Knowledge Base Augmentation Using Tabular Data

Yoones A. Sekhavat, Denilson Barbosa
University of Alberta

Francesco di Paolo, Paolo Merialdo
Roma Tre University

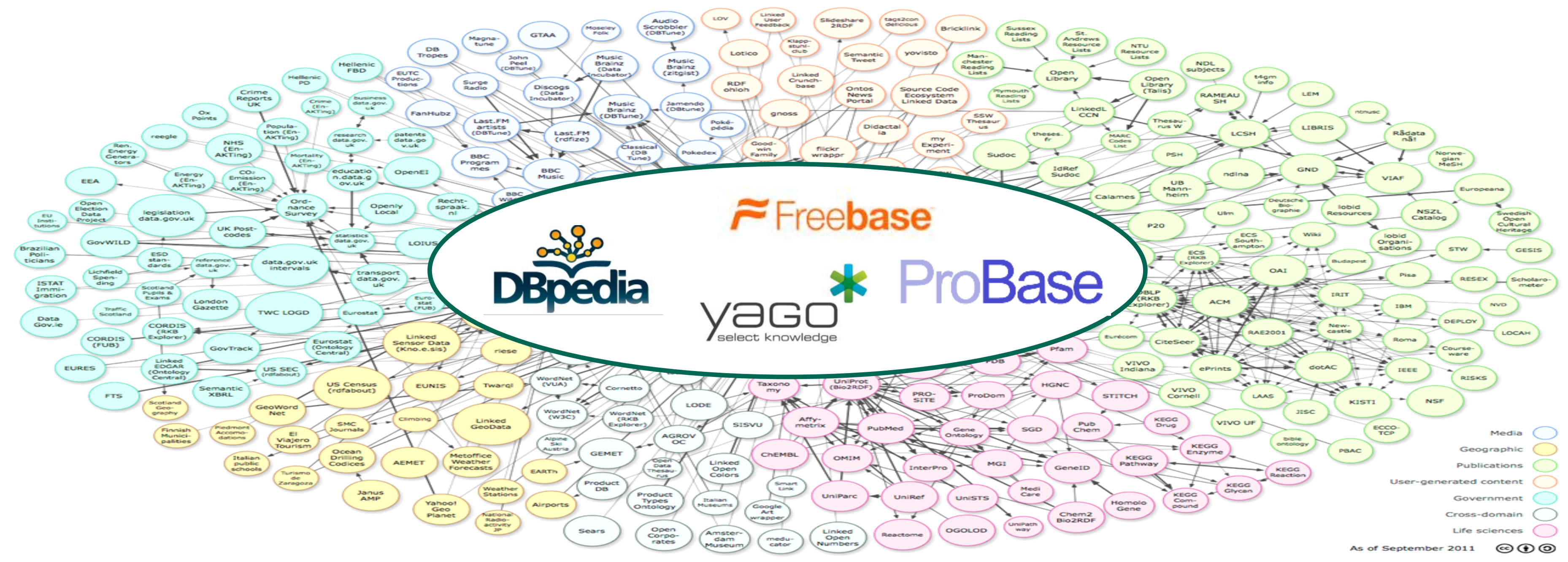
7th International Workshop on Linked Data on the Web

Outline

- Motivation
  - Need for knowledge bases
  - Exploiting semantics of tabular data
- Triple extraction
  - Architecture
  - Probabilistic model
- Implementation
- Experiments and results
- Limitations and future work

All data used to build and test the models in our work can be found at
http://cs.ualberta.ca/~denilson/data/ldow14_ualberta_data.zip

Knowledge Bases at the Core of the LOD Cloud

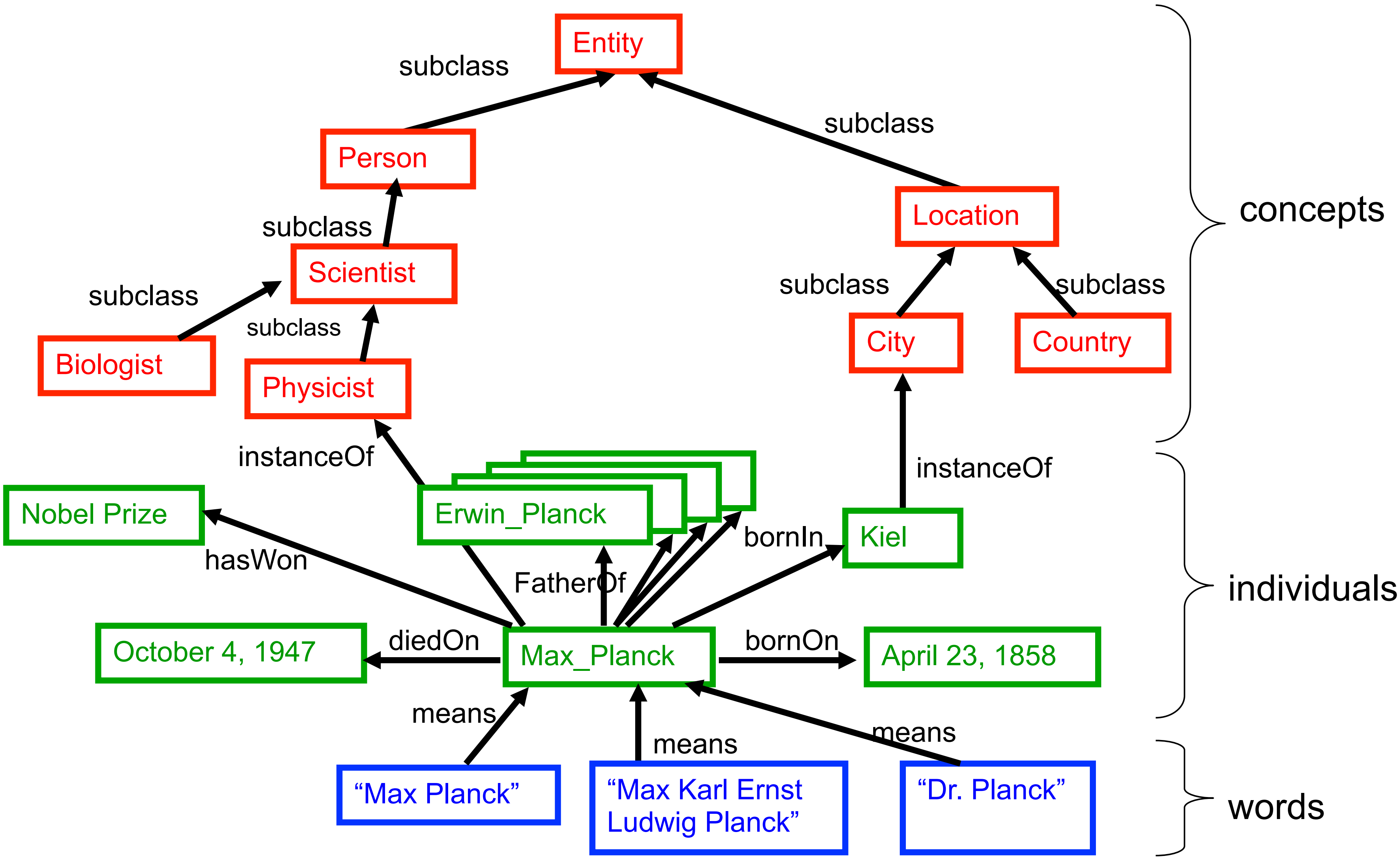


- Semantic query answering
- Information integration
- Data cleaning
- Record linkage

→ Example

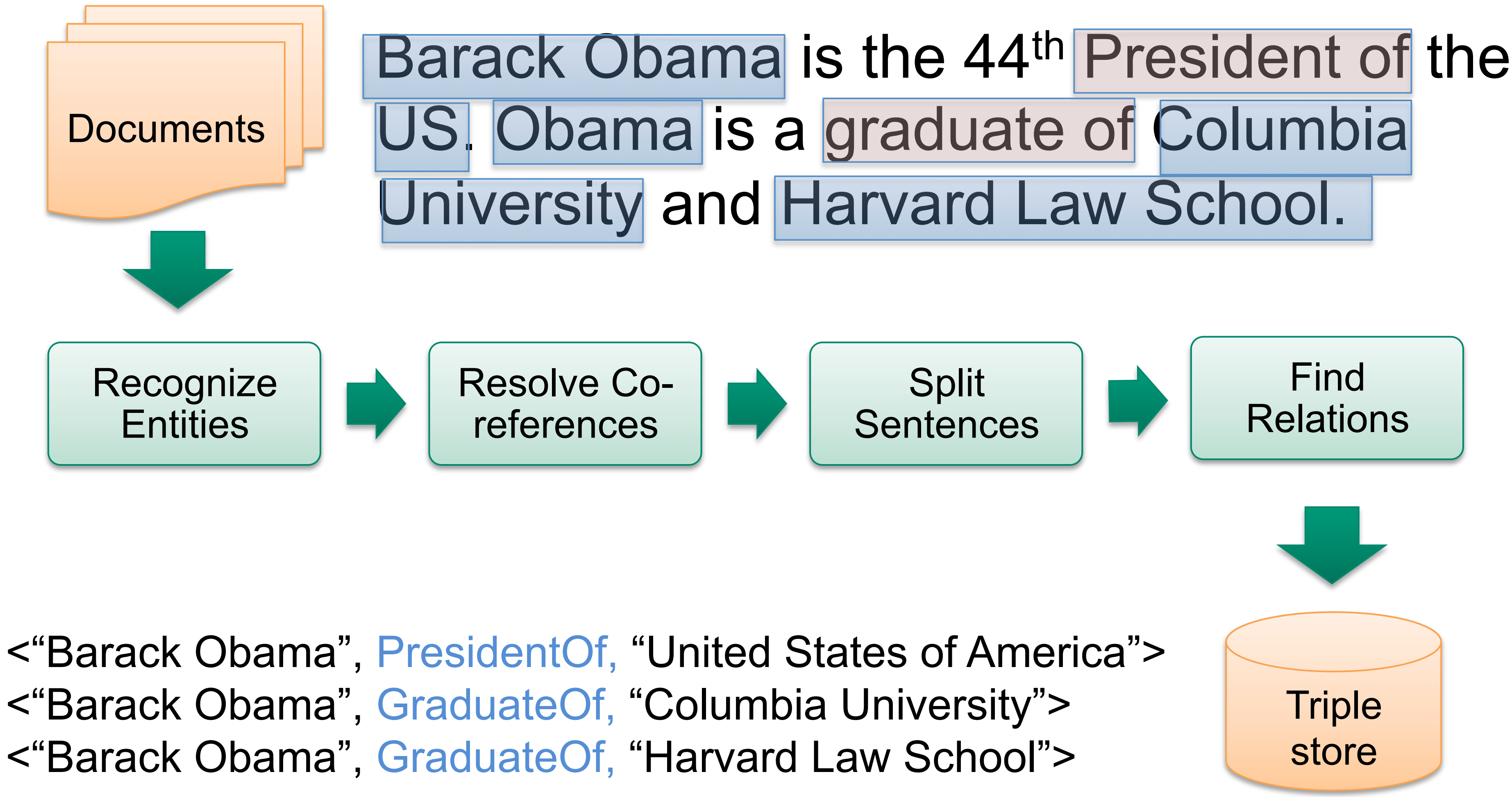
- Artists who are also politicians
- Which artists were born in the same place as John Lennon?

YAGO Knowledge Base



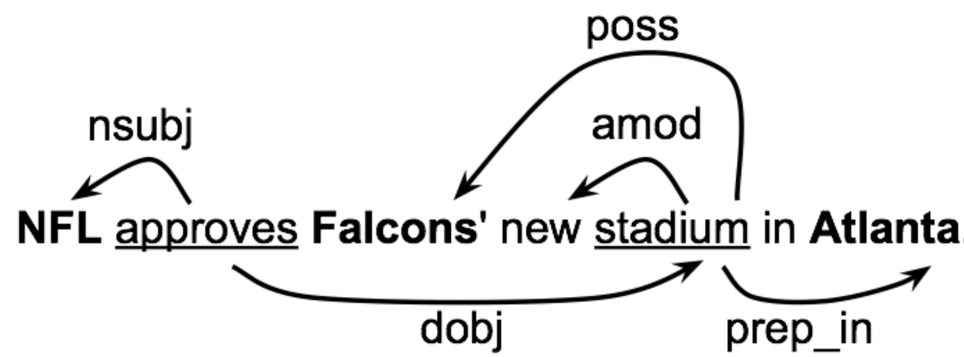
Slide from [Weikum, WSDM2009]

Relation Extraction / Text Mining

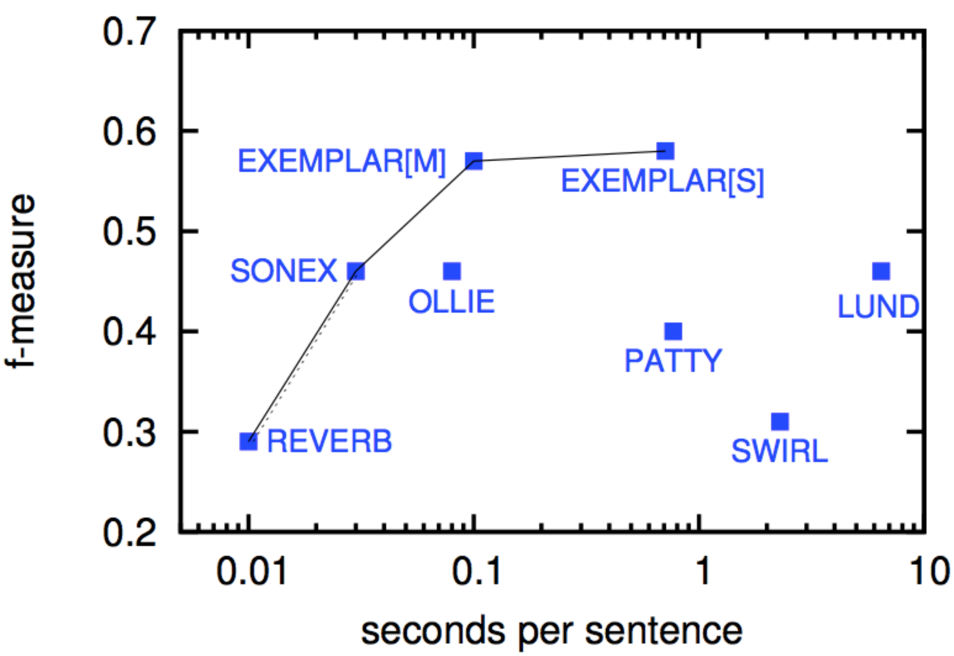


Relation extraction with dependencies

- Comparison of different relation extraction techniques and varying cost/benefit trade-offs [EMNLP'2013]

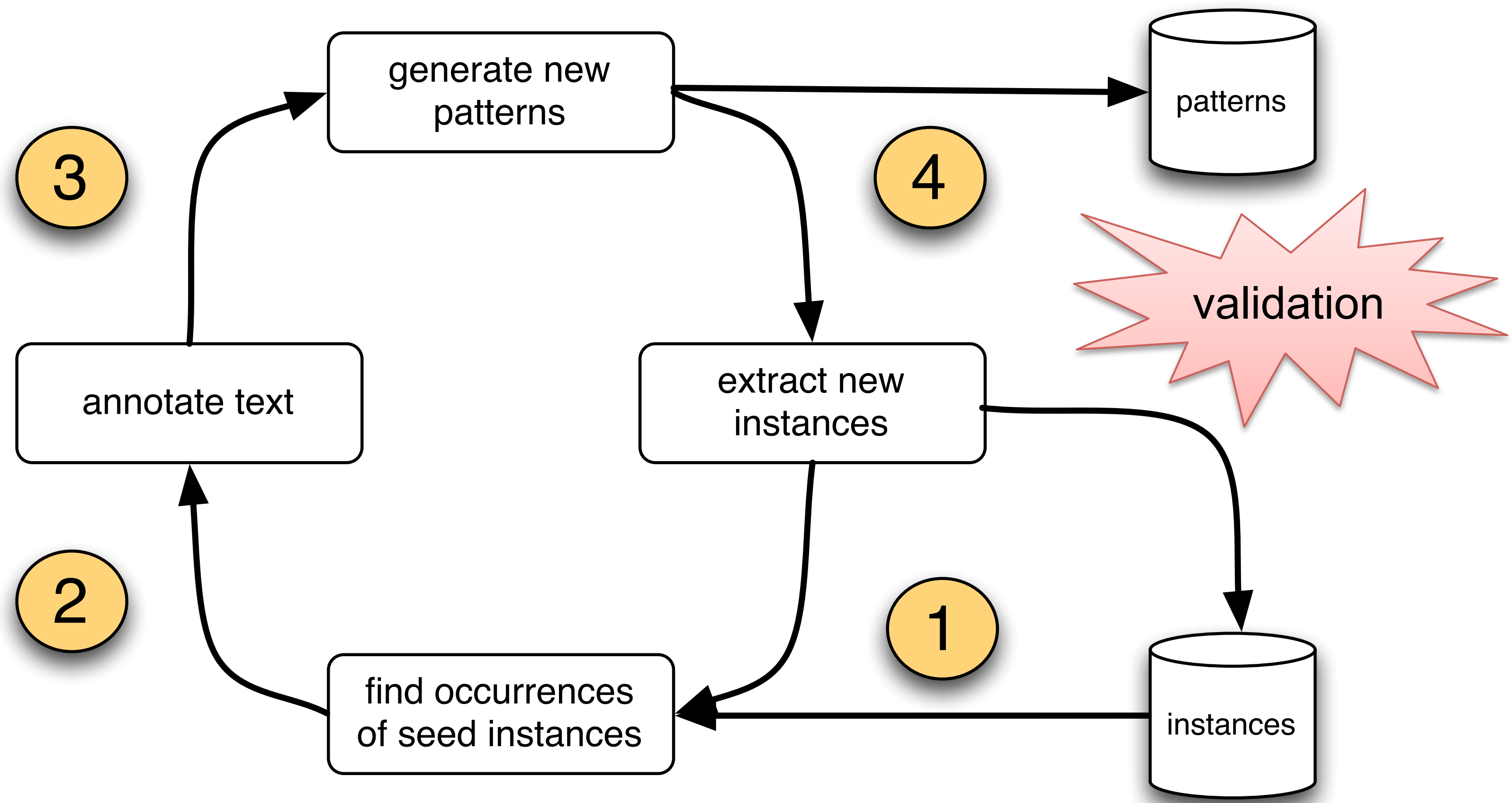


EXEMPLAR
https://github.com/U-Alberta/exemplar/



How does a knowledge base get built?

- Reinforcement cycle: find new instances, generate new patterns, test and repeat!



Problem statement

- Augmenting an existing knowledge base with facts expressed in tabular data on the Web

- Why tabular data?
  - Tables have inherent semantics which are often implicit
  - Tables are everywhere !

Over 400,000 Excel spreadsheets in the Clueweb'09

154 million relational tables in English documents in Cluweb'09

389, 697 raw and geospatial datasets in data.gov

(Cafarella et al. 2008)

147 million relational tables in the 2012 Web Common Crawl

Example

(A snapshot of a table in Wikipedia)

Ronaldinho	Brazil	Barcelona FC
Fabio Cannavaro	Italy	Juventus
Kaka	Brazil	AC Milan
Lionel Messi	Argentina	Barcelona FC

- General approach:
  - link the values in each cell to known entities in a KB
  - identify relations between the linked values.

Best case scenario

- Entities are linked to the same KB
- Relation already exists between entities

Table understanding (e.g., Limaye et al., 2010)

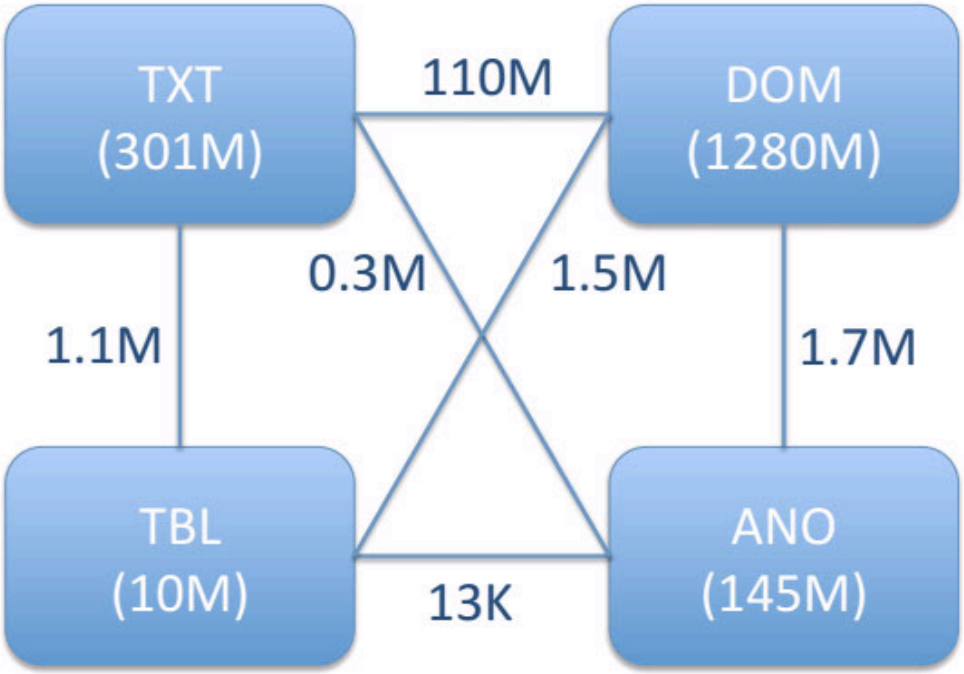
Our take

- Entities in different or unlinked KB
- Entities are not linked to anything yet.

Knowledge base augmentation

Some insight into Google's Knowledge Graph

- Thanks to Xin Luna Dong (Google), from yesterday's talk at DEOS:
- TXT: text extraction
- DOM: deep-web extraction
- ANO: schema.org annotations
- TBL: Web tables
  - Schema matching/table understanding approach



Goal

Augmenting an existing repository with new instances of relations already defined

Idea

- The fact that someone put some literals together in the same rows indicates that there are relationships between them
- Pairs of entities in different rows of two given columns share the same relation

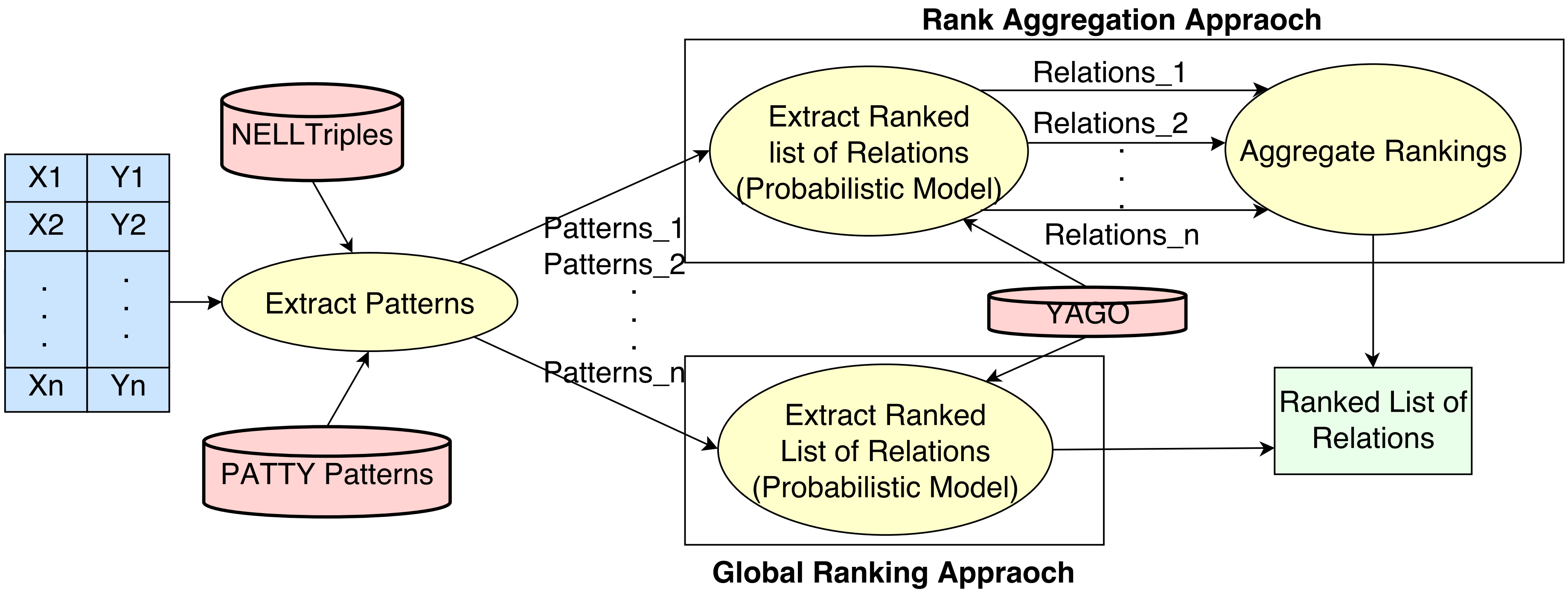
Method

1. Collect all sentences containing both entities from a large text corpus
2. Extract the text in between them
3. Match those texts against the list of patterns
4. Estimate the posterior probability of all candidate relations.

V1.0

- Knowledge base
  - We used YAGO with about 10 million entities and over 120 million facts about them.
- Text corpus
  - Our text corpus is the NELL Subject-Verb-Object (SVO) triple corpus, with about 604 million triples extracted from ClueWeb09 dataset.
  - Clueweb09 is a crawl of the Web with about 1 billion web pages in ten different languages.
- Text patterns
  - We used publicly available patterns from the PATTY project
  - We used 4,357 distinct patterns from PATTY having intersection with NELL (intersection with 108,699,400 triples from NELL)
- Ground truth
  - Facts from YAGO relations where both entities can be matched exactly in the NELL corpus.

Rank Aggregation vs. Global Ranking



A relation in a knowledge base can be represented by different textual patterns

plays-for
"scored for"
"signed contract with"

A pattern may represent more than one relation

"played in"
plays-for
(e.g., "Messi played in 2006 world cup")
performed-at
(e.g., "Pink Floyd played in Pompeii")

Probabilistic model

- We use Bayesian inference to compute the posterior probability of relation *r* given the observed patterns $p_1,\ldots,p_k$.

Evidence variables are conditionally independent

$$Pr(r|p_1,\ldots,p_k) = \frac{Pr(r)Pr(p_1,\ldots,p_k|r)}{Pr(p_1,\ldots,p_k)}$$

$$Pr(r|p_1,\ldots,p_k) = \frac{Pr(r)\prod_{i=1}^{k} Pr(p_i|r)}{Pr(p_1,\ldots,p_k)}$$

- Estimating prior probabilities

$$Pr(r) = |r| / \sum_{r_i \in R} (|r_i|)$$

$$Pr(p|r) = |p| / \sum_{p_i \in PT(r)} |p_i|$$

- *R* is the set of all relations
- *PT(r)* is the set of patterns associated with relation *r*

Results (accuracy)

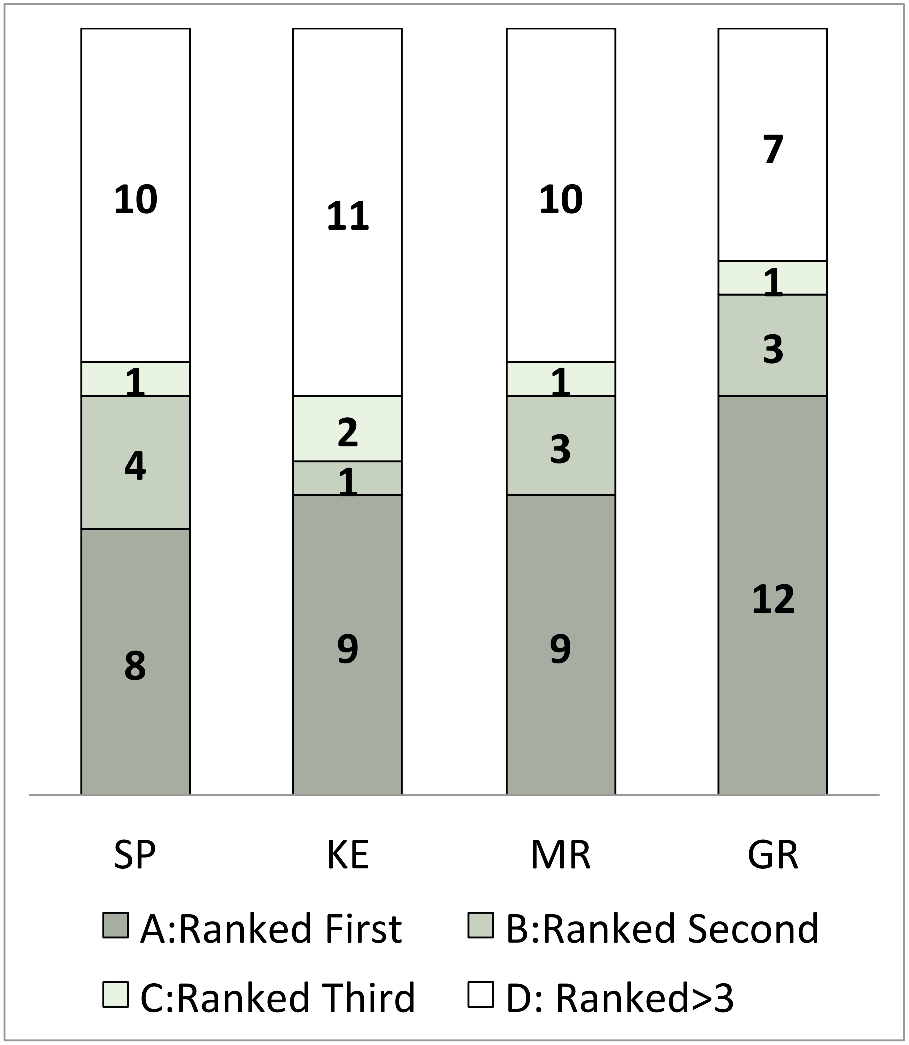
Number of PATTY patterns and resulting rank obtained by each strategy, for each relation.

Relation	Patterns	SP	KE	MR	GR
ismarriedto	1274	1	1	1	1
created	1148	1	1	1	1
haschild	1090	-	-	3	3
influences	694	1	1	2	-
actedin	624	2	2	1	1
graduatedfrom	472	1	1	1	1
isknownfor	452	-	-	-	-
worksat	447	-	-	1	1
holdspoliticalposition	417	-	3	1	1
directed	400	2	-	2	2
playsfor	354	1	1	1	1
diedin	335	3	-	1	2
wasbornin	273	-	-	1	1
islocatedin	249	-	3	-	-
livesin	200	-	-	-	-
isleaderof	156	-	-	-	-
iscitizenof	121	1	1	-	-
haswonprize	81	-	-	3	1
dealswith	59	-	-	-	1
ispoliticianof	49	1	-	-	1
participatedin	33	1	1	2	2
happenedin	12	2	1	-	1
hascapital	1	2	1	-	-

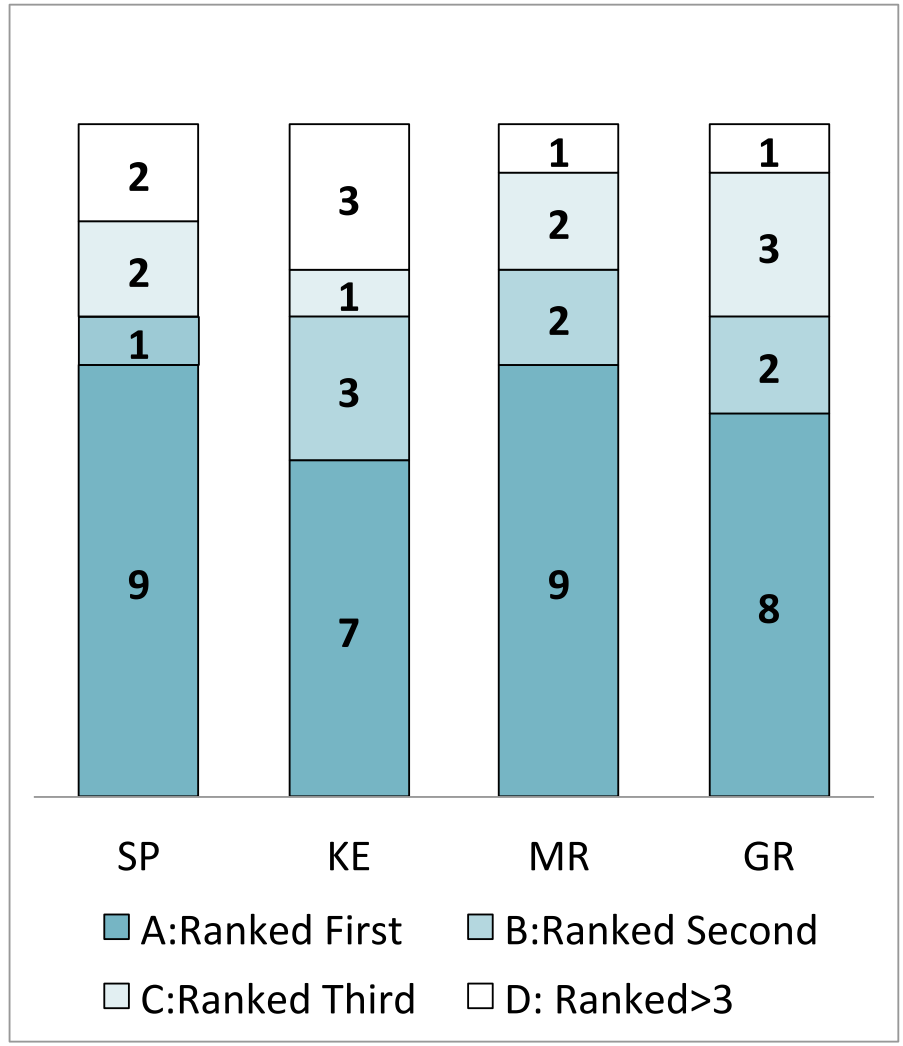
SP: Spearman's Footrule
KE: Kendall's tau
MR: Mean Ranking
GR: Global Ranking

Summary of accuracy results

All relations



Filtered relations



Knowledge augmentation pre-experiment

- Test #1:
  - Input table - a spreadsheet including song data available at http://www.aardvarkdjservices.co.uk
  - Our technique fond 48 triples for the created relation
  - Among those, 31 were already present in YAGO
- Test #2:
  - Input table - a spreadsheet with data about NBA players extracted from http://wwww.espn.go.com
  - Found 100 triples for the plays-for relation,
  - Of these, YAGO had 92 triples in the is-affiliated-to relation

Runtime

- The average execution times (ms) for processing a pair of entities (taken over 20 executions) are:

SP	KE	MR	GR
1688	1868	1729	1719

- There are no considerable differences among the methods

- The majority of the time is spent on matching the entities against the NELL corpus

Summary and Conclusion

- We described a probabilistic approach for augmenting linked open data repositories using tabular data.

- Unlike prior methods that focus on natural language understanding, we started from the (reasonable) assumption that all entities in the same row of a table are related by definition.

- Unlike previous methods that attempt to understand tabular data, we label pairs of columns in the table with relations coming from an established knowledge base.

Summary and Conclusion

- Limitations
  - Small number of YAGO relations ➔ currently experimenting with Freebase
  - Exact entity matching

- Other applications besides knowledge base augmentation
  - Estimating:
    - How many new triples could be extracted from tabular data on the Web?
    - How accurate are they?
  - Using both quantitative and qualitative metrics to chart which websites provide the best data for knowledge base augmentation

V1.1

- Knowledge base
  - YAGO ➔ Freebase
- Text corpus
  - Still NELL
  - Musing about indexing all of Clueweb for this
- Text patterns
  - PATTY ➔ Google's annotated Clueweb with Freebase entities
- Ground truth
  - Facts from YAGO
  - Facts from Freebase (ranging popularity)

Work in progress: (summary of results)

- Filtering
  - Relations associated with less than 1000 or more than 1 million patterns
  - Pattern with length of >12
  - Patterns with frequency below 10
- Ground truth
  - Extracted from freebase facts
  - 50 pairs for each relation
  - Pairs are selected in a way including high and low number of patterns

Per-Domain	Ranked First	Ranked Second	Ranked Third	Ranked >3
Location	7	1	3	7
People	6	2	1	1
Organization	3	1	1	-
Miscellaneous	9	4	2	3
All	17	7	6	21

Knowledge Base Augmentation Using Tabular Data

Yoones A. Sekhavat, Denilson Barbosa
University of Alberta

Francesco di Paolo, Paolo Merialdo
Roma Tre University

NSERC CRSNG

IBM Centres for Advanced Studies

yago select knowledge

Freebase

NELL, CMU

Xin Luna Dong, Google

Work in progress (Improved Probabilistic model)

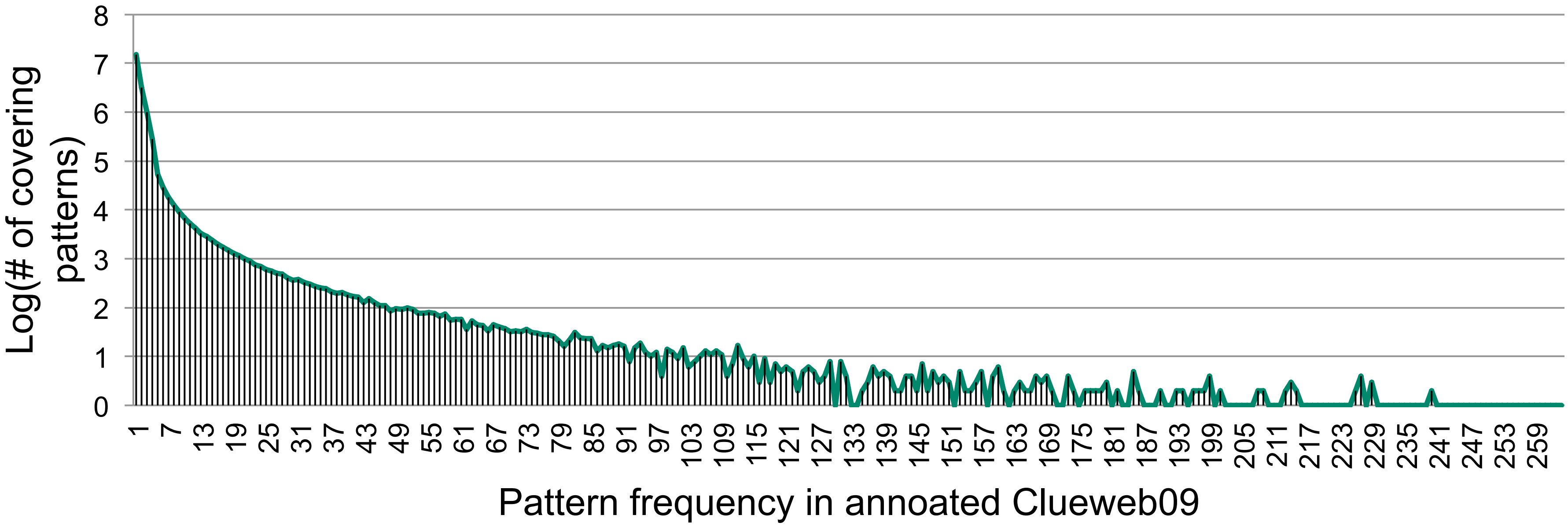
- The types of entities were considered in the model
- We focused on standard NER types: Location, People, Organization and Miscellaneous

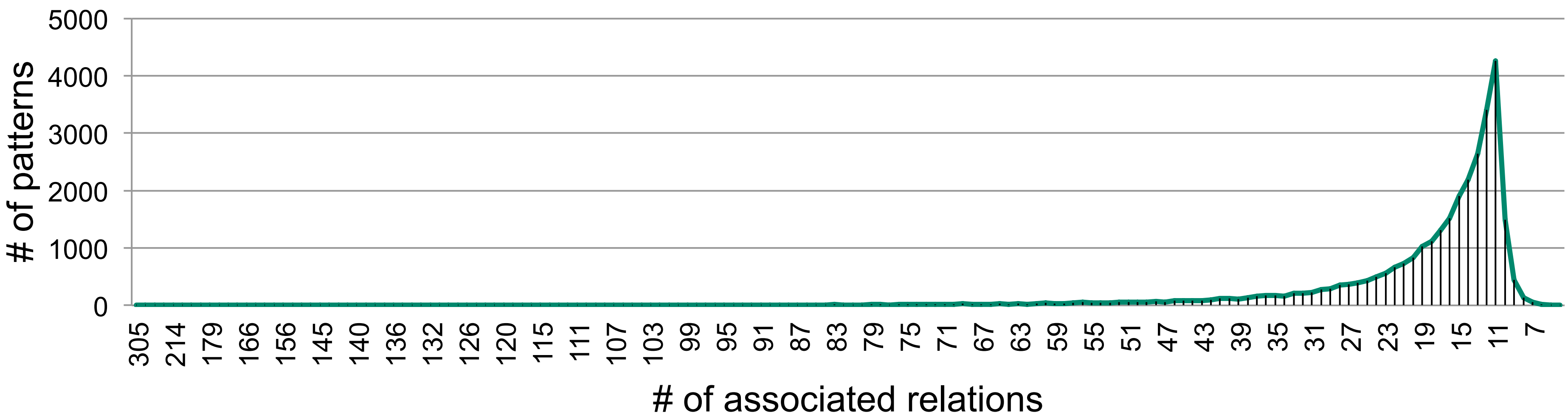
$$Pr(r|p_1, ..., p_k, \langle tx, ty\rangle) = \frac{Pr(r)Pr(p_1, ..., p_k|r)Pr(\langle tx, ty\rangle|r)}{Pr(p_1, ..., p_k)}$$

- We generated quadruples Q = (entity1, pattern, entity2, relation) from annotated clueweb09 using named entities in Freebase
- Improvement in estimating prior probabilities

$$Pr(p|r) = |\{q \in Q|pat(p) \wedge rel(r)\}|/|\{q \in Q|rel(r)\}|)$$

Work in progress (challenges)





Work in progress (challenges)