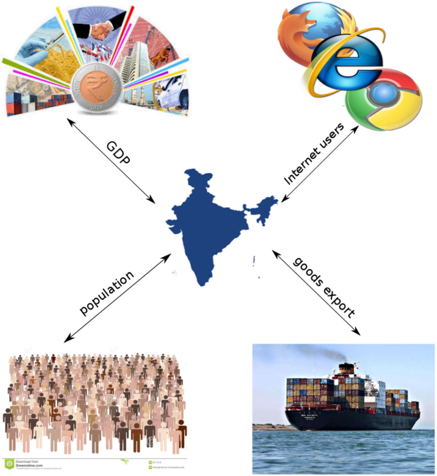
Numerical Relation Extraction with Minimal Supervision

Indian Institute of Technology Bombay

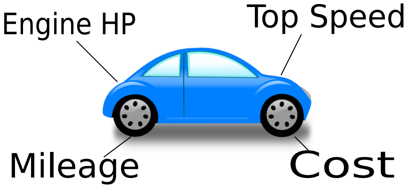
June 2015

Outline

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?

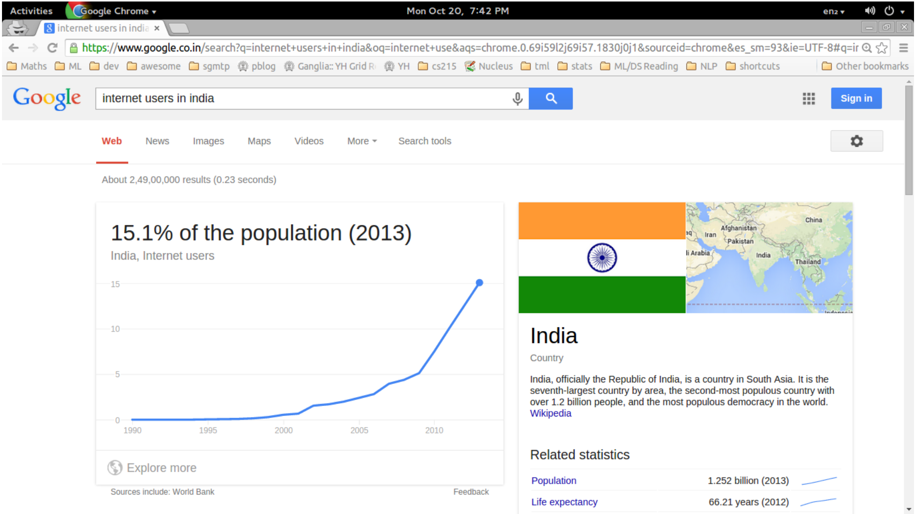
Entities and Numerical Attributes

- 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

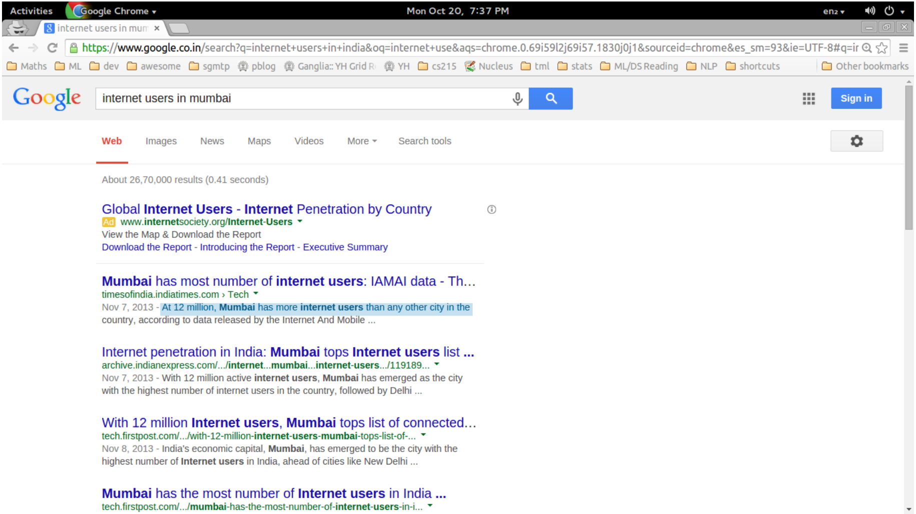
- For popular entities, finding complete knowledge bases is possible.
- data.worldbank.org, Wikipedia infoboxes, freebase ...
- What about less popular entities?
  - What is the population of Arbit Apartments, Powai?
  - What is the GDP of Sugarcane Industry of India?
  - Percent of Internet users in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

Motivation



Activities
Google Chrome
Mon Oct 20, 7:42 PM
internet users in india
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
Maths ML dev awesome sgmtp pblog Ganglia:: YH Grid R YH cs215 Nucleus tml stats ML/DS Reading NLP shortcuts Other bookmarks
Google
internet users in india
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
15
10
5
0
1990 1995 2000 2005 2010
Explore more
Sources include: World Bank
Feedback
China
Afghanistan
Iran
Pakistan
India
Thailand
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
Population 1.252 billion (2013)
Life expectancy 66.21 years (2012)

Motivation



Activities
Google Chrome
Mon Oct 20, 7:37 PM
internet users in mum
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
Maths ML dev awesome sgmtp pblog Ganglia:: YH Grid R YH cs215 Nucleus tml stats ML/DS Reading NLP shortcuts Other bookmarks
Google
internet users in mumbai
Sign in
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
View the Map & Download the Report
Download the Report - Introducing the Report - Executive Summary
Mumbai has most number of internet users: IAMAI 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.

[1] More on this in the coming slides

Motivation

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

[1] More on this in the coming slides

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.

[1] More on this in the coming slides

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** [1] to be similar.

[1] More on this in the coming slides

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** [1] to be similar.
  - Population of India reached 1.3 billion, making it the second largest country in the world.

[1] More on this in the coming slides

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** [1] 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.

[1] More on this in the coming slides

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^2). Alaska, separated from the contiguous United States by Canada, is the 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 miles (28,311 km^2) in area.[136]

- 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)

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.

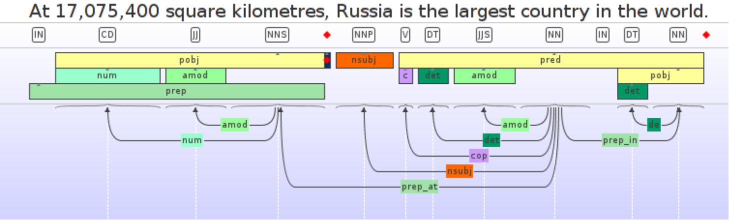
Relation Extraction as a Machine Learning Problem

- **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

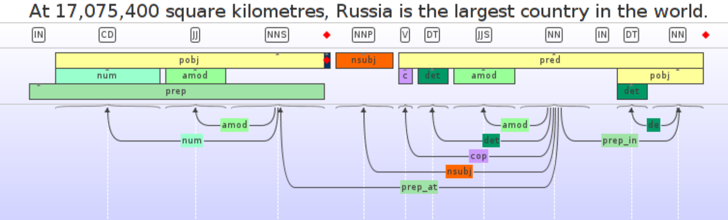
- Redundancy in grammatical features and dependencies of the sentences expressing same relation.

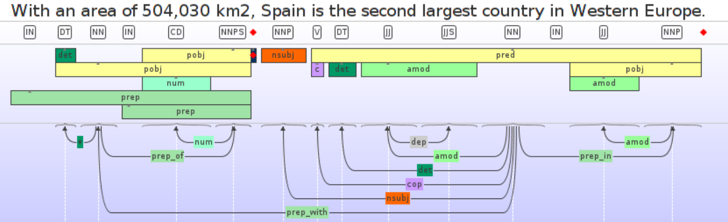
- Redundancy in grammatical features and dependencies of the sentences expressing same relation.



Relation Extraction as a Machine Learning Problem

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





Possible Workflow

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 trees, etc.) for these sentences.

Train a multi-class classifier on this training data.

for *sentence* $s \in$ *Corpus* **do**

    Extract features for s.

    Predict the relation using the model for these features.

    Store the fact into a database.

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 learn high quality extractors

Challenge

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

Peculiarities of Numerical Relation Extraction

Weak Signal from Distant Supervision

- Quantities can appear in far more contexts than typical entities.

- Quantities can appear in far more contexts than typical entities.
  - ("Bill Gates", "Microsoft") vs. ("India", "11%")

- Quantities can appear in far more contexts than typical entities.
  - ("Bill Gates", "Microsoft") vs. ("India", "11%")
- Noise is more for the small whole numbers that are unitless or with popular units (e.g, percent)

Peculiarities of Numerical Relation Extraction

Weak Signal from Distant Supervision

- Quantities can appear in far more contexts than typical entities.
  - ("Bill Gates", "Microsoft") vs. ("India", "11%")
- Noise is more for the small whole numbers that are unitless or with popular units (e.g, percent)
  - 1 or 5% vs. 11.42145 or 330 m/sec

Peculiarities of Numerical Relation Extraction

Weak Signal from Distant Supervision

- Quantities can appear in far more contexts than typical entities.
  - ("Bill Gates", "Microsoft") vs. ("India", "11%")
- Noise is more for the small whole numbers that are unitless or with popular units (e.g, percent)
  - 1 or 5% vs. 11.42145 or 330 m/sec
- Regular IE have fewer cases of entity disambiguation as compared to numerical IE

Peculiarities of Numerical Relation Extraction

Weak Signal from Distant Supervision

- Quantities can appear in far more contexts than typical entities.
  - ("Bill Gates", "Microsoft") vs. ("India", "11%")
- Noise is more for the small whole numbers that are unitless or with popular units (e.g, percent)
  - 1 or 5% vs. 11.42145 or 330 m/sec
- Regular IE have fewer cases of entity disambiguation as compared to numerical IE
  - Entities with name "Michael Jordan" vs. value "10" associated with a single entity

Peculiarities of Numerical Relation Extraction

Match Mines

- When a certain Knowledge Base entry causes an unprecedented number of matches in the corpus.
- Typically happens when *arg1* is a popular entity and *arg2* is a small whole number.
  - Example TODO

Peculiarities of Numerical Relation Extraction

Units

- Unit acts as types for numbers.
- Same quantity may be expressed with different unit
  - 20 kms or 12.4 miles
- Unit extractor needs to perform unit conversions for effective matching and extraction

- Often sentences express the change in the value of a relation instead of, or in addition to, the actual value itself.
  - Amazon stock price increased by $35 to close at $510.
  - India's tiger population sees 30% increase.
  - Ford poised to raise dividend by 20% even as profit declines.

Peculiarities of Numerical Relation Extraction

Relation/Argument Scoping

- Additional modifiers to arguments or relation words may subtly change the meaning and confuse the extractor.
  - *rural* literacy rate of India
  - literacy rate of *rural* India

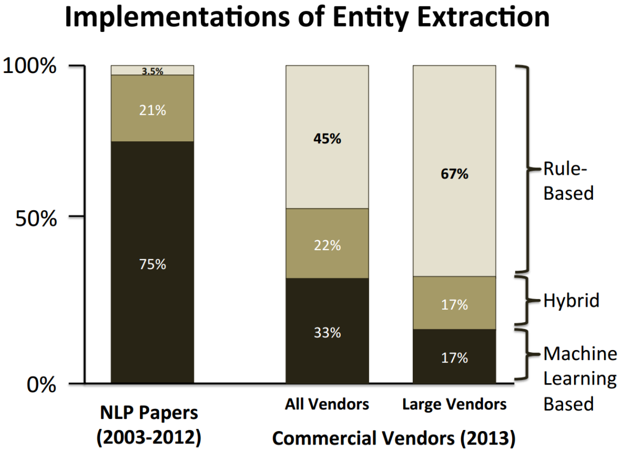
Peculiarities of Numerical Relation Extraction

Keywords

- Many numerical relations are described using one or a handful of keywords.
  - For example, sentences expressing *inflation rate*, *GDP*, *life expectancy* would often use these keywords.

Rule Based IE

Academia vs. Industry

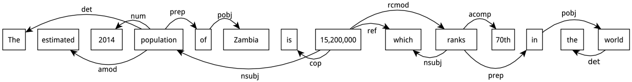


2

[2] from [CLR13]

From [BM05]

If e_1 and e_2 are two entities mentioned in the same sentence such that they are observed to be in a relationship R, our hypothesis stipulates that the contribution of the sentence dependency graph to establishing the relationship $R(e_1, e_2)$ is almost exclusively concentrated in the shortest path between e_1 and e_2 in the undirected version of the dependency graph.



From [BM05]

If e_1 **and** e_2 are two entities mentioned in the same sentence such that they are observed to be in a **relationship R**, our hypothesis stipulates that the contribution of the sentence **dependency graph** to establishing the relationship $R(e_1, e_2)$ is almost **exclusively concentrated in the shortest path** between e_1 and e_2 in the undirected version of the dependency graph.

- When looking for clues for relation extraction, dependency path is a good place to start.
- In the case of Numerical Relations, we already know what to look for: *keywords*.
- Need to take care of modifications to the entities, delta words

- **Keywords** Words that might help in identifying relations.(*GDP*, *internet*, *inflation*)
- **Delta words** Words that indicate that the mention expresses a change, and not the actual relation.
  (*change*, *up*, *down*, *increased*, *changed*, *risen*)
- **Modifiers** A word m is said to be a modifier of the word w if there is a modifying dependency from m to w.
  (*blue* modifies *whale* in **blue** *whale*, **urban** *population*).
- **Augmented Phrase** For a word W, the *Augmented Phrase* W' is formed by concatenating W with words P such that W and P are related via a *modifying dependency*.
  (for the word *whale*, **blue whale** is the augmented phrase)

NumberRule I
Extraction Algorithm

Given a sentence S, let E_S be the set of entities and N_S be the set of numbers that are present in the sentence.
Let R be the set of relations
Let $LegalUnits(r)$ be the set of possible units for a relation r in R
Let $Unit(n)$ be the unit for a given number n as returned by the unit tagger.
Let K be the set of keywords
Let Δ be the set of delta words
for $(e, n) \in (E_S \times N_S)$ **do** //For all entity-number pairs
    $P \leftarrow$ the set of words in the dependency path between e and n
    for $r \in R$ **do**
        if $P \cap K_r = \emptyset$ **then** //keyword is not present
            continue;

NumberRule II

Extraction Algorithm

if $P \cap \Delta \neq \emptyset$ **then** //delta words are present
    continue;
if $Unit(n) \notin LegalUnits(r)$ **then** //incompatible units?
    continue;
Extract $r(e', r', n)$, where e' and r' are the augmented entity and relation phrases.

- "The estimated population for 2014 of the Australian continent is about 36.25 million people"
- Australian $\xrightarrow{amod}$ continent $\xrightarrow{prep\_of}$ 2014 $\xrightarrow{prep\_for}$ **population** $\xrightarrow{nsubj}$ people $\xrightarrow{num}$ million $\xrightarrow{number}$ 36.25
- "The estimated population for 2014 of the Australian continent increased by about 3.25 million people"
- Australian $\xrightarrow{amod}$ continent $\xrightarrow{prep\_of}$ 2014 $\xrightarrow{prep\_for}$ **population** $\xrightarrow{nsubj}$ **increased** $\xrightarrow{prep\_by}$ people $\xrightarrow{num}$ million $\xrightarrow{number}$ 36.25
- As an example of the importance of using augmented phrases, consider the sentence:
- As another example, consider "Female population of urban india is 23 million". The shortest path between the arguments (India, 23) is (Population, million) and thus a vanilla system will yield the wrong extraction **POP(India, 23 million)**, instead of the correct **POP(urban india, female population, 23 million)**.

Sentence	Test
The estimated population of Australia is about 36.25 million people.	-
The estimated population density of Australia is 36.25 million people **per sq km.**	Incompatible Units
The estimated population of Australia **increased** *by about 36.25 million people.*	Delta Word Present
The estimated population of **urban** *Australia is about 36.25 million people.*	Entity is Modified
The estimated **adolescent** *population of Australia is about 36.25 million.*	Entity is Modified
The estimated populations in 2014 are Australia, **100 million** *and New Zealand, 36.25 million.*	100 million is closest to Australia

Table : NumberRule outputs (Australia, Total Population, 36.25 million) only in the first sentence. The second column is test number that fails for other sentences. The input keyword is "population".

For an entity e (India)

- Let S_e be the set of sentences that express the entity e.
- Let Q_e denote the distinct numbers with unit that appear in footnote. We use the unit tagger by [SC14] to identity units of numbers in the text and to convert all unit variants like "mile", "km" to a canonical SI unit, "meter".
- $\forall q \in Q_e$, let $S_{e,q} \subseteq S_e$ denote the sentences that mention e and q.

For each entity e, for each number n_q

- n_r^q, **number nodes** Binary, 1 if the number q is related to e via relation r.

For each mention, $s \in S_{e,q}$

- z_s, **mention nodes** multi-ary, can take values $r \in \mathcal{R} = (R \cup \perp)$, set to $r \in R$ if the sentence expresses any of the R relations, else set to $z_s = \perp$.

NumberTron

Features: Example

Feature type	Features
Fixed Keywords	key: life key: expect
All Keywords	key: life key: expect key: world
Number Features	Num: Billion Num: Integer

Afghanistan , which is mostly rural , has one of the lowest life expectancy rate in the world at 44 year for both man and woman.

The time "44 year" is converted to the SI unit, which comes out to be around 1.3 billion and thus the feature Num: Billion is fired.

- **Mintz Features** Lexical and Synctactic features derived from POS tags and dependency path [MBSJ09]
- **Keyword Features** Derived from a pre-specified list of keywords per relation.
- **Number Features** Capture Information on the magnitude, type (whole, fraction) can also be useful for relation extraction.

NumberTron Training

Perceptron

- The classical perceptron forms the core of our training procedure.

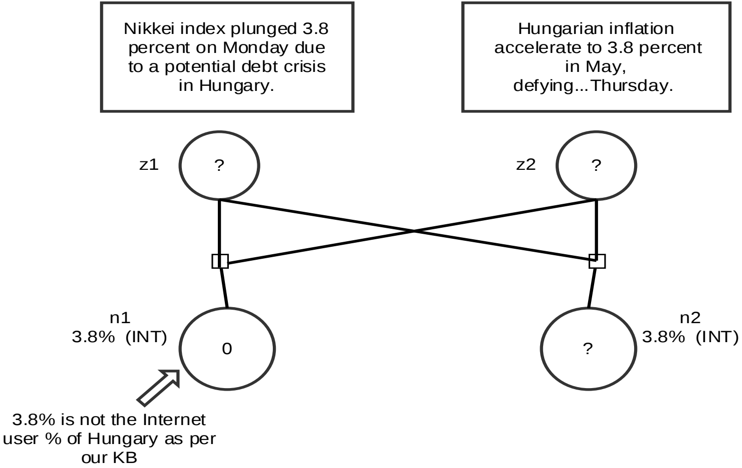
$$\theta \leftarrow \theta + \eta * (t_i - o_i) * x_i \tag{1}$$

$$\boxed{\theta} \leftarrow \boxed{\theta} + \eta * (\boxed{t_i} - \boxed{o_i}) * \boxed{x_i}) \tag{2}$$

- Weights
- True Label
- Observed Label
- Feature (Binary)

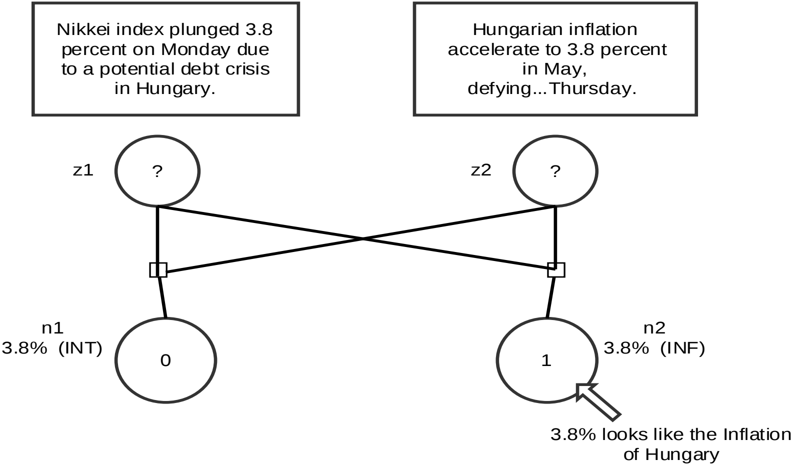
NumberTron Training

True Labels: Distant Supervision



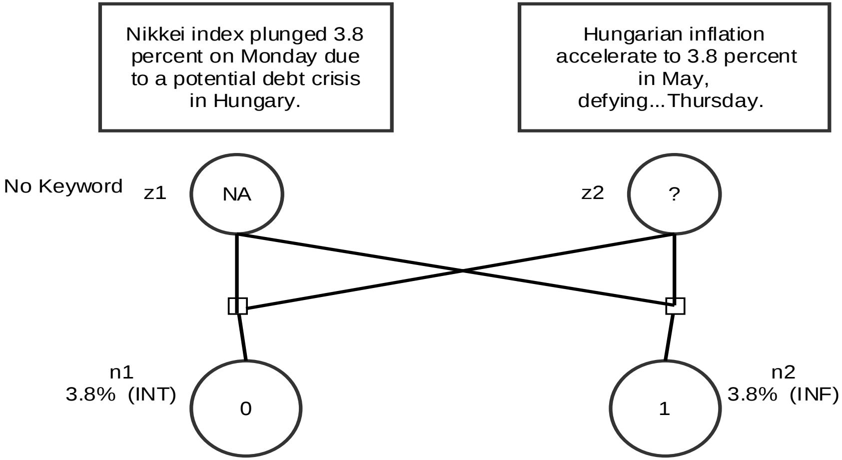
NumberTron Training

True Labels: Distant Supervision



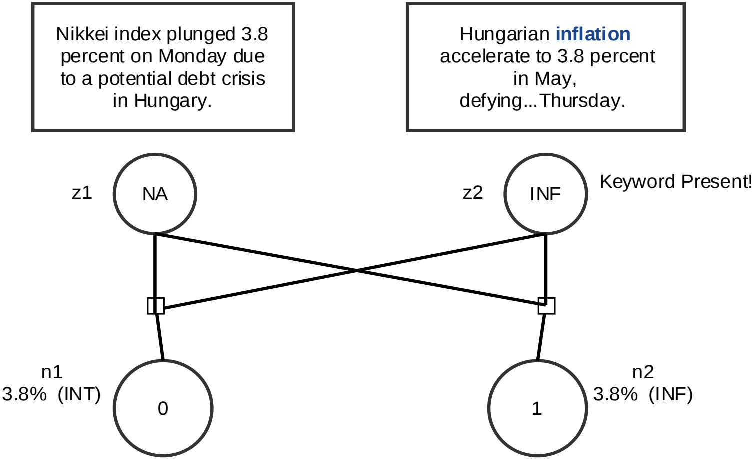
NumberTron Training

True Labels: Distant Supervision



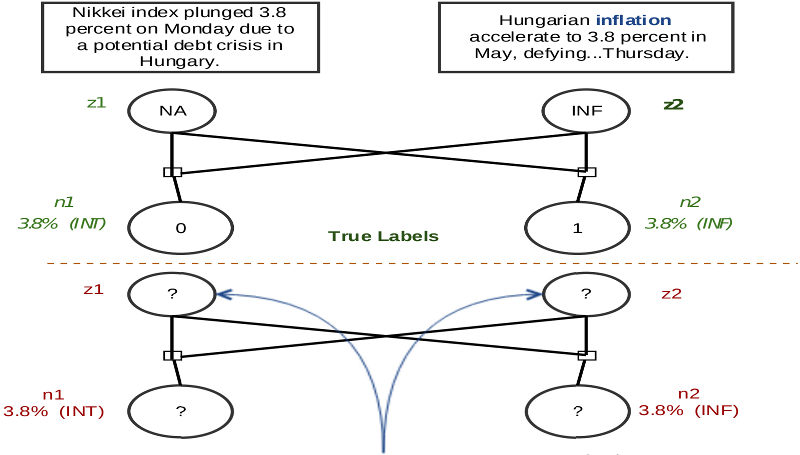
NumberTron Training

True Labels: Distant Supervision



NumberTron Training

Observed Labels: Full Inference

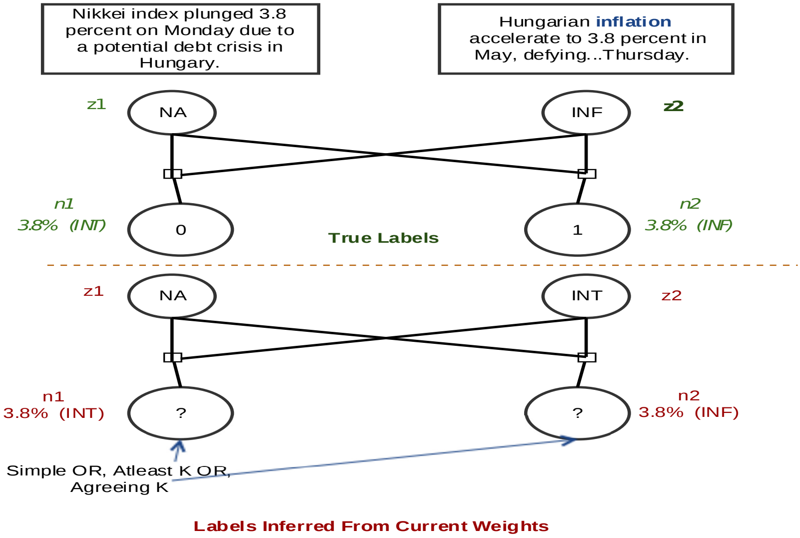


$$\hat{z}_s = \mathrm{argmax}_{r \in \mathcal{R}} \, \theta_s \phi_s(r)$$

Labels Inferred From Current Weights

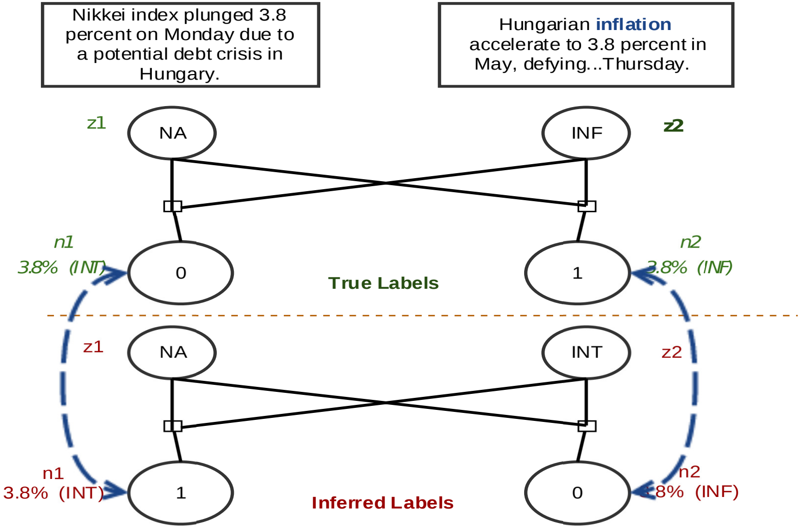
NumberTron Training

Observed Labels: Full Inference



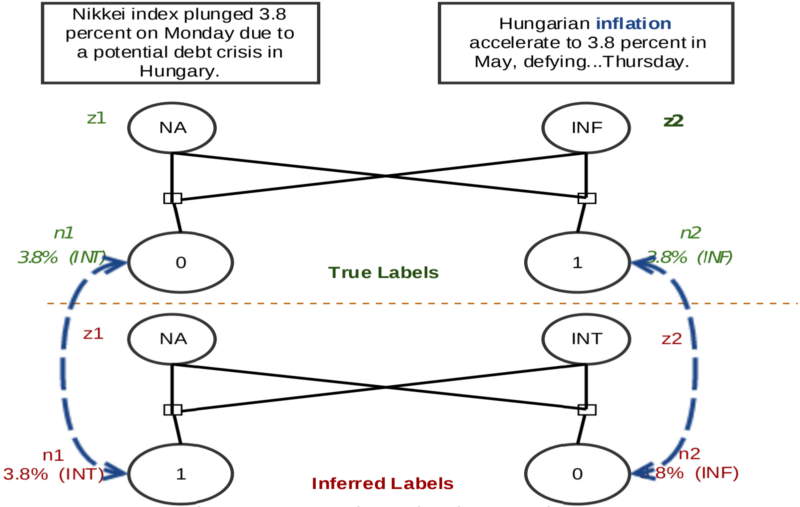
NumberTron Training

Observed Labels: Full Inference



NumberTron Training

Observed Labels: Full Inference



Disagreement, need to update feature weights!

- Let $f_1, f_2, ..., f_k$ be the features fired for *Hungarian inflation accelerate to 3.8 percent in May, defying...Thursday.*
- Examples: **key: inflation**, **Num: Units** and so on.
- $\theta_{f_i}^{INT} \leftarrow \theta_{f_i}^{INT} - 1$
- $\theta_{f_i}^{INF} \leftarrow \theta_{f_i}^{INF} + 1$
- These features actually indicate inflation relation, and not the internet relation!

Experiments

Training Corpus

- Tac KBP 2014 corpus comprising roughly 3 million documents from NewsWire, discussion forums, and the Web.
- Knowledge base is compiled from data.worldbank.org
  - Dataset contains 1,281 numeric indicators for 249 countries, with over 4 million base facts.
  - Dataset is normalized by converting all the values to their SI base unit value.

- Mix of 430 sentences from TAC corpus and sentences from Web search on relation name.

Relation	Units	Positive	Negative
Land Area	Sq. Km	57	17
Population	-	51	300
Inflation	percent	51	84
Internet Users	percent	15	
FDI	\$ (USD)	10	35
GDP	\$ (USD)	8	
Goods Export	\$ (USD)	11	
Life Expectancy	year	15	34
Electricity Production	kWh	13	6
CO_2 Emissions	kiloton	8	16

Table : Test corpus statistics: The third column is the number of instances per relation and the fourth column is the number of "none-on-the-above" (⊥) grouped by relation of the same unit.

Baseline Algorithms

- **Recall –Prior Baseline:** For each unit, predict the relation with the highest *test* prior ignoring the "none-of-the-above" class.
  - All the numbers with the unit "USD" will be labeled 'Goods Exported' since it is most frequent class ignoring the "none-of-the-above" class.

Baseline Algorithms

- **Recall –Prior Baseline:** For each unit, predict the relation with the highest *test* prior ignoring the "none-of-the-above" class.
    - All the numbers with the unit "USD" will be labeled 'Goods Exported' since it is most frequent class ignoring the "none-of-the-above" class.
- **MultiR ++ : Adapting MultiR for Numerical Relations**
    - Added unit tagger as in our algorithms for identifying and normalizing numbers and units.
    - Added our partial matching (using $\pm\delta_r\%$) technique in distant supervision.

Results

Numbertron vs NumberRule vs Baselines

System	Precision	Recall	F1 Score
MultiR++	31.81	28.10	29.84
Recall-Prior	28.18	86.19	42.47
NumberRule	59.30	53.60	56.30
NumberTron	60.93	66.92	63.78

Table : Aggregate results. NumberTron outperforms all other methods.

- Statistical method like NumberTron outperforms NumberRule on increased recall, which jumps from 53.6% to 67
- MultiR++ performs poorly because it does not model peculiarities of numerical relations.

Analysis

NumberTron vs NumberRule

- NumberRule's missed recall is primarily because of not having a keyword on the dependency path.
  - " *Turkey's central bank said Wednesday it expects the annual inflation rate to reach 6.09 percent at the end of 2009 , lower than the official target of 7.5 percent.*"
  - Turkey $\xrightarrow{poss}$ bank $\xrightarrow{nsubj}$ said $\xrightarrow{ccomp}$ expects $\xrightarrow{xcomp}$ reach $\xrightarrow{dobj}$ percent $\xrightarrow{num}$ 6.09
  - Since keyword 'inflation' is not on the shortest dependency path between Turkey and 6.09, NumberRule does not extract.
  - Since NumberTron combines evidences from multiple features such as number's range, presence of 'inflation' in context and dependency path features.

Ablation tests

of various configurations of NumberTron

Distant Supervision	**Simple OR**			**Atleast-K**			**Agreeing-K**		
	P	R	F1	P	R	F1	P	R	F1
KB	43.24	50.93	46.54	40.05	53.93	45.97	35.20	44.52	39.35
Keywords	43.35	73.22	54.46	43.69	73.62	54.83	45.97	70.80	55.74
KB + Keywords	61.56	64.96	63.21	60.93	66.92	63.78	63.46	60.21	61.79

Table : Comparison of various configurations for NumberTron

- Keywords are crucial and KB in conjunction with keyword-based labeling adds significant value.

Ablation tests

of feature templates for NumberTron

Features	Precision	Recall	F1-score
Mintz features only	22.85	36.86	28.21
Keyword features only	51.24	52.55	51.89
Mintz + Keyword	47.10	39.04	42.71
Mintz + Number	17.80	35.03	23.67
Keyword + Number	45.15	69.70	54.80
Mintz + Keyword + Number	*60.93*	*66.92*	*63.78*

Table : Ablation tests of feature templates for NumberTron

- Large set of Mintz features confuses the classifier; Keyword features are much effective in learning.

Relation	NumberTron F1	NumberRule F1
FDI	0	50.00
Life Expectancy	68.96	69.50
Internet Users	55.73	54.54
Electricity Prod.	50.00	62.50
GDP	57.14	42.80
CO_2 Emissions	47.61	53.30
Inflation	88.40	56.25
Goods export	75.00	35.20
Population	49.99	60.30
Land Area	57.44	52.22

Table : Per relation F1 scores for NumberRule and best configuration of NumberTron

Analysis

per relation analysis of NumberTron vs NumberRule

- For FDI relation, NumberTron does not make a extraction because in our corpus sentence expressing the relation are rare.
- Inflation and Population are well represented in training corpus and hence higher recall.

References I

Razvan C. Bunescu and Raymond J. Mooney.
A shortest path dependency kernel for relation extraction.
In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, 2005.

Laura Chiticariu, Yunyao Li, and Frederick R Reiss.
Rule-based information extraction is dead! long live rule-based information extraction systems!
In *EMNLP*, pages 827–832, 2013.

References II

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky.
Distant supervision for relation extraction without labeled data.
In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011, 2009.

Sunita Sarawagi and Soumen Chakrabarti.
Open-domain quantity queries on web tables: annotation, response, and consensus models.
In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 711–720, 2014.