstpost.com/.../mumbai-has-the-most-number-of-internet-users-in-i... ▾

# Motivation

- Web is huge.

# Motivation

- Web is huge.
- Probably, there is some page which contains the information we are looking for.

# Motivation

- Web is huge.
- Probably, there is some page which contains the information we are looking for.
- The way in which you express a fact about an entity depends on the fact, and not the entity.

# Motivation

- Web is huge.
- Probably, there is some page which contains the information we are looking for.
- The way in which you express a fact about an entity depends on the fact, and not the entity.
- We may expect **the sentence structure**<sup>1</sup> to be similar.

# Motivation

- Web is huge.
- Probably, there is some page which contains the information we are looking for.
- The way in which you express a fact about an entity depends on the fact, and not the entity.
- We may expect **the sentence structure**<sup>1</sup> to be similar.
  - Population of India reached 1.3 billion, making it the second largest country in the world.

# Motivation

- Web is huge.
- Probably, there is some page which contains the information we are looking for.
- The way in which you express a fact about an entity depends on the fact, and not the entity.
- We may expect **the sentence structure**<sup>1</sup> to be similar.
  - Population of India reached 1.3 billion, making it the second largest country in the world.
  - Population of Arbit Apartments, Powai reached 1300.

# Outline

- 1 Motivation
- 2 Problem Statement
- 3 Relation Extraction as a Machine Learning Problem
- 4 Distant Supervision
- 5 Distant Supervision Techniques
- 6 TAC Submission
- 7 Numerical Relation Extraction
- 8 Units in Numerical Relation Extraction

# Problem Statement

- Given that we know a lot of facts about some entities, can we train extractors that run over the web and pull similar facts about other entities?

# Problem Statement

- The knowledge is scattered in unstructured text on the Web.

590.56 million people in China were using the internet at mid-2013, an increase of nearly 53 million (or 9.85%) from a year earlier.

The land area of the contiguous United States is 2,959,064 square miles (7,663,941 km<sup>2</sup>). Alaska, separated from the contiguous United States by Canada, is the largest state at 663,268 square miles (1,717,856 km<sup>2</sup>). Hawaii, occupying an archipelago in the central Pacific, southwest of North America, is 10,931 square miles (28,311 km<sup>2</sup>) in area.<sup>[136]</sup>

- Can such facts be extracted automatically?

# Problem Statement

- Formally, train extractors that can harness the Web for numerical relations, where relations are 3-tuples linking an entity to a number
  - (India, **economy**, 1.842 trillion USD)
  - (China, **internet users**, 590.56 million)
  - (USA, **land area**, 2,959,054 square mile)

# Outline

- 1 Motivation
- 2 Problem Statement
- 3 Relation Extraction as a Machine Learning Problem
- 4 Distant Supervision
- 5 Distant Supervision Techniques
- 6 TAC Submission
- 7 Numerical Relation Extraction
- 8 Units in Numerical Relation Extraction

# Relation Extraction as a Machine Learning Problem

- **Structure** and **content** of sentences expressing the same relations can be *expected* to be similar.
  - The population of Australia is estimated to be 23,622,400 as of 7 October 2014.
  - According to an official estimate for 1 June 2014, the population of Russia is 143,800,000.

# Relation Extraction as a Machine Learning Problem

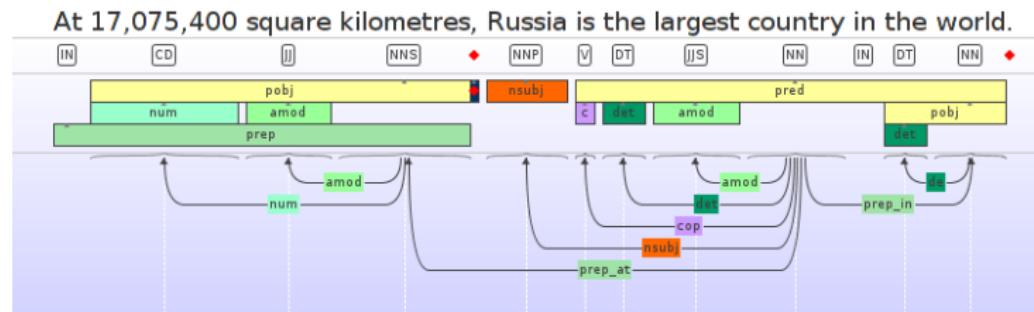
- **Structure** and **content** of sentences expressing the same relations can be *expected* to be similar.
  - At 17,075,400 square kilometres, Russia is the largest country in the world.
  - With an area of 504,030  $km^2$ , Spain is the second largest country in Western Europe.

# Relation Extraction as a Machine Learning Problem

- Redundancy in grammatical features and dependencies of the sentences expressing same relation.

# Relation Extraction as a Machine Learning Problem

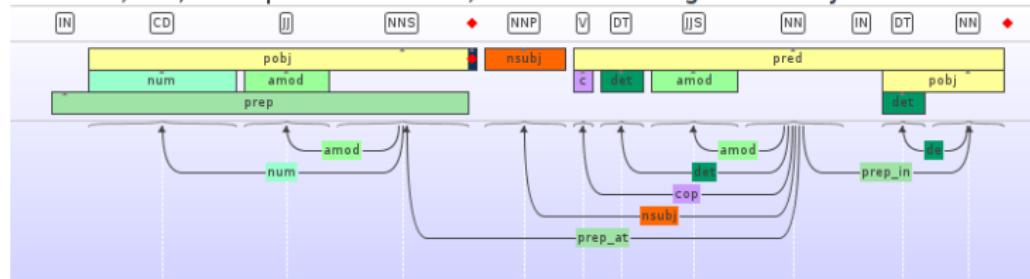
- Redundancy in grammatical features and dependencies of the sentences expressing same relation.



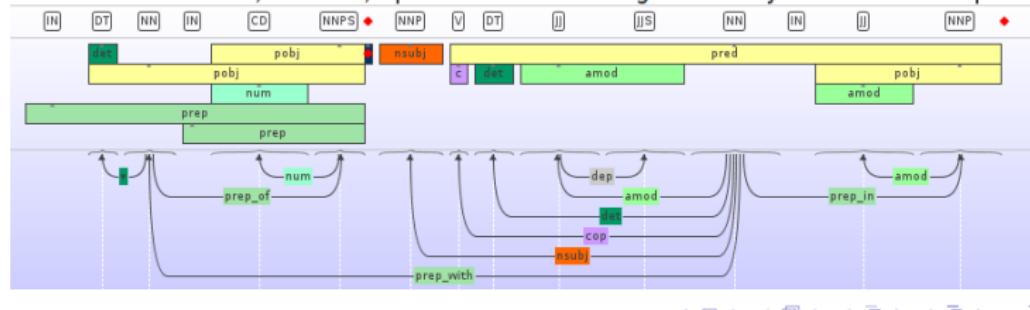
# Relation Extraction as a Machine Learning Problem

- Redundancy in grammatical features and dependencies of the sentences expressing same relation.

At 17,075,400 square kilometres, Russia is the largest country in the world.



With an area of 504,030 km<sup>2</sup>, Spain is the second largest country in Western Europe.



# Possible Workflow

---

## Relation Extraction

---

- 1: **Collect enough examples** for each relation so that there are sufficient patterns and enough redundancy to exploit.
  - 2: **Extract features** (important keywords, grammatical structure, parse trees, etc.) for these sentences.
  - 3: **Train** a multi-class classifier on this training data.
  - 4: **for** *sentence s* **in** *Corpus* **do**
  - 5:   **Extract** features for *s*.
  - 6:   **Predict** the relation using the model for these features.
  - 7:   **Store** the fact into a database.
  - 8: **end for**
-

# Challenge

- Large Corpus ( $\sim$ 16 million sentences), hand labeling is out of question

# Challenge

- Large Corpus ( $\sim$ 16 million sentences), hand labeling is out of question
- Need lots of training data to train high quality extractors

# Challenge

- Large Corpus ( $\sim$ 16 million sentences), hand labeling is out of question
- Need lots of training data to train high quality extractors
- Is there a middle ground?

# Outline

- 1 Motivation
- 2 Problem Statement
- 3 Relation Extraction as a Machine Learning Problem
- 4 Distant Supervision
- 5 Distant Supervision Techniques
- 6 TAC Submission
- 7 Numerical Relation Extraction
- 8 Units in Numerical Relation Extraction

# Distant Supervision

# Distant Supervision

- *Weak Supervision*

# Distant Supervision

- *Weak Supervision*
- Use Heuristics to **align** a **table** of facts with the **corpus**

# Distant Supervision

- *Weak Supervision*
- Use Heuristics to **align** a **table** of facts with the **corpus**
- *Fuzzy training data*

# Distant Supervision

## Example

- Born - In database

Donald Knuth	Wisconsin
Srinivasa Ramanujan	Erode
Alan Turing	London

- Given Sentences

- Srinivasa Ramanujan was born in his maternal grandmother's home in Erode.
- Srinivasa Ramanujan was born in Erode, Tamilnadu, India, on 22nd December, 1887.
- Turing's father was with the Indian Civil Service (ICS) at Chhatrapur, Bihar.
- Alan Turing biopic The Imitation Game named as London film festival opener.

# Distant Supervision

## Example

- Born - In database

Donald Knuth	Wisconsin
Srinivasa Ramanujan	Erode
Alan Turing	London

- Given Sentences

- Srinivasa Ramanujan was born in his maternal grandmother's home in Erode. LABEL : BornIn ✓
- Srinivasa Ramanujan was born in Erode, Tamilnadu, India, on 22nd December, 1887.
- Turing's father was with the Indian Civil Service (ICS) at Chhatrapur, Bihar.
- Alan Turing biopic The Imitation Game named as London film festival opener.

# Distant Supervision

## Example

- Born - In database

Donald Knuth	Wisconsin
Srinivasa Ramanujan	Erode
Alan Turing	London

- Given Sentences

- Srinivasa Ramanujan was born in his maternal grandmother's home in Erode. ✓
- Srinivasa Ramanujan was born in Erode, Tamilnadu, India, on 22nd December, 1887. ✓
- Turing's father was with the Indian Civil Service (ICS) at Chhatrapur, Bihar.
- Alan Turing biopic The Imitation Game named as London film festival opener.

# Distant Supervision

## Example

- Born - In database

Donald Knuth	Wisconsin
Srinivasa Ramanujan	Erode
Alan Turing	London

- Given Sentences

- Srinivasa Ramanujan was born in his maternal grandmother's home in Erode. ✓
- Srinivasa Ramanujan was born in Erode, Tamilnadu, India, on 22nd December, 1887. ✓
- **Turing's father was with the Indian Civil Service (ICS) at Chhatrapur, Bihar X**
- Alan Turing biopic The Imitation Game named as London film festival opener.

# Distant Supervision

## Example

- Born - In database

Donald Knuth	Wisconsin
Srinivasa Ramanujan	Erode
Alan Turing	London

- Given Sentences

- Srinivasa Ramanujan was born in his maternal grandmother's home in Erode. ✓
- Srinivasa Ramanujan was born in Erode, Tamilnadu, India, on 22nd December, 1887. ✓
- Turing's father was with the Indian Civil Service (ICS) at Chhatrapur, Bihar. X
- Alan Turing biopic The Imitation Game named as London film festival opener. ✓

# Distant Supervision

## Example

- Born - In database

Donald Knuth	Wisconsin
Srinivasa Ramanujan	Erode
Alan Turing	London

- Given Sentences

- Srinivasa Ramanujan was born in his maternal grandmother's home in Erode. ✓
- Srinivasa Ramanujan was born in Erode, Tamilnadu, India, on 22nd December, 1887. ✓
- Turing's father was with the Indian Civil Service (ICS) at Chhatrapur, Bihar. X
- Alan Turing biopic The Imitation Game named as London film festival opener.✓ FALSE POSITIVE

# Distant Supervision

# Distant Supervision

- *Distant supervision assumption:* If an entity pair  $E$  appears with a relation  $R$  in the knowledge base, then any sentence containing  $E$  **will** express  $R$

# Distant Supervision

- *Distant supervision assumption:* If an entity pair  $E$  appears with a relation  $R$  in the knowledge base, then any sentence containing  $E$  **will** express  $R$
- Can quickly label huge corpora

# Distant Supervision

- *Distant supervision assumption:* If an entity pair  $E$  appears with a relation  $R$  in the knowledge base, then any sentence containing  $E$  **will** express  $R$
- Can quickly label huge corpora
- Same entity pair can match different relations (Founded(Steve Jobs, Apple) or CEO(Steve Jobs, Apple))

# Distant Supervision

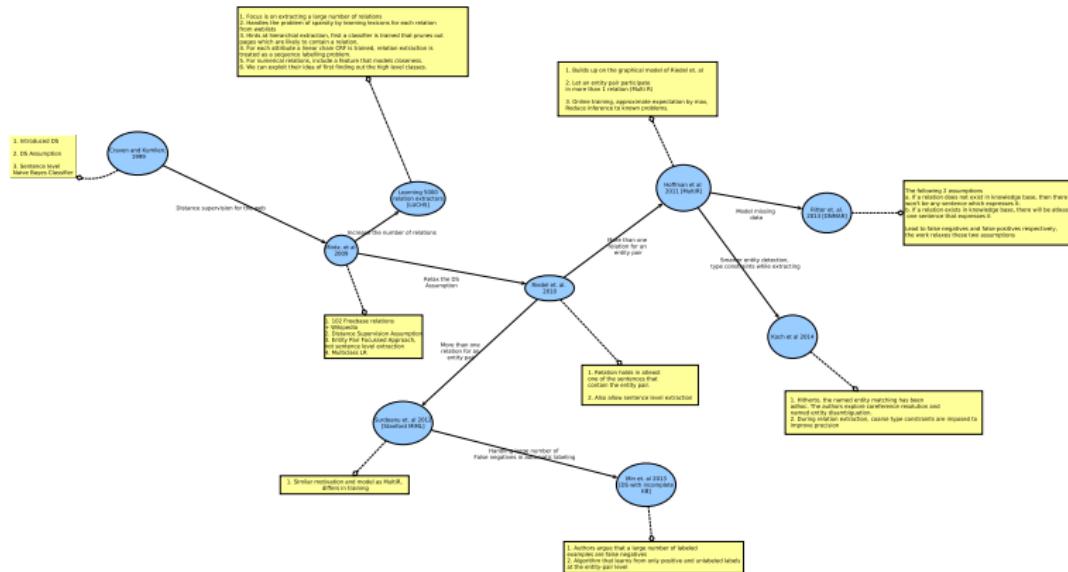
- *Distant supervision assumption:* If an entity pair  $E$  appears with a relation  $R$  in the knowledge base, then any sentence containing  $E$  **will** express  $R$
- Can quickly label huge corpora
- Same entity pair can match different relations (Founded(Steve Jobs, Apple) or CEO(Steve Jobs, Apple))
- False positives, may lead to the model learning wrong patterns for relations (E.g. “Film Festival” an Indicator of BornIn)

# Outline

- 1 Motivation
- 2 Problem Statement
- 3 Relation Extraction as a Machine Learning Problem
- 4 Distant Supervision
- 5 Distant Supervision Techniques
- 6 TAC Submission
- 7 Numerical Relation Extraction
- 8 Units in Numerical Relation Extraction

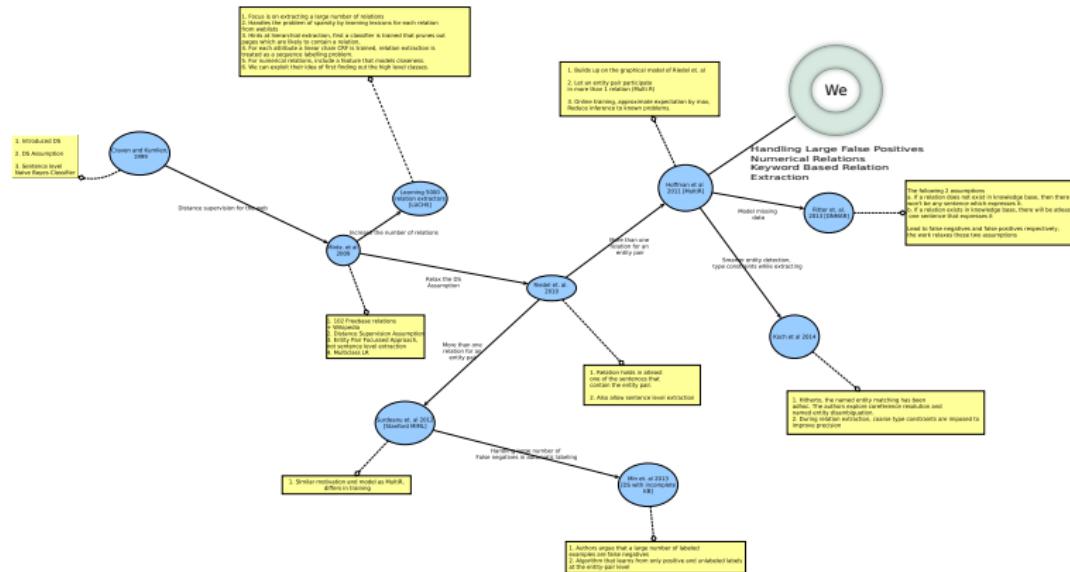
# Distant Supervision Techniques

- First paper in 1999, almost every possibility explored



# Distant Supervision Techniques

- First paper in 1999, *almost* every possibility explored

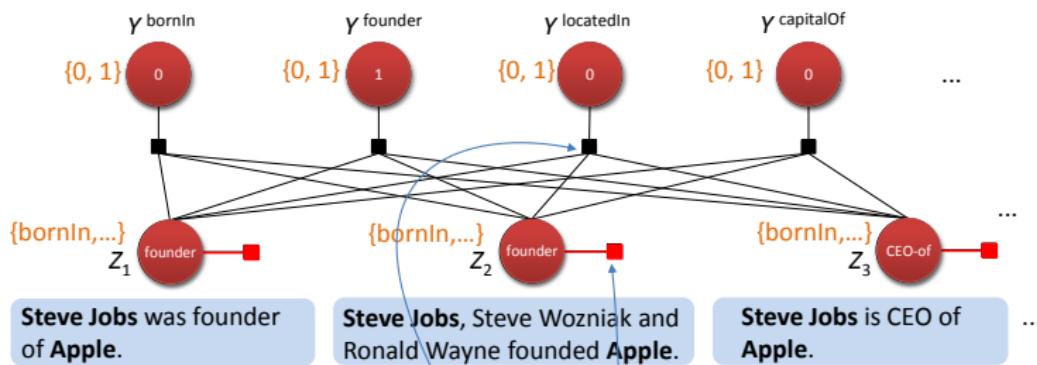


# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Model

Steve Jobs, Apple:



$$p(\mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z} | \mathbf{x}; \theta) \stackrel{\text{def}}{=} \frac{1}{Z_x} \prod_r \Phi^{\text{join}}(y^r, z) \prod_i \Phi^{\text{extract}}(z_i, x_i)$$

$$\Phi^{\text{join}}(y^r, z) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } y^r = \text{true} \wedge \exists i : z_i = r \\ 0 & \text{otherwise} \end{cases}$$

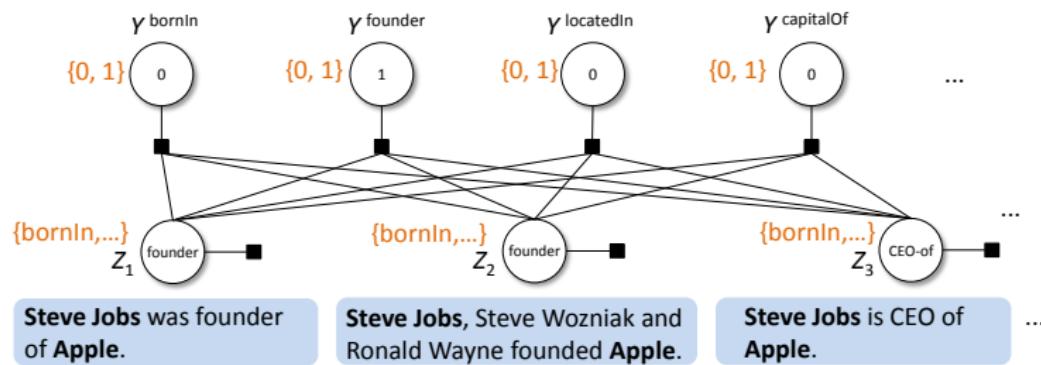
All features at  
sentence-level

(join factors are  
deterministic ORs)

# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Model



- Extraction almost entirely driven by sentence-level reasoning
- Tying of facts  $Y_r$  and sentence-level extractions  $Z_i$  still allows us to model weak supervision for training

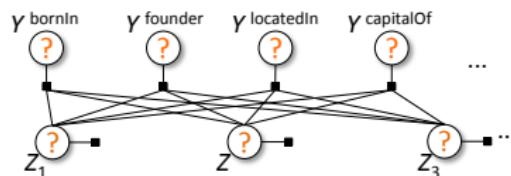
# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Inference

Need:

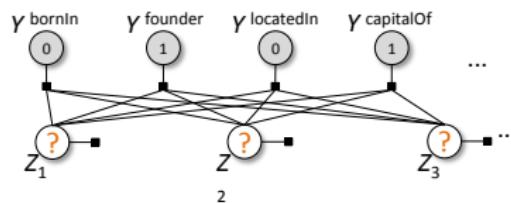
- Most likely sentence labels:



$$\arg \max_{\mathbf{y}, \mathbf{z}} p(\mathbf{y}, \mathbf{z} | \mathbf{x}; \theta)$$

Easy

- Most likely sentence labels *given facts*:



$$\arg \max_{\mathbf{z}} p(\mathbf{z} | \mathbf{x}, \mathbf{y}; \theta)$$

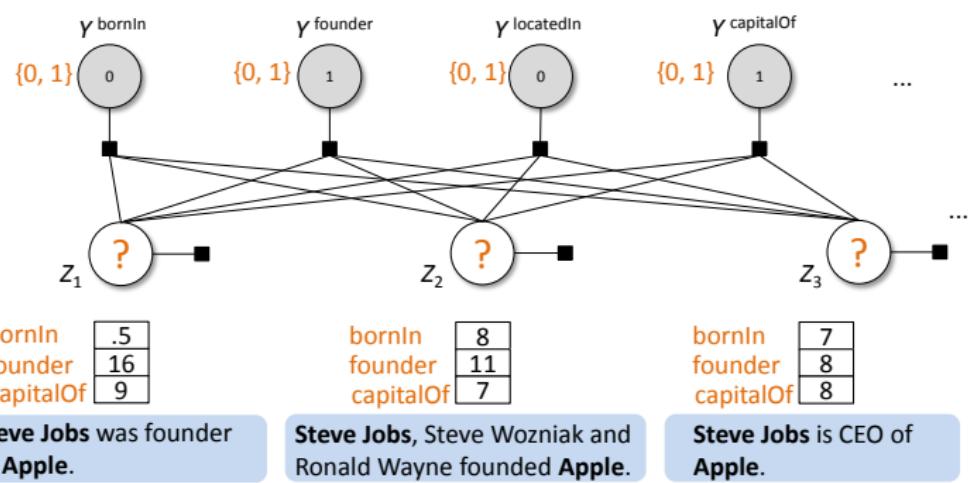
Challenging

# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Inference

- Computing  $\arg \max_z p(z|x, y; \theta)$ :

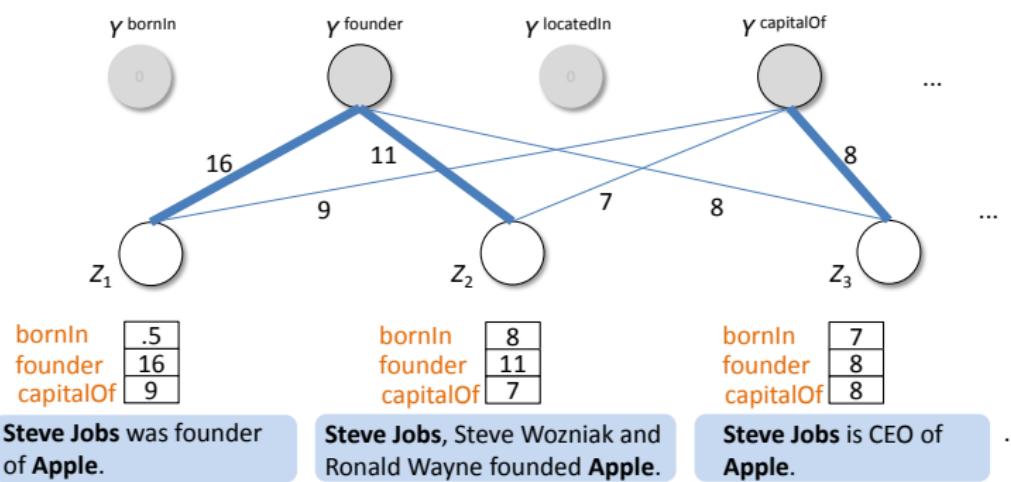


# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Inference

- Variant of the weighted, edge-cover problem:



# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Learning

- Training set  $\{(\mathbf{x}_i, \mathbf{y}_i) | i = 1 \dots n\}$ , where
  - $i$  corresponds to a particular entity pair
  - $\mathbf{x}_i$  contains all sentences with mentions of pair
  - $\mathbf{y}_i$  bit vector of facts about pair from database
- Maximize Likelihood

$$O(\theta) = \prod_i p(\mathbf{y}_i | \mathbf{x}_i; \theta) = \prod_i \sum_{\mathbf{z}} p(\mathbf{y}_i, \mathbf{z} | \mathbf{x}_i; \theta)$$

# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Learning

- Scalability: Perceptron-style additive updates
- Requires two approximations:
  1. Online learning  
For example i (entity pair), define

$$\phi(\mathbf{x}, \mathbf{z}) = \sum_j \phi(x_j, z_j)$$

Use gradient of local log likelihood for example i:

$$\begin{aligned}\frac{\partial \log O_i(\theta)}{\partial \theta_j} &= E_{p(\mathbf{z}|\mathbf{x}_i, \mathbf{y}_i; \theta)} [\phi_j(\mathbf{x}_i, \mathbf{z})] \\ &\quad - E_{p(\mathbf{y}, \mathbf{z}|\mathbf{x}_i; \theta)} [\phi_j(\mathbf{x}_i, \mathbf{z})]\end{aligned}$$

2. Replace expectations with maximizations

# Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

## Learning: Hidden-Variable Perceptron

passes over  
dataset

for each  
entity pair  $i$

most likely  
sentence labels  
and inferred facts  
(ignoring DB facts)

most likely  
sentence labels  
given DB facts

```
initialize parameter vector  $\Theta \leftarrow 0$ 
for  $t = 1 \dots T$  do
    for  $i = 1 \dots n$  do
         $(\mathbf{y}', \mathbf{z}') \leftarrow \arg \max_{\mathbf{y}, \mathbf{z}} p(\mathbf{y}, \mathbf{z} | \mathbf{x}_i; \theta)$ 
        if  $\mathbf{y}' \neq \mathbf{y}_i$  then
             $\mathbf{z}^* \leftarrow \arg \max_{\mathbf{z}} p(\mathbf{z} | \mathbf{x}_i, \mathbf{y}_i; \theta)$ 
             $\Theta \leftarrow \Theta + \phi(\mathbf{x}_i, \mathbf{z}^*) - \phi(\mathbf{x}_i, \mathbf{z}')$ 
        end if
    end for
end for
Return  $\Theta$ 
```

# Outline

- 1 Motivation
- 2 Problem Statement
- 3 Relation Extraction as a Machine Learning Problem
- 4 Distant Supervision
- 5 Distant Supervision Techniques
- 6 TAC Submission
- 7 Numerical Relation Extraction
- 8 Units in Numerical Relation Extraction

## Cold Start Knowledge Base Population, 2014

- Knowledge Base Population (KBP) track of TAC encourages the development of systems that can match entities mentioned in natural texts with those appearing in a knowledge base and extract novel information about entities from a document collection and add it to a new or existing knowledge base.

## Cold Start Knowledge Base Population, 2014

- Knowledge Base Population (KBP) track of TAC encourages the development of systems that can match entities mentioned in natural texts with those appearing in a knowledge base and extract novel information about entities from a document collection and add it to a new or existing knowledge base.
- Some example relations:
  - children of
  - city of birth
  - shareholders
  - countries of residence

## Cold Start Knowledge Base Population

- Modeled the problem using distant supervision.
- Used Freebase as an existing Knowledge base.
- **Freebase:** Freebase is a large collaborative knowledge base consisting of metadata composed mainly by its community members.
- It is an online collection of structured data harvested from many sources, including individual, user-submitted wiki contributions.

# Corpus

- The TAC corpus consisted of three type of documents:

# Corpus

- The TAC corpus consisted of three type of documents:
  - **discussion forums** 99,063 English discussion forum documents selected from the BOLT Phase 1 discussion forums source data releases. Each forum includes at least 5 posts.

- The TAC corpus consisted of three type of documents:
  - **discussion forums** 99,063 English discussion forum documents selected from the BOLT Phase 1 discussion forums source data releases. Each forum includes at least 5 posts.
  - **newswire** 1,000,257 documents selected from English Gigaword Fifth Edition.

- The TAC corpus consisted of three type of documents:
  - **discussion forums** 99,063 English discussion forum documents selected from the BOLT Phase 1 discussion forums source data releases. Each forum includes at least 5 posts.
  - **newswire** 1,000,257 documents selected from English Gigaword Fifth Edition.
  - **web** 999,999 English web documents selected from various GALE web collections.

- The TAC corpus consisted of three type of documents:
  - **discussion forums** 99,063 English discussion forum documents selected from the BOLT Phase 1 discussion forums source data releases. Each forum includes at least 5 posts.
  - **newswire** 1,000,257 documents selected from English Gigaword Fifth Edition.
  - **web** 999,999 English web documents selected from various GALE web collections.
- We have submitted the knowledge base populated using our techniques on the above corpus and waiting for results.

# Outline

- 1 Motivation
- 2 Problem Statement
- 3 Relation Extraction as a Machine Learning Problem
- 4 Distant Supervision
- 5 Distant Supervision Techniques
- 6 TAC Submission
- 7 Numerical Relation Extraction
- 8 Units in Numerical Relation Extraction

# Distant supervision for Numerical Relation Extraction

## Knowledge Base

- Derived from [data.worldbank.org](http://data.worldbank.org), 4371979 numerical facts about 249 countries, 1281 attributes

Freebase Entity	Value	Relation
/m/04g5k	3126000130	Electricity Production
/m/02k8k	1969.179	$CO_2$ Emission
/m/06nnj	332315	Total Population
/m/019rg5	55.020073	Life Expectancy
/m/05sb1	19974.148	$CO_2$ Emission
/m/05v8c	10000000000	Electricity Production
/m/03spz	7639000100	Electricity Production
/m/06vbd	44249.688	$CO_2$ Emission
/m/0d060g	51.3	Internet Users(%)
/m/05qkp	62.298927	Life Expectancy

# Selected Relations

Relation Name	Relation Code
Land area (sq. km)	AG.LND.TOTL.K2
Foreign direct investment, net (current US\$)	BN.KLT.DINV.CD
Goods exports (current US\$)	BX.GSR.MRCH.CD
Electricity production (kWh)	EG.ELC.PROD.KH
CO2 emissions (kt)	EN.ATM.CO2E.KT
Pump price for diesel fuel (US\$ per liter)	EP.PMP.DESL.CD
Inflation, consumer prices (annual %)	FP.CPI.TOTL.ZG
Internet users (per 100 people)	IT.NET.USER.P2
GDP (current US\$)	NY.GDP.MKTP.CD
Life expectancy at birth, total (years)	SP.DYN.LE00.IN
Population (Total)	SP.POP.TOTL

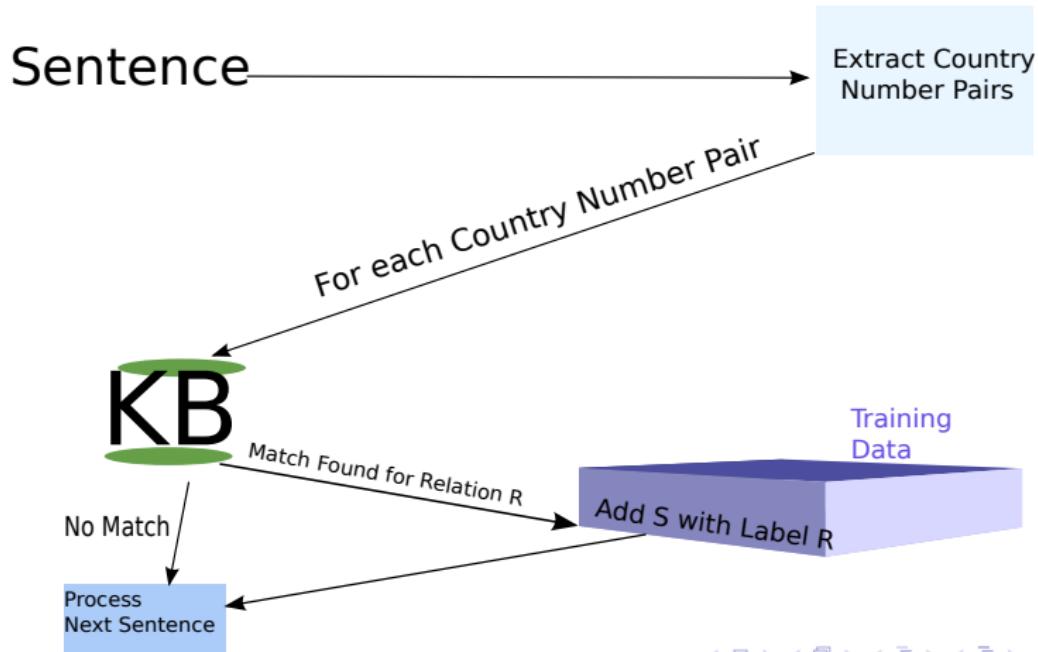
# Corpus

- Subset of the tac corpus
- 268, 036 Documents, X sentences having a country and a number
- List of countries augmented manually by adding all possible synonyms (Dutch, Netherlands) and inflections (Ireland, Irish)

# Distant supervision process for numerical relation extraction

- Extract a country and number from a sentence, go to kb and check if there is a match.
- Can be creative during matching:
  - Distance based matching
  - Time based matching

# First Attempt: Vanilla Matching



# Vanilla Matching

- A large number of matches
- Most of them false positives

# Vanilla Matching: Results

Relation	Total Matches	Sampled Matches	True Matches	Precision(%)
Land Area	1884	15	<b>1</b>	<b>6.7</b>
Foreign Direct Investment	0	0	0	0
Goods Export	0	0	0	0
Electricity Production	381	10	0	0
CO <sub>2</sub> Emission	0	0	0	0
Diesel Prices	8491	15	0	0
Inflation(%)	8689	15	0	0
Internet Users(%)	182319	40	0	0
GDP(\$)	0	0	0	0
Life Expectancy	267	10	0	0
Total Population	0	0	0	0

# Sample Matches for Vanilla Matching

Country	Relation	Value	Sentence
South Africa	Internet Users(%)	24	Montjane, <b>24</b> , was crowned on Sunday evening at the Superbowl at Sun City, in <b>South Africa</b> 's North West province.
Croatia	Inflation	500	<b>Croatian</b> police say more than <b>500</b> guns turned in after the country's 1991 war were stolen from a police depot and sold on the black market.
France	Life Expectancy	75	<b>France</b> : Hugo Lloris, Eric Abidal, Patrice Evra, William Gallas, Bacary Sagna, Abou Diaby, Yoann Gourcuff ( <b>Florent Malouda, 75</b> ), Jeremy Toulalan, Nicolas Anelka (Thierry Henry, 72), Sidney Govou (Andre-Pierre Gignac, 85), Franck Ribery.
France	Diesel prices	1.72	Lotte Friis of Denmark won the women's 800 freestyle in 8:23.27, followed at 0.73 seconds by Ophelie Cyriell Etienne of <b>France</b> and Federica Pellegrini of Italy, <b>1.72</b> seconds back.
Bermuda	Land Area(sq. km)	50	Forecasters were also closely tracking the path of Tropical Storm Fiona about 360 miles (580 km) south of <b>Bermuda</b> , with wind speeds of up to <b>50</b> miles (85 km) per hour.

## Key Observations on vanilla matching

- Units for precision, not just recall
- Dates and Sports articles lead to a lot of false positives
- Small numbers generate larger false positives; finer numbers generate fewer false positives.
  - Intuitively, it makes sense that we'll see a lot of 2s, 3s, 71s in different contexts than 1,000,232,112. Due to similar reasons, getting an exact match for 23.14152 is more difficult than matching 23.

# False Positives

Numbers are weak entities

- Vanilla numerical relation matching is bound to attract humoungous amounts of false positives;

# False Positives

Numbers are weak entities

- Vanilla numerical relation matching is bound to attract humoungous amounts of false positives;
- Stems from the fact that numbers don't have an identity of their own.

# False Positives

Numbers are weak entities

- Vanilla numerical relation matching is bound to attract humoungous amounts of false positives;
- Stems from the fact that numbers don't have an identity of their own.
- Consider India and Mumbai Vs. India and 19

# False Positives

Numbers are weak entities

- Vanilla numerical relation matching is bound to attract humoungous amounts of false positives;
- Stems from the fact that numbers don't have an identity of their own.
- Consider India and Mumbai Vs. India and 19
- Mumbai is a strong entity, 19 is a **weak** entity.

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can appear with India in several contexts:

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %
  - Billion dollars invested by a company

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %
  - Billion dollars invested by a company
  - % of people below the poverty line

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %
  - Billion dollars invested by a company
  - % of people below the poverty line
  - date (if we are not careful)

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %
  - Billion dollars invested by a company
  - % of people below the poverty line
  - date (if we are not careful)
  - number of medals won by Indian athletes...

# False Positives

Numbers are weak entities

- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %
  - Billion dollars invested by a company
  - % of people below the poverty line
  - date (if we are not careful)
  - number of medals won by Indian athletes...
- Can we expect the situation to be even worse for certain types of numbers?

# Outline

- 1 Motivation
- 2 Problem Statement
- 3 Relation Extraction as a Machine Learning Problem
- 4 Distant Supervision
- 5 Distant Supervision Techniques
- 6 TAC Submission
- 7 Numerical Relation Extraction
- 8 Units in Numerical Relation Extraction

# Numbers are incomplete without units

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth [US\\$1.842 trillion](#); it is the eleventh-largest economy by market exchange rates, and is, at [US\\$4.962 trillion](#), the third-largest by [purchasing power parity](#), or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reaching

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth [US\\$1.842 trillion](#); it is the eleventh-largest economy by market exchange rates, and is, at [US\\$4.962 trillion](#), the third-largest by [purchasing power parity](#), or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reac-

- Units help in improving recall.

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth [US\\$1.842 trillion](#); it is the eleventh-largest economy by market exchange rates, and is, at [US\\$4.962 trillion](#), the third-largest by [purchasing power parity](#), or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reaching

- Units help in improving recall.

- In above example 1.842 as number alone would not match with the fact regarding economy in knowledge base.

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth [US\\$1.842 trillion](#); it is the eleventh-largest economy by market exchange rates, and is, at [US\\$4.962 trillion](#), the third-largest by [purchasing power parity](#), or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reaching

- Units help in improving recall.

- In above example 1.842 as number alone would not match with the fact regarding economy in knowledge base.
- But with the unit trillion USD, we can normalize the value and then it would match existing facts in knowledge base.

# Numbers are incomplete without units

## Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth US\$1.842 trillion; it is the eleventh-largest economy by market exchange rates, and is, at US\$4.962 trillion, the third-largest by purchasing power parity, or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reac-

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth US\$1.842 trillion; it is the eleventh-largest economy by market exchange rates, and is, at US\$4.962 trillion, the third-largest by purchasing power parity, or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reac-

- Units help in improving recall.

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth US\$1.842 trillion; it is the eleventh-largest economy by market exchange rates, and is, at US\$4.962 trillion, the third-largest by purchasing power parity, or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reac-

- Units help in improving recall.
  - In above example 1.842 as number alone would not match with the fact regarding economy in knowledge base.

# Numbers are incomplete without units

- Apart from being a constant quantity, a number usually makes sense when presented along with units.

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth US\$1.842 trillion; it is the eleventh-largest economy by market exchange rates, and is, at US\$4.962 trillion, the third-largest by purchasing power parity, or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reac-

- Units help in improving recall.

- In above example 1.842 as number alone would not match with the fact regarding economy in knowledge base.
- But with the unit trillion USD, we can normalize the value and then it would match existing facts in knowledge base.

# Numbers are incomplete without units

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth US\$1.842 trillion; it is the eleventh-largest economy by market exchange rates, and is, at US\$4.962 trillion, the third-largest by purchasing power parity, or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reaching

- Units help in reducing false positives and hence improving precision.

# Numbers are incomplete without units

According to the International Monetary Fund (IMF), as of 2013, the Indian economy is nominally worth US\$1.842 trillion; it is the eleventh-largest economy by market exchange rates, and is, at US\$4.962 trillion, the third-largest by purchasing power parity, or PPP.<sup>[9]</sup> With its average annual GDP growth rate of 5.8% over the past two decades, and reaching

- Units help in reducing false positives and hence improving precision.
- If there is a fact, e.g, **inflation(India, 1.842%)** in knowledge base, then ignoring units can cause an incorrect match which leads in learning towards noisy patterns.

# Unit Extraction is not easy!

# Unit Extraction is not easy!

- Different ways to represent a single unit.

# Unit Extraction is not easy!

- Different ways to represent a single unit.
  - Tunisia occupies an area of 163,610 **square kilometres**, of which 8,250 are water.

# Unit Extraction is not easy!

- Different ways to represent a single unit.
  - Tunisia occupies an area of 163,610 **square kilometres**, of which 8,250 are water.
  - With an area of about 9.6 **million km<sup>2</sup>**, the People's Republic of China is the 3rd largest country in total area behind Russia and Canada, and very similar to the United States.

# Unit Extraction is not easy!

- Different ways to represent a single unit.
  - Tunisia occupies an area of 163,610 **square kilometres**, of which 8,250 are water.
  - With an area of about 9.6 **million km<sup>2</sup>**, the People's Republic of China is the 3rd largest country in total area behind Russia and Canada, and very similar to the United States.
- Multiple units to represent a single numerical fact.

# Unit Extraction is not easy!

- Different ways to represent a single unit.
  - Tunisia occupies an area of 163,610 **square kilometres**, of which 8,250 are water.
  - With an area of about 9.6 **million km<sup>2</sup>**, the People's Republic of China is the 3rd largest country in total area behind Russia and Canada, and very similar to the United States.
- Multiple units to represent a single numerical fact.
  - Vatican City, a walled enclave within the city of Rome, with an area of approximately **44 hectares (110 acres)**, and a population of 842, is the smallest internationally recognized independent state in the world by both area and population.

# Overview of Unit Extraction System

# Overview of Unit Extraction System

- A discriminative context free grammar with scores attached to each possible production in the grammar.

# Overview of Unit Extraction System

- A discriminative context free grammar with scores attached to each possible production in the grammar.
- A production  $P$  in the grammar is of the form  $R ::= R_1 R_2$ , scored as

$$score(P) = \mathbf{w} \cdot \mathbf{f}(P, x, i, j, k),$$

where  $(i, j)$  and  $(j + 1, k)$  are text spans in  $x$  that  $R_1$  and  $R_2$  cover.

# Overview of Unit Extraction System

- Some of the features that grammar uses to assign the best scores to various parses are as belows:

# Overview of Unit Extraction System

- Some of the features that grammar uses to assign the best scores to various parses are as belows:
  - Matches with Unit Catalog

# Overview of Unit Extraction System

- Some of the features that grammar uses to assign the best scores to various parses are as belows:
  - Matches with Unit Catalog
  - Lexical Clues

# Overview of Unit Extraction System

- Some of the features that grammar uses to assign the best scores to various parses are as belows:
  - Matches with Unit Catalog
  - Lexical Clues
  - Relative Frequency - Prior of the word to be present as unit, then as an non-unit word. This is derived from WordNet ontologies.

# Overview of Unit Extraction System

- Some of the features that grammar uses to assign the best scores to various parses are as belows:
  - Matches with Unit Catalog
  - Lexical Clues
  - Relative Frequency - Prior of the word to be present as unit, then as an non-unit word. This is derived from WordNet ontologies.
  - Co-occurrence statistics - presence of strongly co-occurring words in the text can help in disambiguating the various candidate units

# Units Based Matching: Results

Relation	Total Matches	Sampled Matches	True Matches	Precision(%)
<b>Land Area</b>	98	40	32	80
Foreign Direct Investment	791	40	1	2.5
Goods Export	816	40	3	7.5
Electricity Production	19	19	0	0
CO <sub>2</sub> Emission	196	40	2	5
Diesel Prices	2	2	2	100
Inflation(%)	27598	40	0	0
Internet Users(%)	24639	40	0	0
GDP(\$)	1790	40	0	0
Life Expectancy	3081	40	0	0
<b>Total Population</b>	5225	40	11	27.5

# Sample Matches for Unit Based Matching

Country	Relation	Value	Sentence
Lebanon	Land Area	$1.02 \times 10^{11}$ sq m	Lebanon lies in the eastern Mediterranean and covers about <b>4,030 square miles</b> (10,450 square kilometers) – smaller than the U.S. state of Connecticut.
Russia	Goods Export	$10^{10}$ USD	The IMF agrees to offer a loan of <b>US\$10.2 billion</b> to Russia over the next three years to help Russians transform their economy.
Australia	Internet Users(%)	0.6%	South Korea's main index added <b>0.6 percent</b> , China's Shanghai's benchmark climbed 1.5 percent and <b>Australia's</b> index advanced 1.1 percent.
Israel	Internet Users(%)	20.0%	<b>Israel's</b> Arab community numbers 1.3 million, about <b>20 percent</b> of the population.
Pakistan	Life Expectancy(years)	60	Why does a small elite still control vast swaths of land more than <b>60 years</b> after <b>Pakistan</b> became a nation?

# Key Observations on Unit Based Matching

- When it comes to numbers, distant supervision assumption is weak as can be evident from the examples above. Since second argument of a relation is number, this number can be related to the entity in multiple number of ways, which then gives lot of false positives.
- For the attributes like inflation, percentage of internet user, whose values are in percentage are affected by stock data that is heavily present in news corpus.
- For relations FDI, Goods Export, GDP we have almost similar values for three values. Hence a (entity, number) pair matches all of the three relations. We can improve our matching function for closely related attributes.

# A case for Keywords

- A large number of false matches had no reference to the relation involved.
- Eg. No mention of Population in the following matches:
  - The website of China's Ministry of Defense (MOD) has attracted around 1.25 billion visits in the three months since its opening, with the United States topping the source countries for foreign visits, website editor-in-chief Ji Guilin said.
  - Insulza, for his part, said the Organization of American States expects to raise 10 million dollars for Haiti's recovery.
  - Koloini and others brought 10 million euros, probably 15 million, back from Iraq at the time, Falter quoted from the diary.
- No reference to Co2 emission:
  - China's iron ore imports surged 41.6 percent to 627.8 million tonnes in 2009, with the value falling 17.4 percent as prices were hit by the global downturn, customs data shows.

# A case for Keywords

## Good News

Sentences expressing a numerical relation can be expected to have keywords that denote the relation

- Take all the labeled sentences, prune out sentences that don't have atleast one of the relevant keywords

Internet User %	"Internet"
Land Area	"area", "land", "land area"
Population	"Population"
Diesel	"diesel"
GDP	"Gross domestic", "GDP"
CO2	"Carbon", "Carbon Emission", "CO2"
Inflation	"Inflation", "Price Rise"
FDI	"Foreign", "FDI"
Goods Export	"goods"
Life Expectancy	"life", "life expectancy"
Electricity Production	"Electricity"

# A case for Keywords

- Numbers are the second entity in our setup ( $\text{Relation}(\text{Country}, \text{Number})$ )
- Unlike real world entities, numbers don't have an identity of their own, sentences should have words (keywords!) indicating what the number stands for
- Manual inspection of 400 sentences pruned out after applying keyword based filter backs this conjecture, not even one false negative
- The keywords are created manually, can this process be automated?

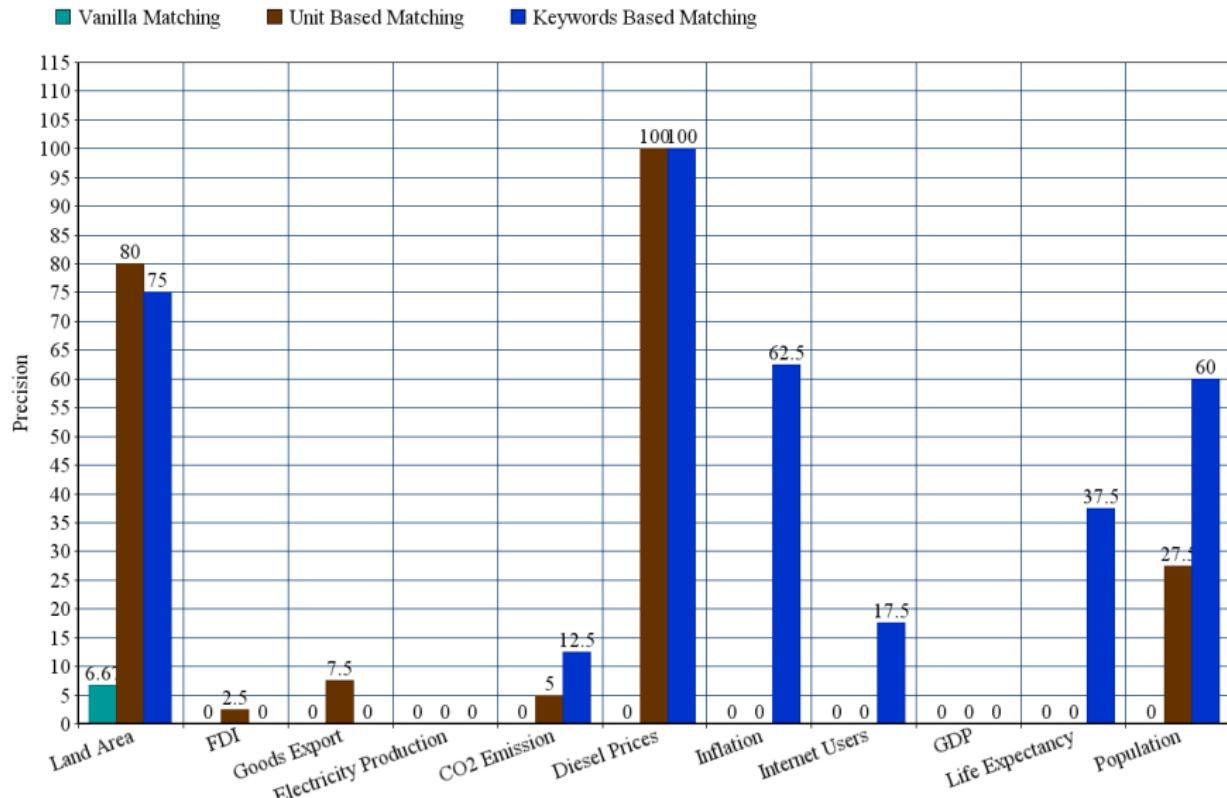
# Keywords + Units Based Matching: Results

Relation	Total Matches	Sampled Matches	True Matches	Precision(%)
<b>Land Area</b>	61	40	<b>30</b>	<b>75</b>
Foreign Direct Investment	8	0	0	0
Goods Export	4	4	0	0
Electricity Production	0	0	0	0
CO <sub>2</sub> Emission	16	16	2	12.5
Diesel Prices	2	2	2	100
<b>Inflation(%)</b>	3853	40	25	62.5
Internet Users(%)	308	40	7	17.5
GDP(\$)	0	0	0	0
<b>Life Expectancy</b>	99	40	15	37.5
<b>Total Population</b>	607	40	24	60

# Sample Matches for Keywords Based Matching

Country	Relation	Value	Sentence
India	Land Area	$3.00 \times 10^{12}$ sq m	With an <b>area</b> of <b>2.98 million square km</b> , <b>India</b> is the largest country in South Asia.
Bulgaria	$CO_2$ Emission	$6 \times 10^7$ ton	Depending on the Commission's ruling on the country's challenging of the quotas, <b>Bulgaria</b> would end up with at least <b>60 million tonnes</b> of <b>CO2</b> or 60 million EUAs to trade, which are worth several hundred million euros.
Malawi	Life Expectancy(years)	40 years	<b>Malawi</b> , like other southern African countries, has seen its <b>life expectancy</b> drop from about 60 years in the early 1990s to below <b>40 years</b> presently due to the HIV-AIDS pandemic.
Iceland	Total Population	330000	home to <b>320,000 people</b> , <b>Iceland</b> officially apply to join the EU at the end of July.
Japan	Inflation(%)	8.2%	<b>Japan</b> 's economy decline by <b>8.4 percent</b> , after adjustment for <b>inflation</b> , from the first quarter of 2008 to the first quarter of 2009.

# Precision Comparison



## Relative Recall

- Different Heuristics improved some aspect of matching for distant supervision.
- But we need to be careful, that with aggressive pruning of incorrect sentences, we are not leaving some good sentences.
- So for all the Heuristics, we calculate the percentage of true matches that were extracted by each Heuristics. We call this **Relative Recall**

Total Sentences for evaluation	676
Total True Positives	138
Vanilla Matching	2
Units + Distance Based Matching	79
Keywords + Units + Distance Based Matching	112

# Dependency Path Based Relation Extraction

## Peculiarity of Numerical Relations

- Analyzing a number of sentences expressing numerical relations lead to several insights as already discussed.
- **Keywords** We can expect presence of certain keywords that might help in identifying relations.
- **Modifiers** A large number of false positives stem out of mentions where a change in the numerical attribute is mentioned.

# Dependency Path Based Relation Extraction

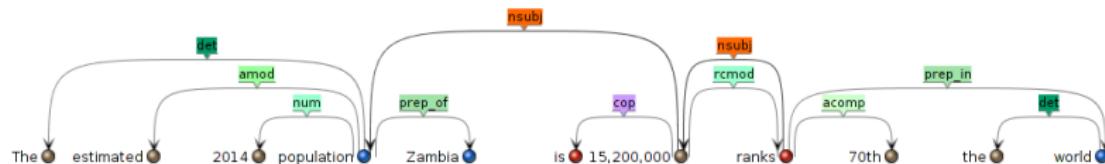
## Dependencies

- Dependencies: Grammatical relation between two words, governor and dependent.
- “The red ball was lost”
  - **amod(ball,3,red,2)** “Red” is an adjective for “ball”
  - **det(ball,3,The,1)** “the” is a determiner of “ball”
  - **nsubjpass(lost,5,ball,3)** “ball is the subject of lost”
  - **auxpass(lost,5,was,4)** “was is an auxiliary of lost”



# Dependency Path Based Relation Extraction

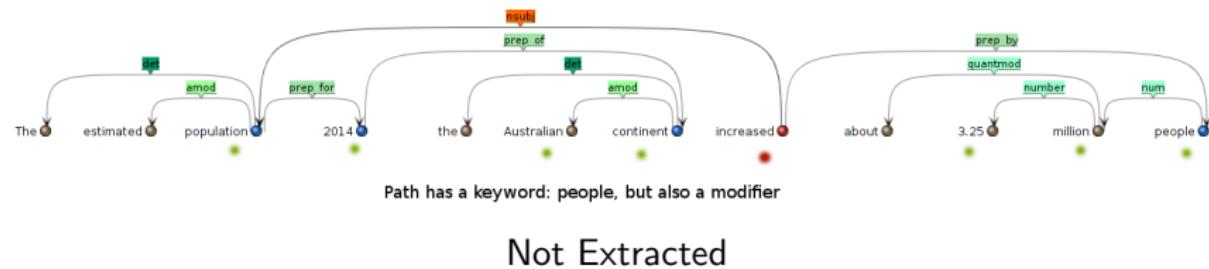
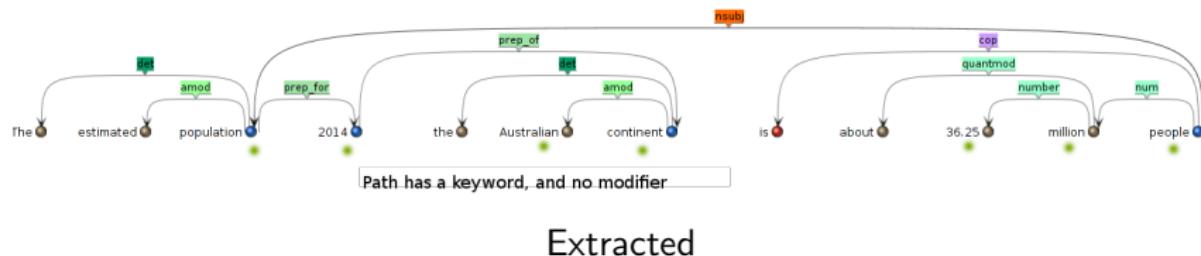
- Given a Country-Number pair, extract the shortest undirected path between them in the dependency graph.



- Path(Zambia - 15,200,000) = {Zambia, population, 15,200,000}
- For a match, the path:
  - Should have one of the keywords
  - Should not have a modifier

# Dependency Path Based Relation Extraction

## Example



- The extractor was applied to 30 sentences expressing 23 different relations.



	Relations Present	Relations not Present (False positives)
Extracted	16	17
Not Extracted	7	N/A

- Precision: 48.4%
- Recall: 69.6%
- The precision will increase further on applying unit based pruning.