Numerical Relation Extraction with Minimal Supervision

Aman Madaan [1] Ashish Mittal [2] Mausam [3] Ganesh Ramakrishnan [4] Sunita Sarawagi [4]

[1]Visa Inc

[2]IBM Research

[3]IIT Delhi

[4]IIT Bombay

Most of the work done while Aman and Ashish were graduate students at IIT Bombay

Introduction

Motivation

- Relation Extraction has been around for a while (MUC 1991).
- Distant Supervision Based Solutions.
- First distant supervision paper came out in 1999 [CK99].

Preface: Distant Supervision

Quick Introduction

- Given a knowledge base for a relation, in the example ”born in”

Donald Knuth	Wisconsin
Srinivasa Ramanujan	Erode
Alan Turing	London

- Label the corpora by aligning with the KB
  - 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

- Born - In KB

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

- Born - In KB

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

- Born - In KB

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

Motivation

- The problem of relation extraction has been focused on entity-entity pairs (persons, organizations, locations).
- An important subset of numbers has received some attention [HZW10], [KZBA14], [RVR15], [DR10]
- Numbers as first class objects in the relation extraction setting.

Numerical Relations?

- A 2004 EU entrant of 38 million people, Poland is almost entirely reliant on coal for electricity and heat.
- About half of Greenland 's 60,000 people be native to the icebound island .
- Uranium is a chemical element with symbol U and atomic number 92.

Goal

- Build Information Extractors that given a sentence expressing a numerical relation, extract the fact tuples, with the second argument a number.
  - Population(Poland, 38million)
  - Population(Greenland, 60000)
  - Atomic Number(Uranium, 92)

Plan

Introduction

Peculiarities of Numerical Relation Extraction

NumberRule: Rule Based Relation Extraction

NumberTron: Probabilistic Relation Extraction

Results

Plan

Introduction

Peculiarities of Numerical Relation Extraction

NumberRule: Rule Based Relation Extraction

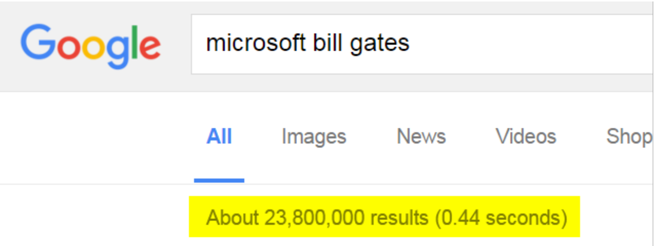
NumberTron: Probabilistic Relation Extraction

Results

Peculiarities of Numerical Relation Extraction

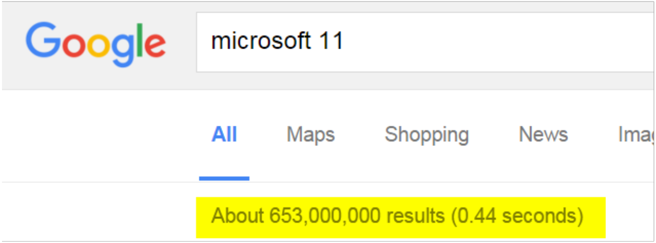
Numbers are more ambiguous

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



microsoft bill gates

About 23,800,000 results (0.44 seconds)



microsoft 11

About 653,000,000 results (0.44 seconds)

Peculiarities of Numerical Relation Extraction

Units

- Unit acts as types for numbers.
- Unit extractor[1] needed to perform unit conversions for correct matching and extraction.

[1] we use the open source unit tagger by [SC14]

Peculiarities of Numerical Relation Extraction

Delta Words

- Not uncommon to find sentences expressing change in the value of a relation (instead of, or in addition to, the actual value).
  - 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: Modifiers

- Additional modifiers to arguments or relation words may subtly change the meaning and confuse the extractor.
  - *rural* literacy rate of India
  - literacy rate of *south* India
- A word m is said to be a modifier of the word w if there is a modifying dependency from m to w.

Peculiarities of Numerical Relation Extraction

Keywords

- Sentences expressing many numerical relations usually include one or a handful of keywords.
- Sentences expressing the GDP of a country **without** mentioning the term *GDP*? Sentences expressing inflation without mentioning inflation?
- *Founder of* relation **without the phrase** *founder of*?
  - Bill Gates is the founder of Microsoft
  - Bill Gates founded Microsoft
  - Bill Gates is the father of Microsoft
  - Bill Gates laid the foundation stone of Microsoft
  - Bill Gates started Microsoft

Plan

Introduction

Peculiarities of Numerical Relation Extraction

NumberRule: Rule Based Relation Extraction

NumberTron: Probabilistic Relation Extraction

Results

NumberRule

Problem Statement

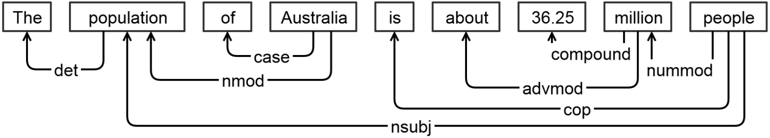
- Given:
  - A sentence S, with an entity **e** and a number **n**.
  - A set of numerical relations R
- Using:
  - A set of **keywords** for each of the numerical relations $r \in R$ (*GDP*, *internet*, *inflation* etc.) and **delta words** (*increased*, *changed* etc.)
  - Information about units for relations $r \in R$.
- Answer: Are **e** and **n** connected by one of the numerical relations $r \in R$?

NumberRule

Motivation

- When looking for clues for relation extraction, dependency path is a good place to start [BM05].
- 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

Dependency Path?



NumberRule

Extraction Algorithm

Create the dependency path P, and in P, check that:

C1. Keyword is present ✗

Australia has 36.25 million SUVs

NumberRule

Extraction Algorithm

Create the dependency path P, and in P, check that:

C1. Keyword is present ✓

C2. Delta words are not present ✗

The population of Australia **increased** by about 36.25 million.

NumberRule

Extraction Algorithm

Create the dependency path P, and in P, check that:

C1. Keyword is present ✓

C2. Delta words are not present ✓

C3. Units are compatible ✗

The population density of Australia is 36.25 million people **per sq km**.

NumberRule

Extraction Algorithm

Create the dependency path P, and in P, check that:

C1. Keyword is present ✓

C2. Delta words are not present ✓

C3. Units are compatible ✓

C4. Keyword is not modified/scoped ✗

The **adolescent** population of Australia is about 36.25 million people.

NumberRule

Extraction Algorithm

Create the dependency path P, and in P, check that:

C1. Keyword is present ✓

C2. Delta words are not present ✓

C3. Units are compatible ✓

C4. Keyword is not modified/scoped ✓

C5. Entity is not modified/scoped ✗

The population of **urban** Australia is about 36.25 million people.

NumberRule

Extraction Algorithm

Create the dependency path P, and in P, check that:

C1. Keyword is present ✓

C2. Delta words are not present ✓

C3. Units are compatible ✓

C4.Keyword is not modified/scoped ✓

The population of Australia is about 36.25 million people.

C5.Entity is not modified/scoped ✓

→ All good! add extraction population(Australia, 36.25 million)

Plan

Introduction

Peculiarities of Numerical Relation Extraction

NumberRule: Rule Based Relation Extraction

NumberTron: Probabilistic Relation Extraction

Results

NumberTron

Problem Statement

- Given
  - An Unlabeled Corpus (Sentencified, pruned to retain sentences having a country and a number)
  - A knowledge base of numerical facts.
  - A set of keywords
- Build Numerical Extractors.

NumberTron

Graphical Model Overview

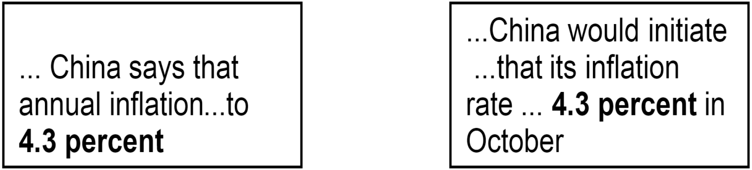
- One possibly disjoint graph per entity, θ shared across the graphs.

- Collect:
  - S_e: sentences that have a mention of e.
  - Q_e: all the numbers with units present in S_e.

- For each entity e and relation r, create:
  - n, number nodes, binary, capture the confidence that the number is a valid member of the relation $r(e, n)$.
  - z, sentence nodes, binary, confidence that the sentence can express the relation r for e.

NumberTron Training

True Labels: Distant Supervision

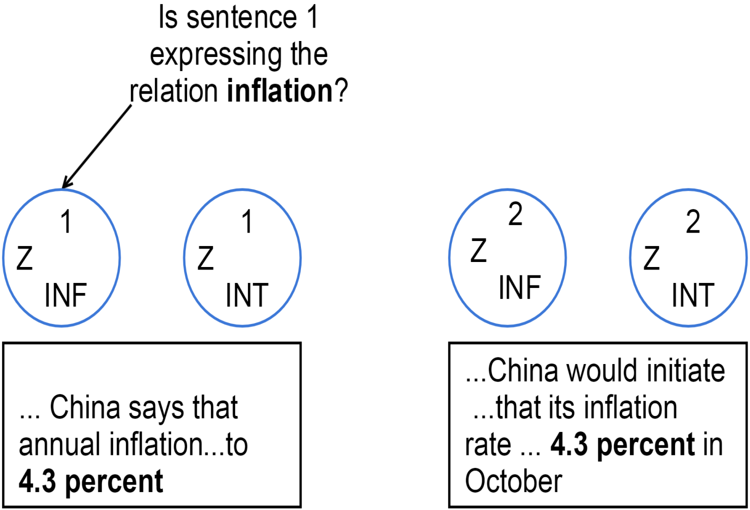


... China says that annual inflation...to **4.3 percent**

...China would initiate ...that its inflation rate ... **4.3 percent** in October

NumberTron Training

True Labels: Distant Supervision



Is sentence 1 expressing the relation **inflation**?

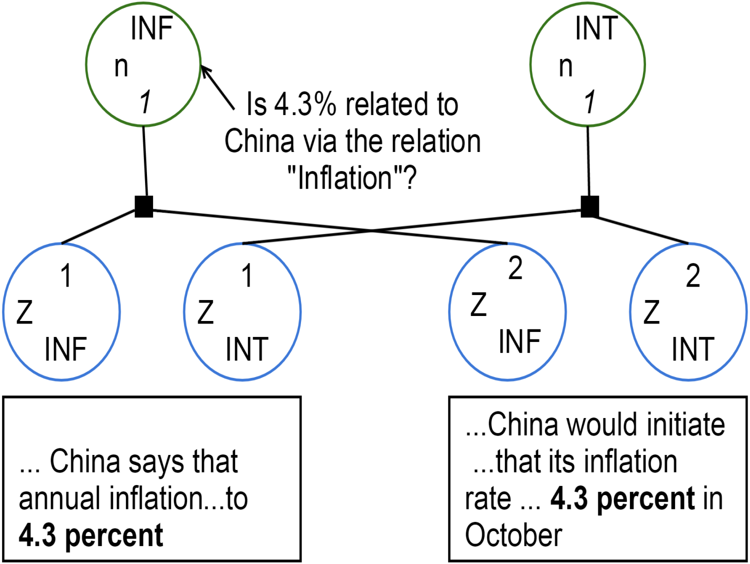
Z^{1}_{INF} Z^{1}_{INT} Z^{2}_{INF} Z^{2}_{INT}

... China says that annual inflation...to **4.3 percent**

...China would initiate ...that its inflation rate ... **4.3 percent** in October

NumberTron Training

True Labels: Distant Supervision



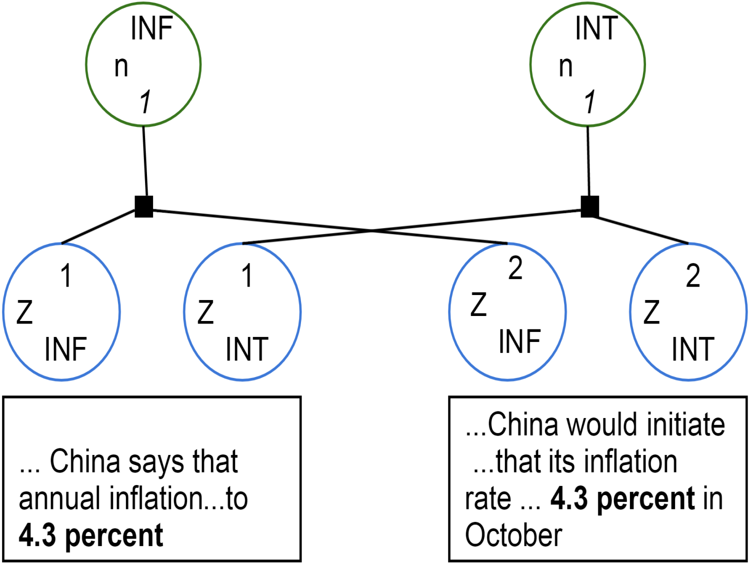
Is 4.3% related to China via the relation "Inflation"?

... China says that annual inflation...to **4.3 percent**

...China would initiate ...that its inflation rate ... **4.3 percent** in October

NumberTron Training

True Labels: Distant Supervision

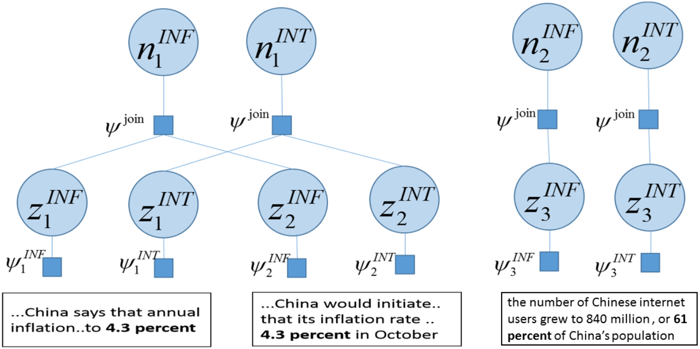


... China says that annual inflation...to **4.3 percent**

...China would initiate ...that its inflation rate ... **4.3 percent** in October

NumberTron

Graphical Model



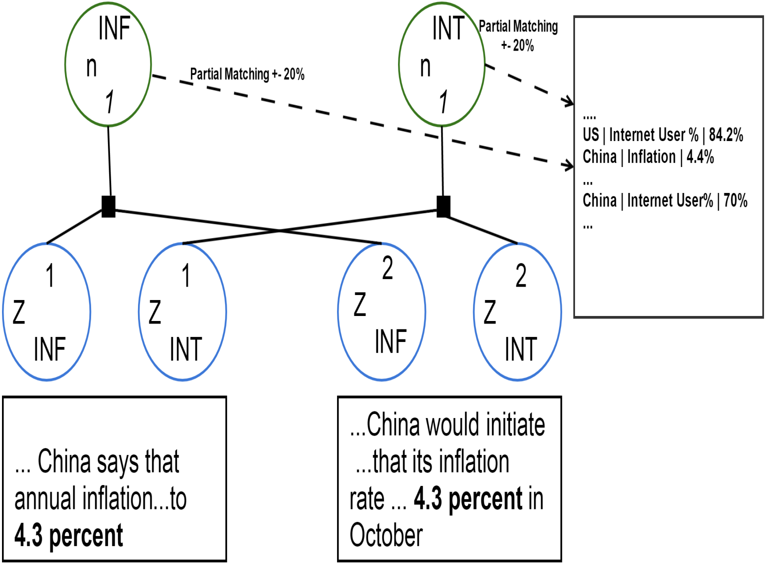
...China says that annual inflation..to **4.3 percent**

...China would initiate.. that its inflation rate .. **4.3 percent** in October

the number of Chinese internet users grew to 840 million , or **61 percent** of China's population

NumberTron Training

True Labels: Distant Supervision



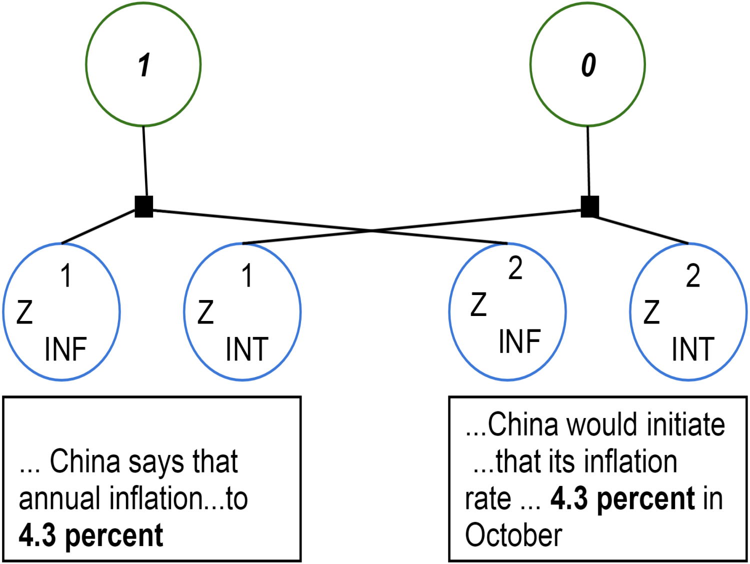
....
US | Internet User % | 84.2%
China | Inflation | 4.4%
...
China | Internet User% | 70%
...

... China says that annual inflation...to **4.3 percent**

...China would initiate ...that its inflation rate ... **4.3 percent** in October

NumberTron Training

True Labels: Distant Supervision

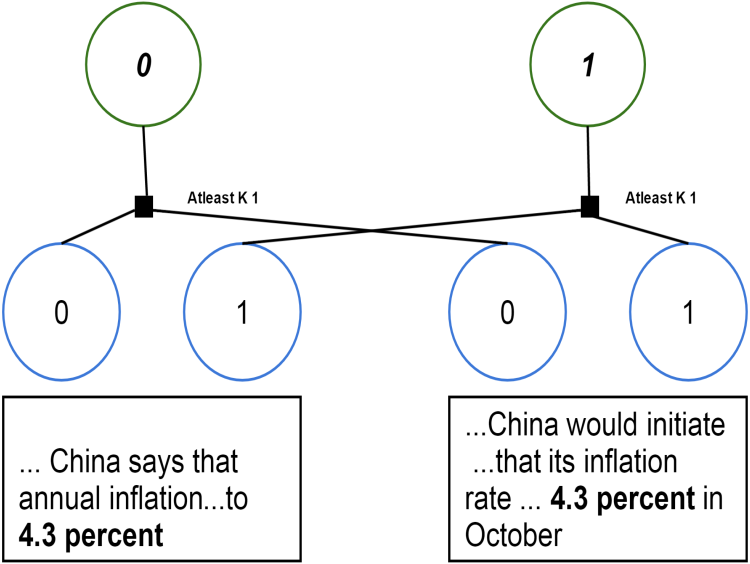


... China says that annual inflation...to **4.3 percent**

...China would initiate ...that its inflation rate ... **4.3 percent** in October

NumberTron Training

True Labels: Distant Supervision



... China says that annual inflation...to **4.3 percent**

...China would initiate ...that its inflation rate ... **4.3 percent** in October

NumberTron

Features

- Lexical and Syntactic features derived from POS tags and dependency path [MBSJ09] (...str:rural[rcmod]$->$ |LOCATION|[nsubj]...).
- **Keyword Features** Derived from a pre-specified list of keywords per relation (key: life key: expect).
- **Number Features** Magnitude, type (whole, fraction) of the number (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.

Plan

Introduction

Peculiarities of Numerical Relation Extraction

NumberRule: Rule Based Relation Extraction

NumberTron: Probabilistic Relation Extraction

Results

Experiments

- **Training Corpus**: Tac KBP 2014 corpus 3 million documents from NewsWire, discussion forums, and the Web.
- Knowledge base derived from data.worldbank.org, values normalized to their SI base unit value, selected 10 relations for the experiments.
- Test Set: Mix of 430 sentences from TAC corpus and sentences from Web search on relation name.
- Unit tagging done using the open source unit tagger by [Sarawagi and Chakrabarti 2014].
- Extractions are sentence level.

Experiments

KB and the Set of keywords

China	4.091616e+17	ELEC
Ukraine	9.27261850301	INF

Table: KB, for each relations the SI unit is used

Relation	Keywords
Internet User %	internet
Land Area	area, land
Population	population, people, inhabitants
GDP	gross, domestic, GDP
CO_2 emission	carbon, emission, CO2
Inflation	inflation
Goods Export	goods, export
Life Expectancy	life, expectancy
Electricity Production	electricity

Table: Set of Keywords

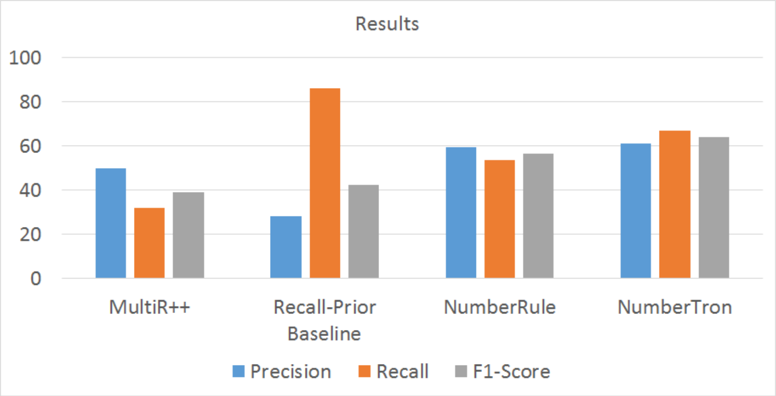
Baselines

- **MultiR ++**[HZL$^+$11]
  - Added unit tagger for identifying and normalizing numbers and units.
  - Added partial matching (using $\pm\delta_r\%$) technique in distant supervision.
- **Recall –Prior Baseline** Unit based prediction, relation with the highest frequency for a given relation wins.

Inflation	percent	51 ✓
Internet Users	percent	15

Results

Baselines vs NumberRule vs Numbertron



- NumberTron, statistical, outperforms NumberRule on increased recall (53.6% to 67%)

Ablation tests

of feature templates for NumberTron

Features	Precision	Recall	F1-score
Mintz features only	22.85	36.86	28.21
Mintz + Keyword	47.10	39.04	42.71
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.

Summary

- Numerical relation extraction has several peculiarities, more challenging than standard IE.
- **NumberRule**, a rule based system that can extract any numerical relation given input keywords for that relation.
- **NumberTron**, a probabilistic graphical model, that employs novel task-specific features and can be trained via distant supervision or other heuristic labelings.
- NumberTron aggregates evidence from multiple features and produces higher recall at a precision comparable to NumberRule.
- Both systems vastly outperform baselines and non-numeric IE systems, with NumberTron yielding over 33 point F-score improvement.

Thanks!

- Code, KB, and test data at: **https://github.com/NEO-IE**

Questions?

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.

Mark Craven and Johan Kumlien.
Constructing biological knowledge bases by extracting information from text sources.
In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, August 6-10, 1999, Heidelberg, Germany*, pages 77–86, 1999.

References II

Dmitry Davidov and Ari Rappoport.
Extraction and approximation of numerical attributes from the web.
In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 1308–1317, 2010.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S. Zettlemoyer, and Daniel S. Weld.
Knowledge-based weak supervision for information extraction of overlapping relations.
In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 541–550, 2011.

References III

- Raphael Hoffmann, Congle Zhang, and Daniel S. Weld.
  Learning 5000 relational extractors.
  In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 286–295, 2010.

- Nate Kushman, Luke Zettlemoyer, Regina Barzilay, and Yoav Artzi.
  Learning to automatically solve algebra word problems.
  In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 271–281, 2014.

References IV

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.

Subhro Roy, Tim Vieira, and Dan Roth.
Reasoning about quantities in natural language.
TACL, 3:1–13, 2015.

References V

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.