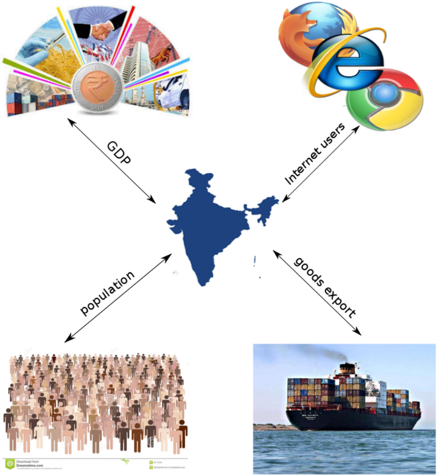
Outline

Motivation



Motivation

- Repositories of facts containing this information can be found at many places, like data.worldbank.org, Wikipedia infoboxes etc.

Motivation

- Repositories of facts containing this information can be found at many places, like data.worldbank.org, Wikipedia infoboxes etc.
- Countries are popular and finite, finding complete knowledge bases is possible.

Motivation

- Repositories of facts containing this information can be found at many places, like data.worldbank.org, Wikipedia infoboxes etc.
- Countries are popular and finite, finding complete knowledge bases is possible.
- What about less popular entities?

Motivation

- Repositories of facts containing this information can be found at many places, like data.worldbank.org, Wikipedia infoboxes etc.
- Countries are popular and finite, finding complete knowledge bases is possible.
- What about less popular entities?
  - What is the population of Arbit Apartments, Powai?

Motivation

- Repositories of facts containing this information can be found at many places, like data.worldbank.org, Wikipedia infoboxes etc.
- Countries are popular and finite, finding complete knowledge bases is possible.
- What about less popular entities?
  - What is the population of Arbit Apartments, Powai?
  - What is the GDP of Sugarcane Industry of India?

Motivation

- Repositories of facts containing this information can be found at many places, like data.worldbank.org, Wikipedia infoboxes etc.
- Countries are popular and finite, finding complete knowledge bases is possible.
- 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

- Good News!!!

Motivation

- Good News!!!
- Web is huge, probably, there is some page which contains the information we are looking for.

Motivation

- Good News!!!
- 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

- Good News!!!
- 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 to be similar.

Motivation

- Good News!!!
- 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 to be similar.
  - Population of India reached 1.3 billion, making it the second largest country in the world.

Motivation

- Good News!!!
- 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 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

Problem Statement

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

Introduction

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

> 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.[9] With its average annual GDP growth rate of 5.8% over the past two decades, and reac

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

- Can such facts be extracted automatically?

Relation Extraction: Problem

- Extract 3-tuples which consists of an entity and a numerical value that are bound by some relation.
  - (India, **economy**, 1.842 trillion USD)
  - (China, **internet users**, 590.56 million)
  - (USA, **land area**, 2,959,054 square mile)

Outline

Relation extraction as a Machine Learning Problem

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

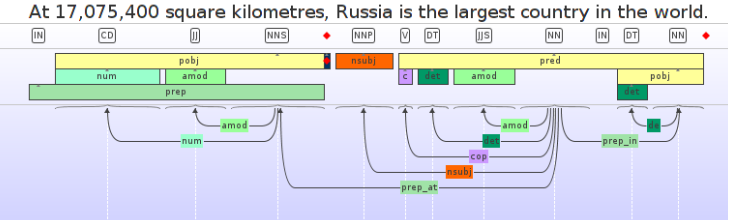
- Structure and content of sentences expressing the same relations can be expected to be similar.
  - At 17,075,400 square kilometres, Russia is the largest country in the world.
  - With an area of 504,030 km^2, Spain is the second largest country in Western Europe.

Relation Extraction as a Machine Learning Problem

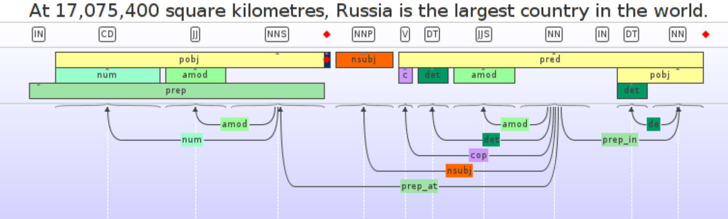
- Redundancy in grammatical features and dependencies of the sentences expressing same relation.

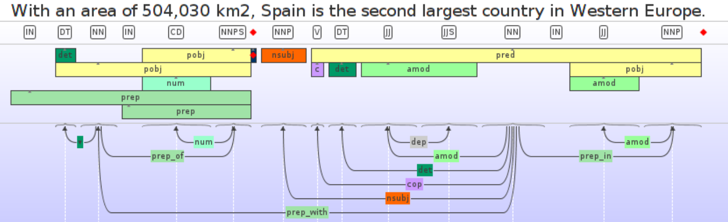
Relation Extraction as a Machine Learning Problem

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



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





Relation Extraction as a Machine Learning Problem

- There is lot of redundancy in ways in which a relation is expressed in sentence.

Relation Extraction as a Machine Learning Problem

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.

Relation Extraction as a Machine Learning Problem

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.
  - grammatical patterns - POS tags, dependency parse.

Relation Extraction as a Machine Learning Problem

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.
    - grammatical patterns - POS tags, dependency parse.
    - keywords for the relations.

Relation Extraction as a Machine Learning Problem

- There is lot of redundancy in ways in which a relation is expressed in sentence.
- So for every relation learn the patterns that express it.
  - grammatical patterns - POS tags, dependency parse.
  - keywords for the relations.
- This forms the relation extraction as a multi-class classification problem.

Possible Workflow

Possible Workflow

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.

Possible Workflow

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.
- Extract features (important keywords, grammatical structure, parse tress, etc.) for these sentences.

Possible Workflow

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.
- Extract features (important keywords, grammatical structure, parse tress, etc.) for these sentences.
- Learn a multi-class classifier on this training data (Explained later).

Possible Workflow

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.
- Extract features (important keywords, grammatical structure, parse tress, etc.) for these sentences.
- Learn a multi-class classifier on this training data (Explained later).
- Once the model is learnt, for every sentence

Possible Workflow

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.
- Extract features (important keywords, grammatical structure, parse tress, etc.) for these sentences.
- Learn a multi-class classifier on this training data (Explained later).
- Once the model is learnt, for every sentence
    - Extract features for the sentence

Possible Workflow

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.
- Extract features (important keywords, grammatical structure, parse tress, etc.) for these sentences.
- Learn a multi-class classifier on this training data (Explained later).
- Once the model is learnt, for every sentence
  - Extract features for the sentence
  - Predict the relation using the model for these features

Possible Workflow

- Collect enough examples for each relation so that there are sufficient patterns and enough redundancy to exploit.
- Extract features (important keywords, grammatical structure, parse tress, etc.) for these sentences.
- Learn a multi-class classifier on this training data (Explained later).
- Once the model is learnt, for every sentence
  - Extract features for the sentence
  - Predict the relation using the model for these features
  - store the fact into database.

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).
- It is very hard to go through the entire corpus and label each sentence to one of the relations.

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).
- It is very hard to go through the entire corpus and label each sentence to one of the relations.
- For model to generalize well, we need lot of training data.

Challenge

- The size of corpus is enormous (e.g, 5 million sentences).
- It is very hard to go through the entire corpus and label each sentence to one of the relations.
- For model to generalize well, we need lot of training data.
- What to do then?

Outline

Distant Supervision

Distant Supervision

- Manual labeling of the entire corpus is not possible

Distant Supervision

- Manual labeling of the entire corpus is not possible
- Weak Supervision as a middle ground

Distant Supervision

- Manual labeling of the entire corpus is not possible
- Weak Supervision as a middle ground
- Use Heuristics to align a table of facts with the corpus

Distant Supervision

- Manual labeling of the entire corpus is not possible
- Weak Supervision as a middle ground
- Use Heuristics to align a table of facts with the corpus
- Fuzzy training

- 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 was born in Paddington, London, while his father was on leave from his position 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 was born in Paddington, London, while his father was on leave from his position 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 was born in Paddington, London, while his father was on leave from his position 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 was born in Paddington, London, while his father was on leave from his position 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 was born in Paddington, London, while his father was on leave from his position with the Indian Civil Service (ICS) at Chhatrapur, Bihar
  - Alan Turing biopic The Imitation Game named as London film festival opener.

Distant Supervision

Distant Supervision

- Distant supervision assumption, any sentence containing the entity pair will express the corresponding relation

Distant Supervision

- Distant supervision assumption, any sentence containing the entity pair will express the corresponding relation
- Can quickly label huge corpora

Distant Supervision

- Distant supervision assumption, any sentence containing the entity pair will express the corresponding relation
- Can quickly label huge corpora
- Same entity pair can match different relations (Founded(Steve Jobs, Apple) or CEO(Steve Jobs, Apple))

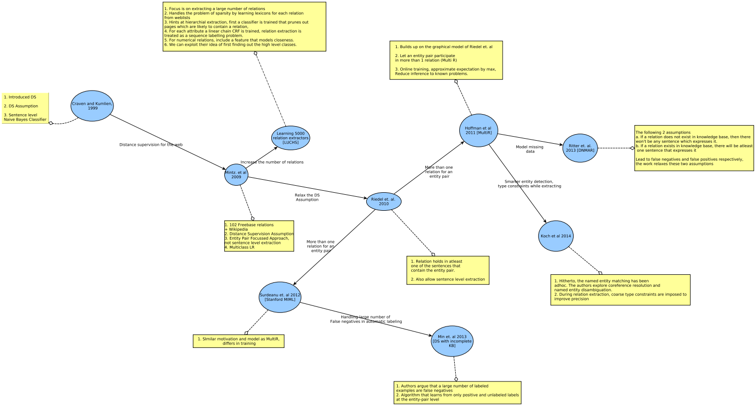
Distant Supervision

- Distant supervision assumption, any sentence containing the entity pair will express the corresponding relation
- 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 model learning wrong patterns for relations

Outline

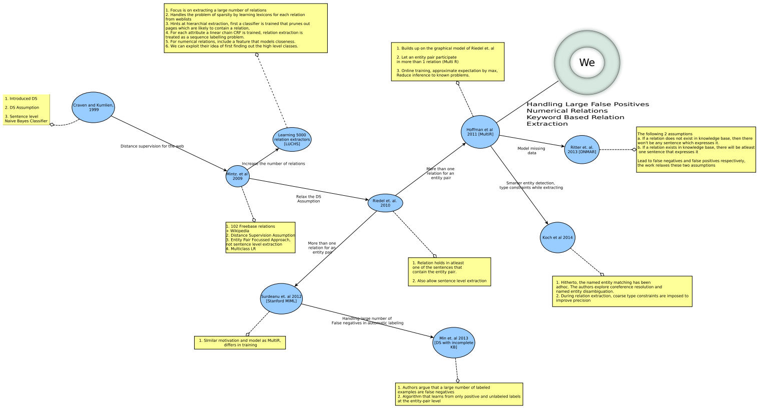
Distant Supervision Techniques

- First paper in 1999, almost every possibility explored



Distant Supervision Techniques

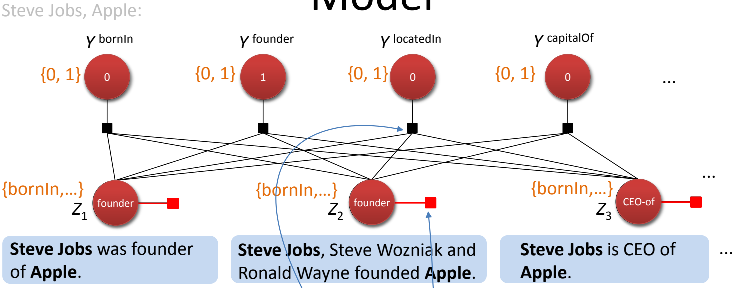
- First paper in 1999, *almost* every possibility explored



Relation Extraction Using MultiR

From `raphaelhoffmann.com/publications`

Model



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

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

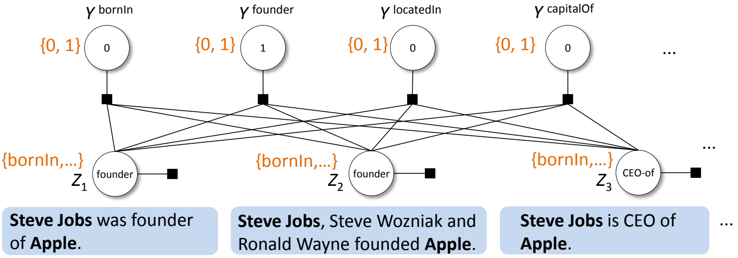
All features at sentence-level

(join factors are deterministic ORs)

Relation Extraction Using MultiR

From raphaelhoffmann.com/publications

Model

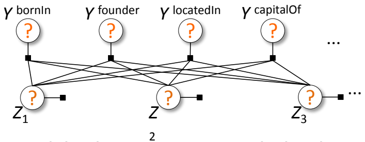


- Extraction almost entirely driven by sentence-level reasoning
- Tying of facts Y_r and sentence-level extractions Z_i still allows us to model weak supervision for training

Inference

Need:

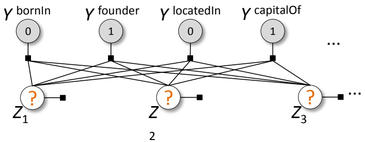
- Most likely sentence labels:



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

Easy

- Most likely sentence labels *given* facts:

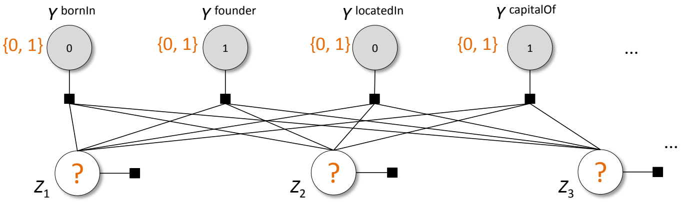


$\arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}, \mathbf{y}; \theta)$

Challenging

Inference

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



bornIn	.5
founder	16
capitalOf	9

Steve Jobs was founder of **Apple**.

bornIn	8
founder	11
capitalOf	7

Steve Jobs, Steve Wozniak and Ronald Wayne founded **Apple**.

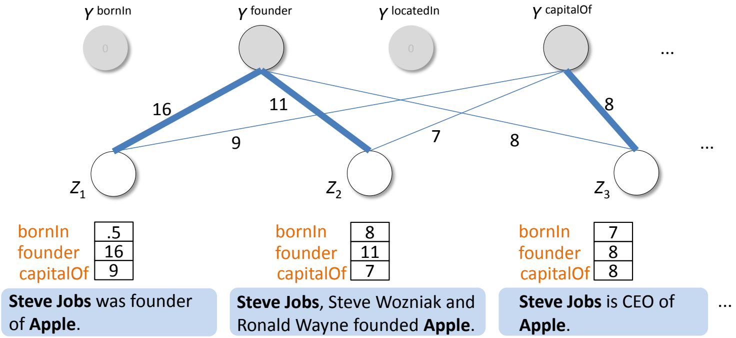
bornIn	7
founder	8
capitalOf	8

Steve Jobs is CEO of **Apple**.

...

Inference

- Variant of the weighted, edge-cover problem:



Learning

- Training set $\{(\mathbf{x}_i, \mathbf{y}_i) | i = 1 \ldots n\}$, 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

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

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:
     $$\frac{\partial \log O_i(\theta)}{\partial \theta_j} = E_{p(\mathbf{z}|\mathbf{x}_i, \mathbf{y}_i; \theta)}[\phi_j(\mathbf{x}_i, \mathbf{z})] - E_{p(\mathbf{y}, \mathbf{z}|\mathbf{x}_i; \theta)}[\phi_j(\mathbf{x}_i, \mathbf{z})]$$
  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 $\mathbf{\Theta} \leftarrow \mathbf{0}$

for $t = 1 \ldots T$ **do**

  for $i = 1 \ldots n$ **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**

      $\mathbf{z}^* \leftarrow \arg\max_{\mathbf{z}} p(\mathbf{z} | \mathbf{x_i}, \mathbf{y_i}; \theta)$

      $\mathbf{\Theta} \leftarrow \mathbf{\Theta} + \phi(\mathbf{x_i}, \mathbf{z}^*) - \phi(\mathbf{x_i}, \mathbf{z}')$

    end if

  end for

end for

Return Θ

Outline

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.

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.
  - **newswire** 1,000,257 documents selected from English Gigaword Fifth Edition.

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.
  - **newswire** 1,000,257 documents selected from English Gigaword Fifth Edition.
  - **web** 999,999 English web documents selected from various GALE web collections.

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.
  - **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

- Derived from `data.worldbank.org`, 4371979 numerical facts about 249 countries, 1281 attributes

/m/04g5k	3126000130	EG.ELC.PROD.KH
/m/02k8k	1969.179	EN.ATM.CO2E.KT
/m/06nnj	332315	SP.POP.TOTL
/m/019rg5	55.020073	SP.DYN.LE00.IN
/m/05sb1	19974.148	EN.ATM.CO2E.KT
/m/05v8c	10000000000	EG.ELC.PROD.KH
/m/03spz	7639000100	EG.ELC.PROD.KH
/m/06vbd	44249.688	EN.ATM.CO2E.KT
/m/0d060g	51.3	IT.NET.USER.P2
/m/05qkp	62.298927	SP.DYN.LE00.IN

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

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 be 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

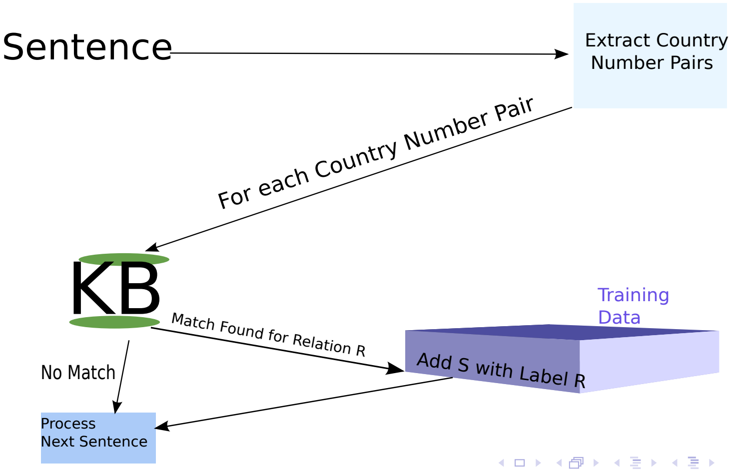
- Compare the number of sentences in which India and Mumbai appear together, vs the number of sentences in which India and 19 appear together.
- Just 19 can be appear with India in several contexts:
  - Internet user %
  - Billion dollars invested by a company
  - % of people below the poverty line
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

One Thousand Words



Outline

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.[9] 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.[9] 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.[9] 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.[9] 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 exisiting 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.[9] 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.[9] 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.[9] 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.[9] 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 exisiting 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.[9] 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.[9] 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^2, 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^2, 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^2, 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 grammmar with scores attached to each possible production in the grammar.

Overview of Unit Extraction System

- A discriminative context free grammmar 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

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

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-occuring words in the text can help in disambiguating the various candidate units

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^2, 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^2, 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^2, 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.

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 (Relation(Country, 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?

Outline