



Domain Adaptive Relation Extraction for Big Text Data Analytics

Feiyu Xu



Outline



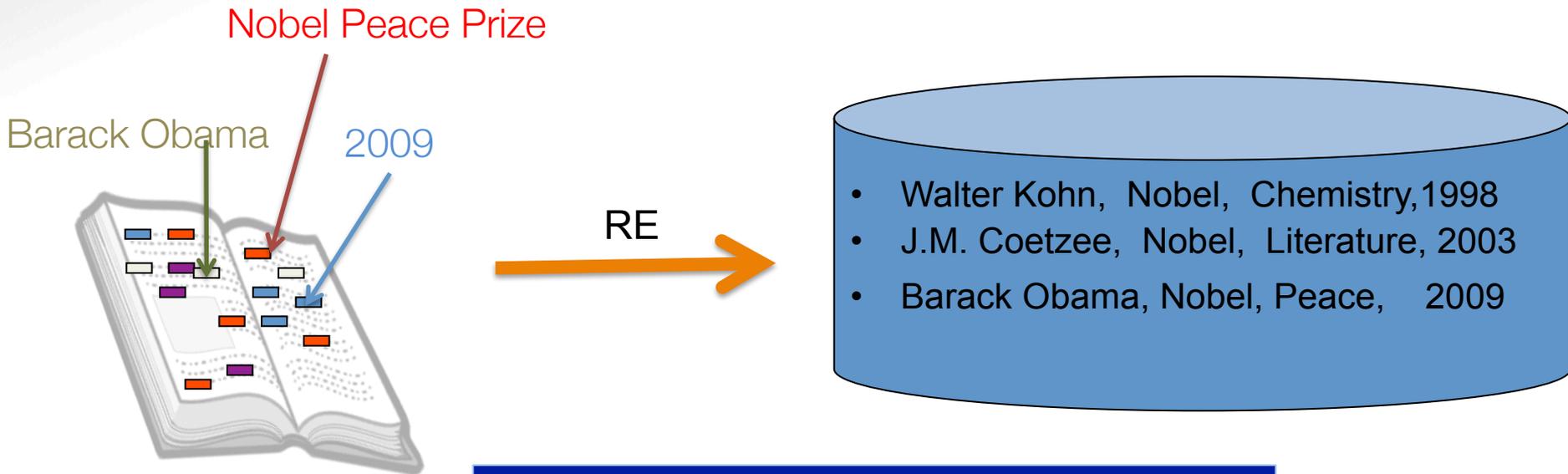
- Introduction to relation extraction and its applications
- Motivation of domain adaptation in big text data analytics
- Solutions
- Conclusion and future work



What is Relation Extraction



- Given an **unstructured** text, a relation extraction (RE) tool should be able to automatically recognize and extract relations among the relevant entities or concepts that are salient to the user's needs



Linguistic Patterns:

- <prize>* be awarded to *<person>*
- <person>* win *<prize>* in *<year>*
-

Example in Opinion Mining



Mitten in der Euro-Krise geht **Altkanzler Helmut Kohl** mit **Angela Merkel** äußerst hart ins Gericht

-- Welt online, 25.08.2011

Opinion Holder

Opinion Target

Polarity



General application task 1:



☆ Information access for information finder

mapping unstructured textual queries of users to more structured formal query for search and answer engines

The screenshot shows a Google search interface. At the top left is the Google logo. To its right is a search bar containing the text "Where is New York". Below the search bar, the word "Search" is displayed in red, followed by the text "About 4,460,000,000 results (0.28 seconds)". On the left side, there is a vertical menu with the following items: "Everything" (highlighted in red), "Images", "Maps", "Videos", "News", "Shopping", and "More". The main content area features a map of New York City and its surrounding areas, with labels for various locations such as Paterson, Yonkers, White Plains, Newark, Rochelle, New York, Elizabeth, and Long Beach. To the right of the map is a link that says "New York, NY" in blue, with "maps.google.com" in green below it. Below the map and link, there is a list of search results. The first result is a list of links: "Hotels - Restaurants - Empire State Building - Top of the Rock Observation Deck - Metropolitan Museum of Art - Statue of Liberty - Rockefeller center - Nyc & Company". The second result is "New York City - Wikipedia, the free encyclopedia" in blue, with a magnifying glass icon to its right. Below this link is the URL "en.wikipedia.org/wiki/New_York_City" in green, followed by "- Cached". The third result is a snippet of text: "New York is the most populous city in the United States and the center of the New York Metropolitan Area, one of the most populous metropolitan areas in the ...". Below this snippet is another list of links: "Demographics - Borough - Neighborhoods - List of tallest buildings in New York City".

General application task 2:



☆ Information acquisition for information provider

extract structured information from big amount free texts to construct knowledge bases

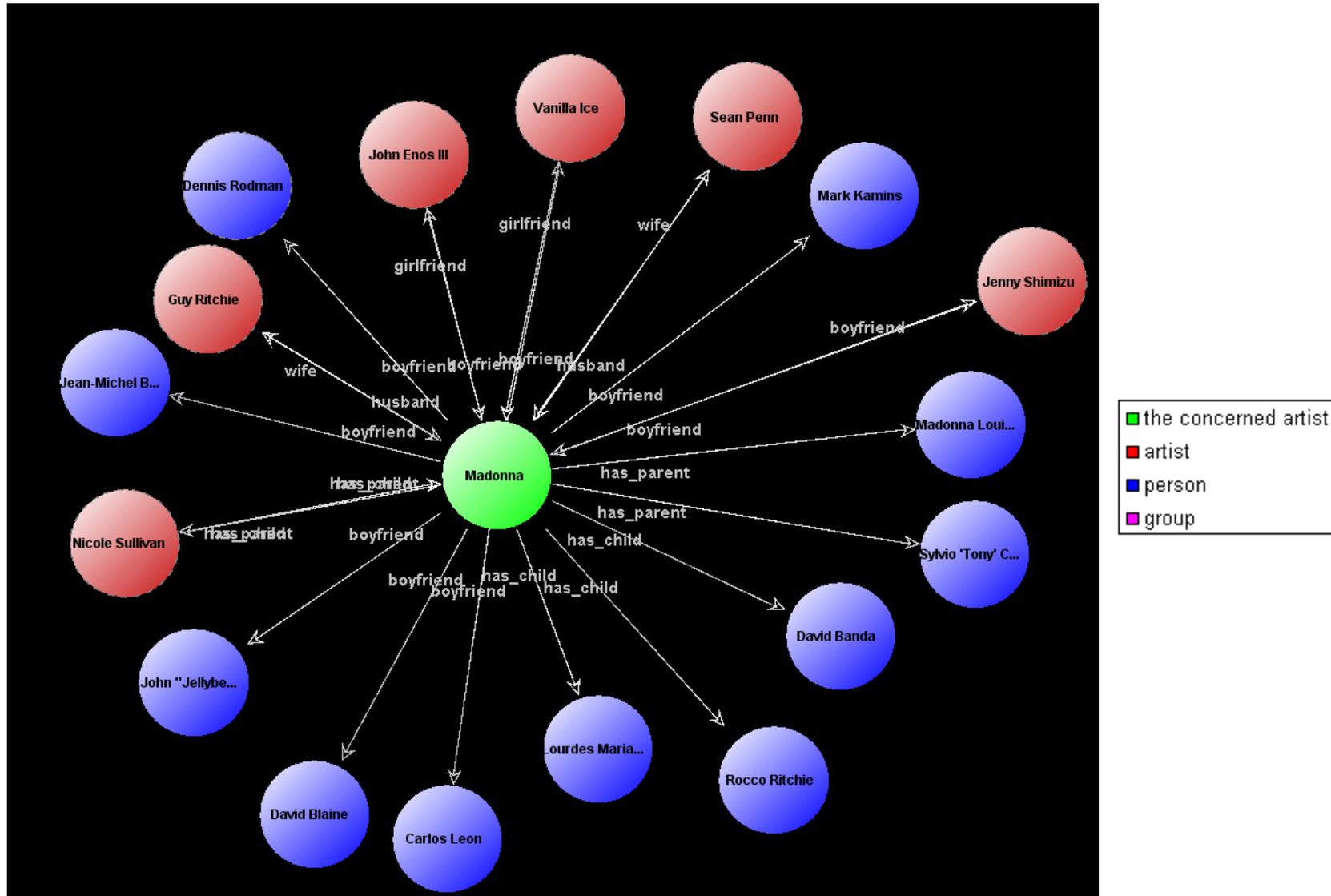


RE

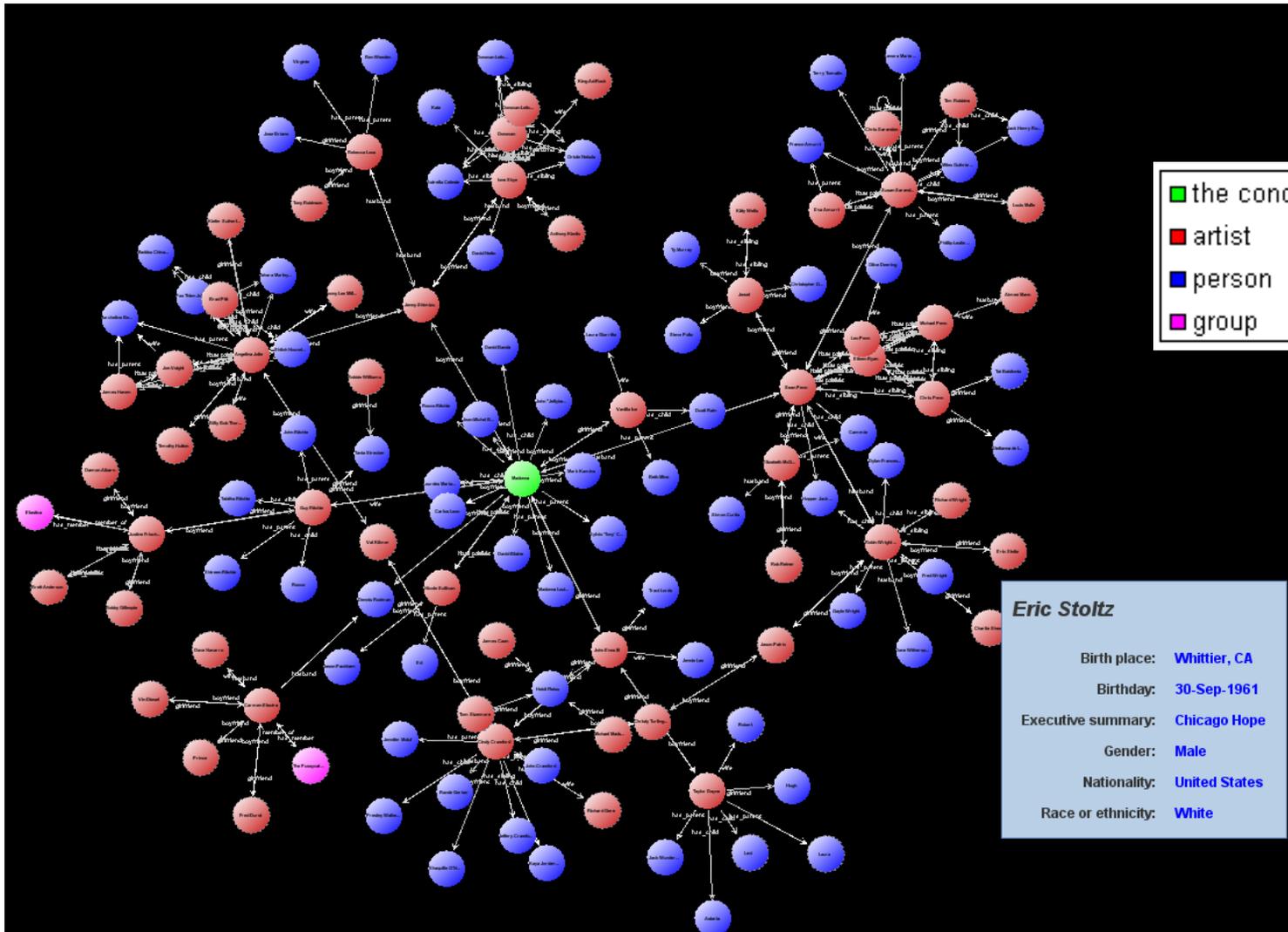


Acquisition of Social Network of Pop Stars from Web

Social Network of "Madonna" (Depth = 1)



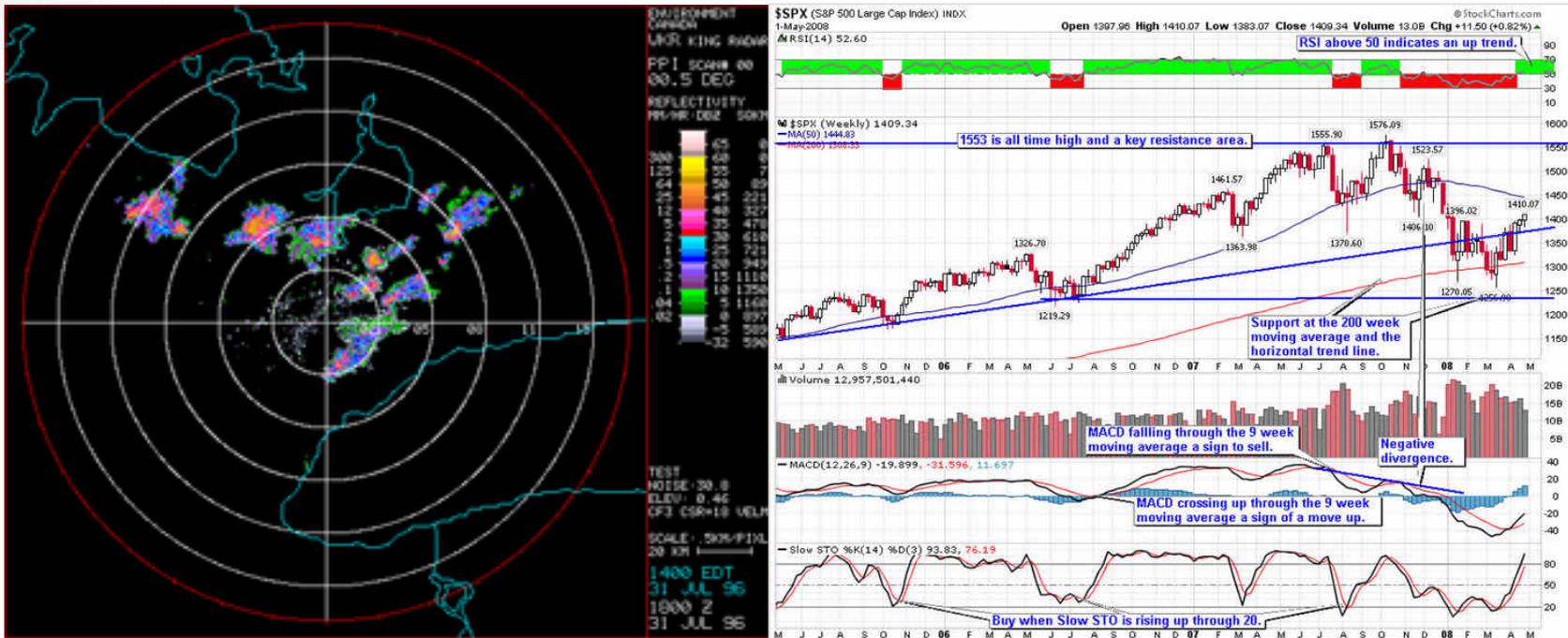
Social Network of "Madonna" (Depth = 3)



General application task 3: Big Data Analytics



- Enabling the linking between structured and unstructured data
 - Large-scale information monitoring
 - Analytics: analyses of areas, markets, trends
 - Watch: Scanning for relevant new developments



Example: Network of Innovation Keyplayers



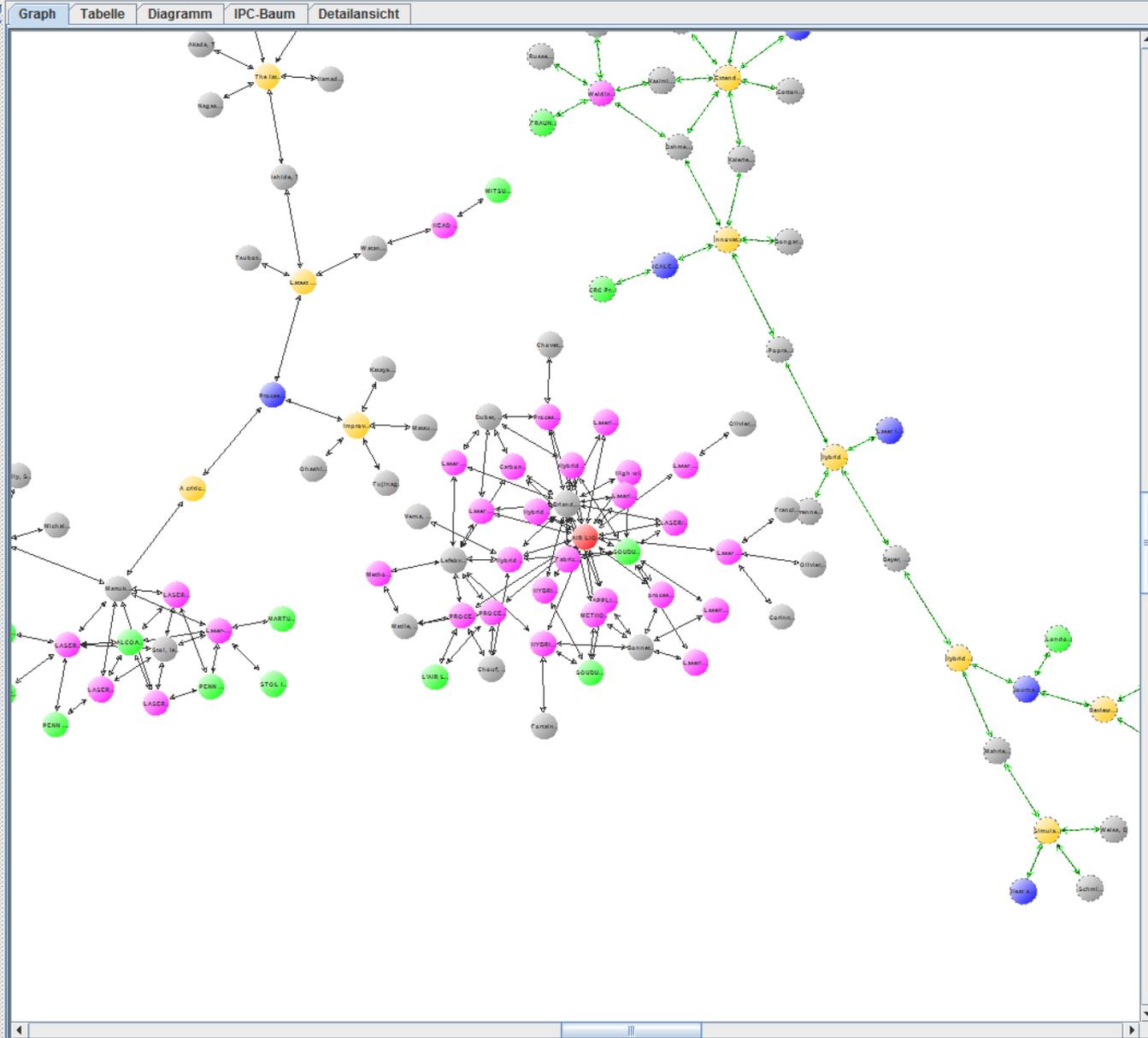
http://techwatchtool.dfki.de/thyssenkruppsteel/concept?queryId=33&queryName=hybrid+beam+welding&command=show

Ergebnisse für "hybrid beam v

- Patent (62)
- Publikation (43)
- Person (207)
- Organisation (52)
- AIR LIQUIDE
- ALCOA INC
- ALCOA INC.
- AMERICAN WELDING SOC
- AT-FACHVERLAG
- Aedermannsdorf, Switzerland: Tran
- American Welding Society
- BOEING CO

AIR LIQUIDE

- All Patents in Database:**
- [METHOD FOR LASER WELDING USING A NOZZLE CAPABLE OF STABILISING THE KEYHOLE](#) (2009-09-03)
 - [Laser Beam Welding Method with a Metal Vapour Capillary Formation Control](#) (2009-05-28)
 - [Laser beam welding method with a metal vapour capillary formation control](#) (2008-10-22)
 - [Process for laser-ARC hybrid welding aluminized metal workpieces](#) (2008-01-17)
 - [Laser arc hybrid welding method for surface coated metal parts, the surface coating containing aluminium](#) (2008-01-16)
 - [Laser/MIG hybrid welding process with a high wire speed](#) (2007-10-30)
 - [Laser or hybrid arc-laser welding with formation of a plasma on the back side.](#) (2006-11-29)
 - [LASER/MIG HYBRID WELDING PROCESS WITH A HIGH WIRE SPEED](#) (2006-05-19)
 - [Hybrid laser-Metal in Gas welding with a elevated welding and filler wire supply speeds and a high](#)



Legende

- Auswahl
- Patent
- Publikation
- Person
- Organisation
- Journal

Ansicht

- Vergrößern (+)
- Verkleinern (-)
- Normalgröße
- Größe an Fenster a...

Graph anpassen

- Fast Organic
- Circle

Outline



- Introduction to relation extraction and its applications
- Motivation of domain adaptation in big text data analytics
- Solutions
- Conclusion and future work



Text Analytics for Big Textual Data



- Three main features of big data
 - *Volume*: large-scale in volume
 - *Variety*: with respect to heterogeneous domains and formats
 - *Velocity*: because of its rapid and steady growing.

- Requirements of text analytics technologies for big data
 - *efficient*
 - *robust*
 - *scalable*
 - *domain-adaptive*



Domain Adaptation is Essential for Big Data!



- Among the three big data features, **variety** and **velocity** are even more challenging than the sheer size **volume**



Reasons:



- New domains have been constantly emerging, rapidly growing in size.
- Domains can differ in
 - **topics** (e.g., medicine, chemistry or mechanics)
 - **genres** (e.g., news, novels, blogs, scientific publications or patents)
 - **targets** (e.g., different relations such as marriage, person-parent relation, disease-symptom relation)
 - **data internal properties** (e.g., size or redundancy or connectivity).
- Systems, methods or strategies developed or trained for so-called general purpose or one specific domain can often not be directly taken over by other domains, because
 - each domain needs its own domain knowledge and
 - each application data has its own special properties.



Relevant Strategies for Domain Adaptation



- Minimally dependent on the labeled training data
 - Minimally or weakly supervised machine learning methods
- Strategies for
 - confidence estimation of automatically learned information and knowledge
 - filtering of irrelevant and wrong information
- Domain adaptation of generic systems for specific applications



Outline



- Introduction to relation extraction and its applications
- Motivation of domain adaptation in big text data analytics
- **Solutions**
- Conclusion and future work



Our solutions (1)



- **minimally supervised and distantly supervised automatic learning of domain-specific grammar-based pattern rules for n-ary RE: DARE and Web-DARE Systems**
 - Feiyu Xu, Hans Uszkoreit, Hong Li, “A Seed-driven Bottom-up Machine Learning Framework for Extracting Relations of Various Complexity (2007)”. In ACL 2007.
 - Hans Uszkoreit, Feiyu Xu, Hong Li. “Analysis and Improvement of Minimally Supervised Machine Learning for Relation Extraction”. In NLDB 2009.
 - Sebastian Krause, Hong Li, Hans Uszkoreit, Feiyu Xu, “Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web”. In Proceedings of the 11th International Semantic Web Conference (ISWC 2012).



Our solutions (2)



- **Various filtering and confidence estimation methods for high-performance and large-scale relation extraction**
 - Sebastian Krause, Hong Li, Hans Uszkoreit, Feiyu Xu, “Large-Scale Learning of Relation-Extraction Rules with Distant Supervision from the Web”. In Proceedings of the 11th International Semantic Web Conference (ISWC 2012)
 - Andrea Moro, Hong Li, Sebastian Krause, Feiyu Xu, Roberto Navigli, Hans Uszkoreit, “Semantic rule filtering for web-scale relation extraction”. In Proceeding of International Semantic Web Conference (ISWC 2013).
 - Feiyu Xu, Hans Uszkoreit, Sebastian Krause, Hong Li. Boosting Relation Extraction with Limited Closed-World Knowledge. COLING 2009, Poster.



Our solutions (3)



- **Automatic adaptation and improvement of generic parsing results for specific domains**
 - Peter Adolphs, Feiyu Xu, Hans Uszkoreit, Hong Li, “Dependency Graphs as a Generic Interface between Parsers and Relation Extraction Rule Learning”. In Proceedings of KI 2011, pp. 50-62, 2011.
 - Feiyu Xu, Hong Li, Yi Zhang, Hans Uszkoreit, Sebastian Krause, “Parse reranking for domain-adaptive relation extraction”. Journal of Logic and Computation, doi: 10.1093/logcom/exs055, Oxford University Press, 2012.



Our solutions (4)



- **Automatic generation of domain-specific linguistic knowledge resources**
 - Hans Uszkoreit and Feiyu Xu, “From Strings to Things, SAR-Graphs: A New Type of Resource for Connecting Knowledge and Language”. In Proceedings of 1st International Workshop on NLP and DBpedia volume 1064, Sydney, NSW, Australia, CEUR Workshop Proceedings, 10/2013
 - **Open source: sargraph.dfki.de**





Web-DARE

Distant-supervised Web-scale RE



Web-DARE: Distant Supervision based RE



- Large number of RE rules are automatically learned by using Freebase as seed knowledge and Web as training corpus
- Goal:
 - covering most linguistic variants for expressing a relation
 - thus solving the notorious long-tail problem of real-world NL applications



Data Set



- rules learned for 39 relations
 - n-ary relations $n \geq 2$
- three domains: business, awards and people
- 2.8 million relation instances retrieved from Freebase as seed
- 20 million web documents as training corpus



Example in Nobel Prize Award Domain



- Seed example

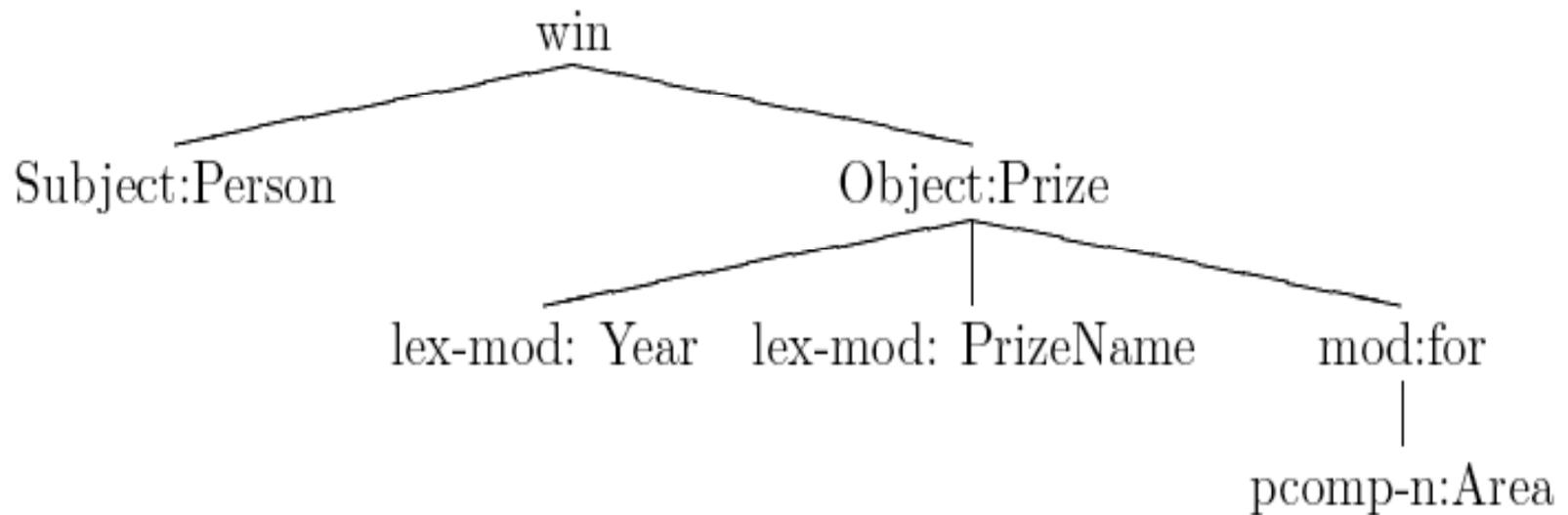
<Mohamed ElBaradei/Person, Nobel/Prize, Peace/Area, 2005/Year>

- Sentence matched with the seed

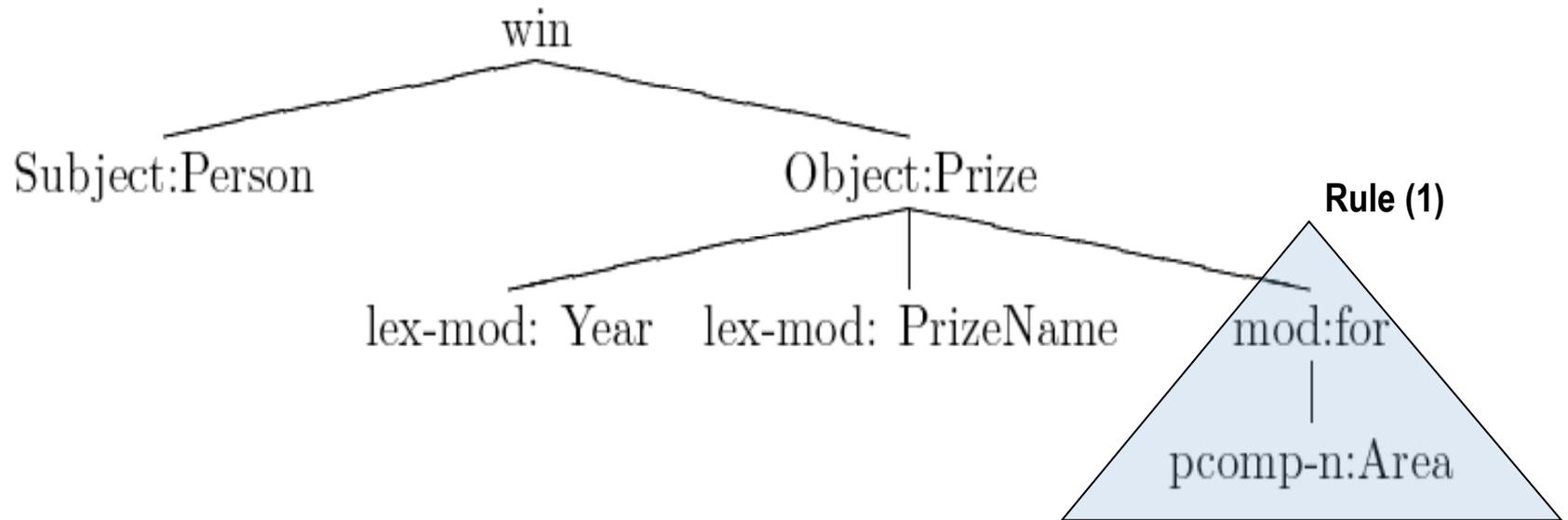
Mohamed ElBaradei won the 2005 Nobel Prize for Peace on Friday ...



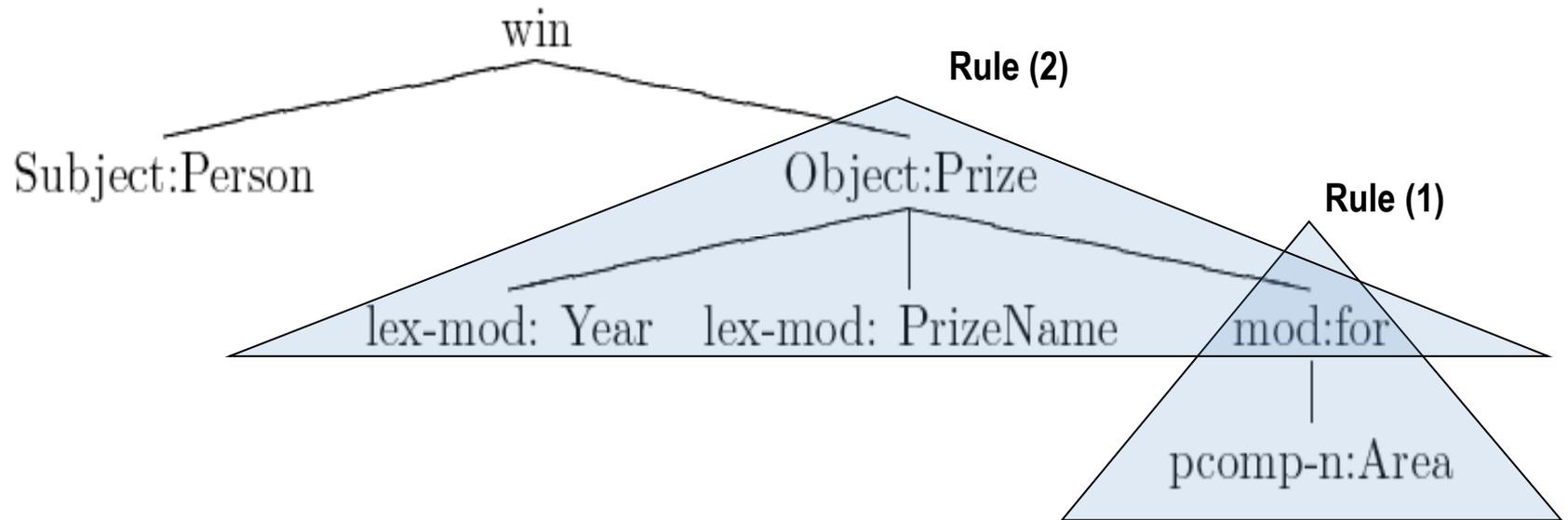
Dependency Parse Result



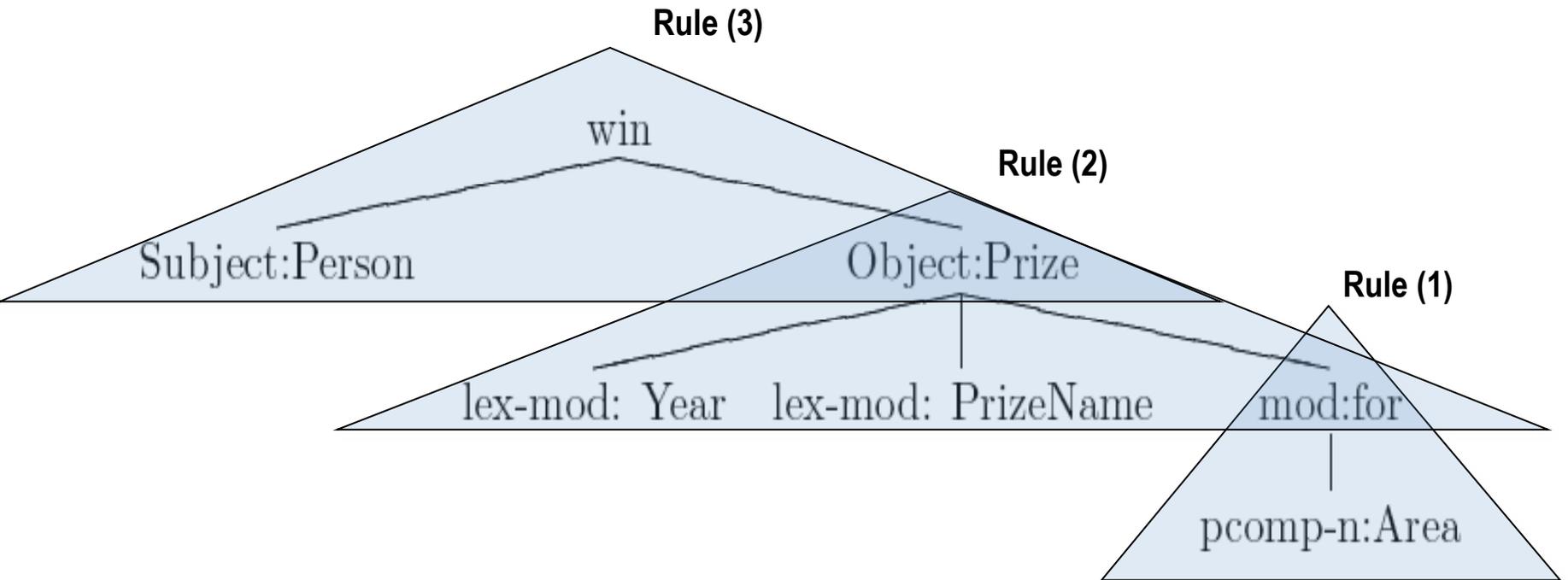
Bottom Up Rule Learning



Bottom Up Rule Learning



Bottom Up Rule Learning



Web-DARE Architecture

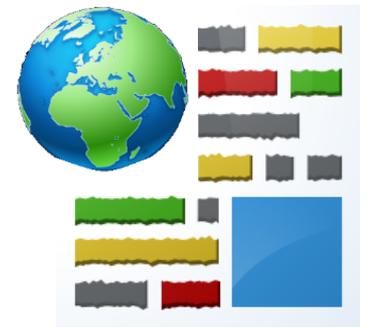


Google
bing™

Queries

Facts

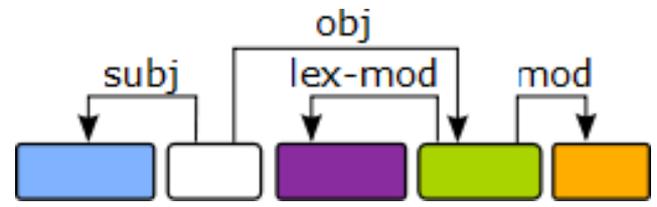
Web pages



Filtered Patterns

Sentence Mentions of Facts

Patterns



Some Statistics of Web-DARE Rules



Relation	# Sentences used	# Sentences w/ a learned rule	# Rules	# Rules w/o duplicates
<i>award nomination</i>	13,966	13,149	23,987	7,800
<i>award honor</i>	50,550	49,001	106,550	40,578
<i>hall of fame induction</i>	31,244	28,278	44,920	17,450
<i>organization relationship</i>	46,331	42,824	60,379	28,903
<i>acquisition</i>	63,967	60,903	96,747	50,544
<i>organization merger</i>	2,996	1,521	3,243	1,758
<i>company name change</i>	9,433	9,132	15,619	6,910
<i>spin off</i>	5,247	5,094	8,319	4,798
<i>marriage</i>	342,895	335,313	557,478	176,949
<i>sibling relationship</i>	167,611	160,893	255,788	69,596
<i>romantic relationship</i>	155,335	152,878	229,393	74,895
<i>person parent</i>	192,610	186,834	390,878	119,238
average of 39 relations	66,545	66,509	109,435	41,620



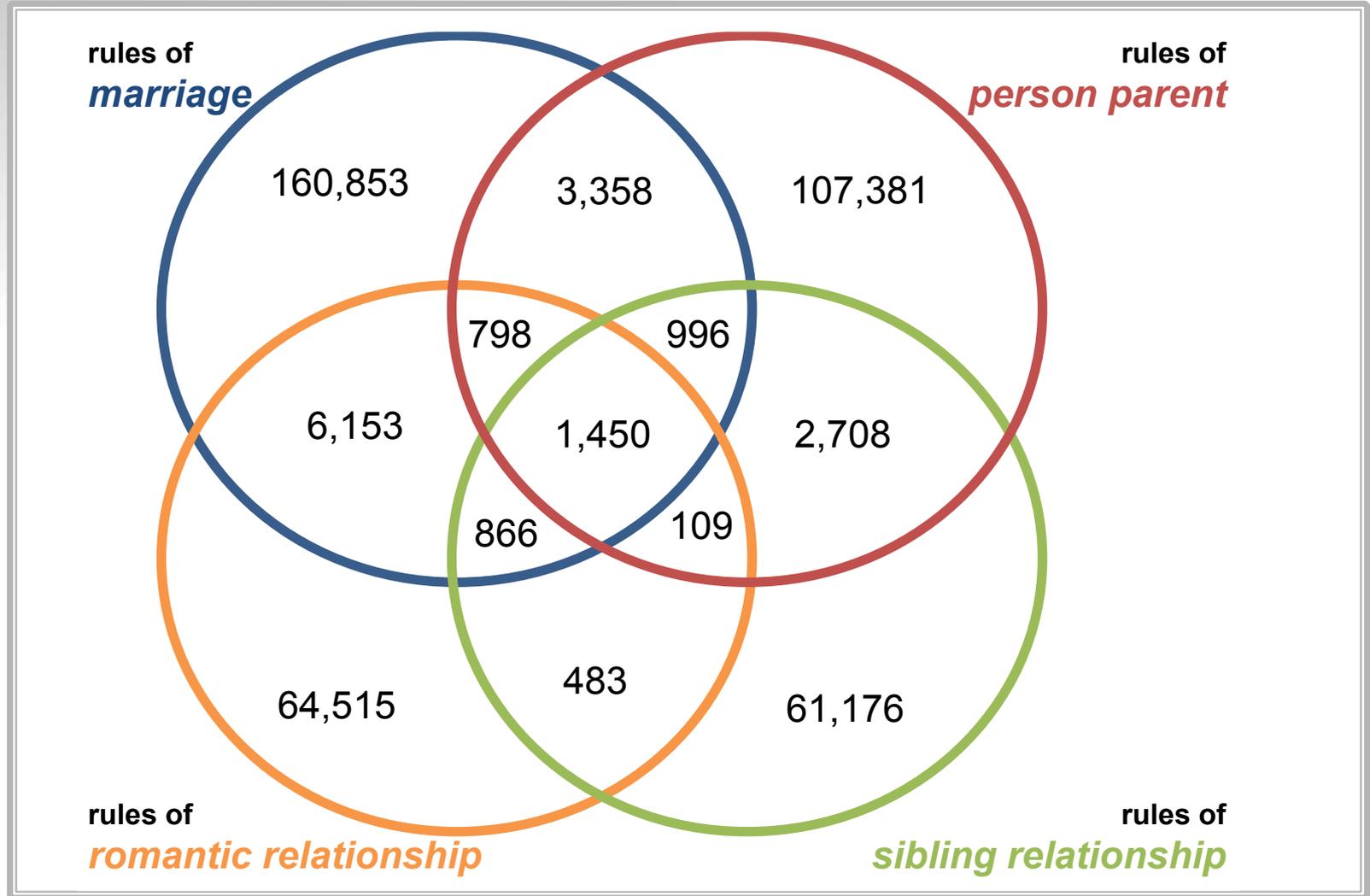
Problems of Large-Scale Approach



- Very low precision
 - a lot of noisy rules
 - many rules are learned from more than one relation



Euler Diagram for Four People-Relations





Various Filtering Strategies for High-Performance Web-Scale RE



Frequency-Driven Rule Filters

- Merged Filter:

$$valid_m^{\mathcal{R}}(r) = valid_{freq}^{\mathcal{R}}(r) \wedge valid_{inter}^{\mathcal{R}}(r)$$

- 1) **absolute frequency filtering**: a threshold to exclude rules with low occurrency



Rule Frequency Driven Filters

- Merged Filter:

$$valid_m^{\mathcal{R}}(r) = valid_{freq}^{\mathcal{R}}(r) \wedge valid_{inter}^{\mathcal{R}}(r)$$

- 1) **absolute frequency filtering**: a threshold to exclude rules with low occurrence
- 2) **inter-relation filter (Overlap Filter – FO Filter)**:
 - based on mutual exclusiveness of relations with similar entity-type signatures.
 - a rule is only valid for a relation, if its relative frequency is higher than any other relations with similar entity type signatures.

$$valid_{inter}^{\mathcal{R}}(r) = \begin{cases} true & \text{if } \forall \mathcal{R}' \in \mathbb{R} \setminus \{\mathcal{R}\} : rf_{r, \mathcal{R}} > rf_{r, \mathcal{R}'} \\ false & \text{otherwise} \end{cases}$$



Weakness of Filtering with Frequency



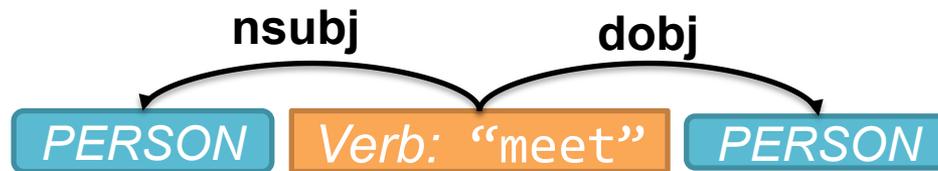
- Undetected low-quality patterns:
 - high frequency in target relation, low frequency in coupled relations



Weakness of Filtering with Rule Frequency



- Undetected low-quality patterns:
 - high frequency in target relation, low frequency in coupled relations



X ?



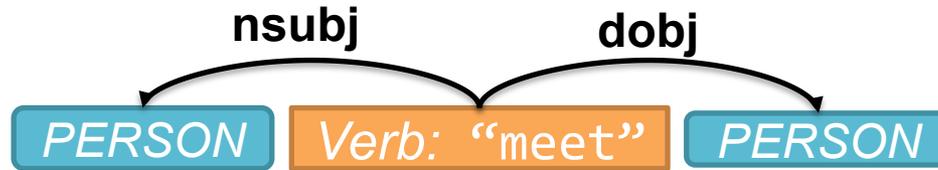
✓ ?



Weakness of Filtering with Rule Frequency



- Undetected low-quality patterns:
 - high frequency in target relation, low frequency in coupled relations



X ?

- Erroneously-deleted good patterns:
 - infrequent patterns



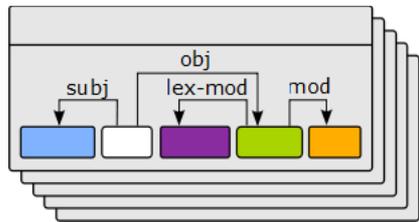
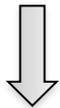
✓ ?



Lexical Semantics can help!



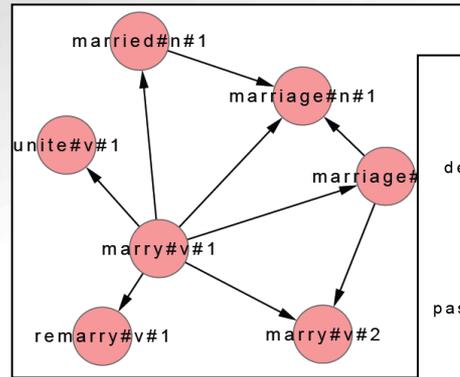
World Wide Web



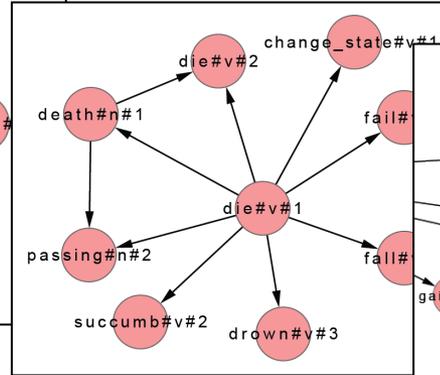
✓ ? ✗ ?

Candidate RE Patterns

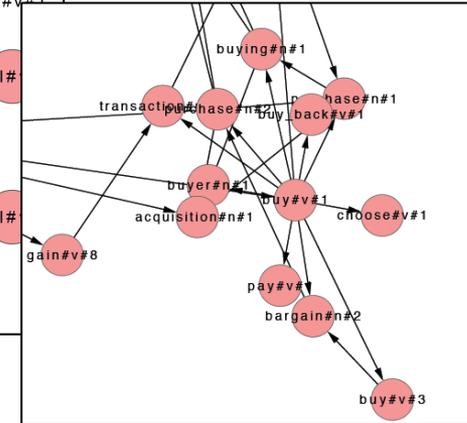
marriage



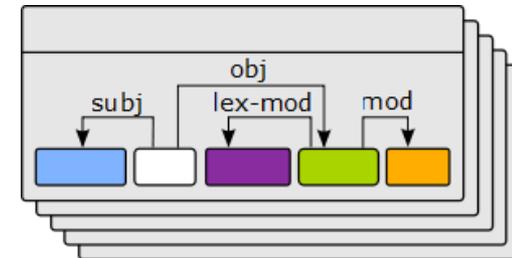
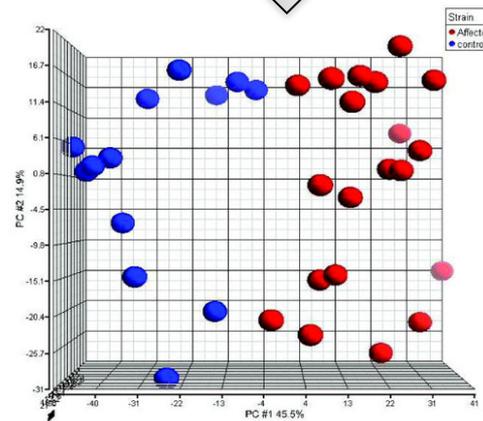
person-death



acquisition



Relation-specific lexical semantic graphs



✓ !!!

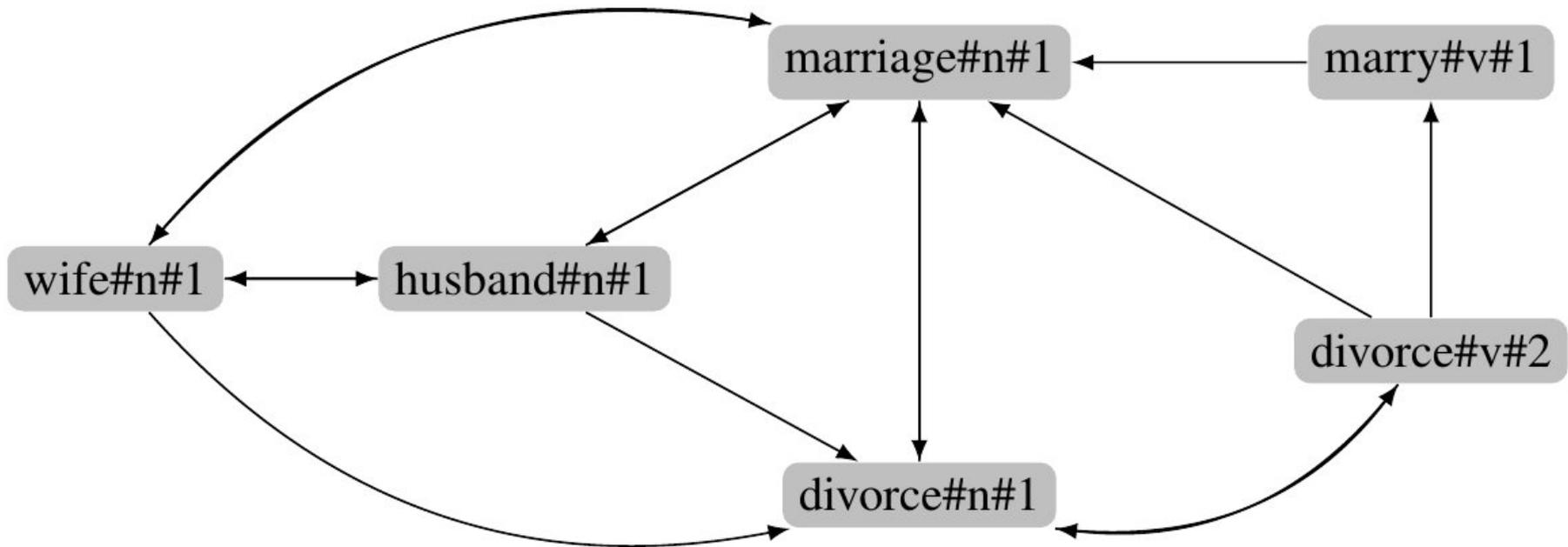
Unsupervised Classification

High-quality RE Patterns

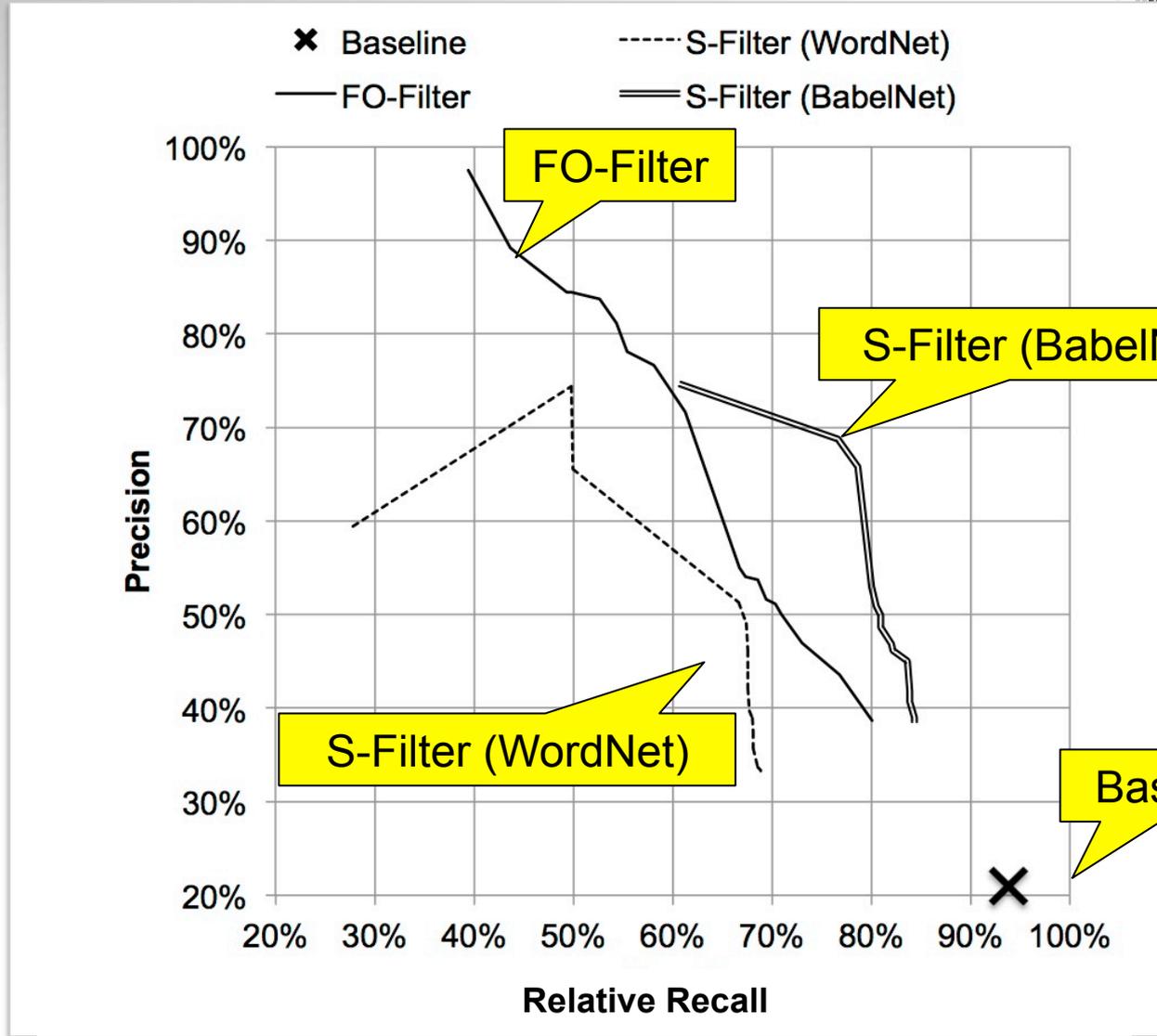
The Relation-Specific Semantic Graph



An excerpt of the semantic graph for the relation *marriage*



Extrinsic Eval. – Web-DARE





Parse-Reranking for Domain-adaptive RE



Error Types of Extracted Wrong Instances



Content	Modality	Named Entity Recognition (NER)	Parsing	NER & Parsing	DARE Rules
11.8%	17.6%	5.9%	38.2%	11.8%	14.7%

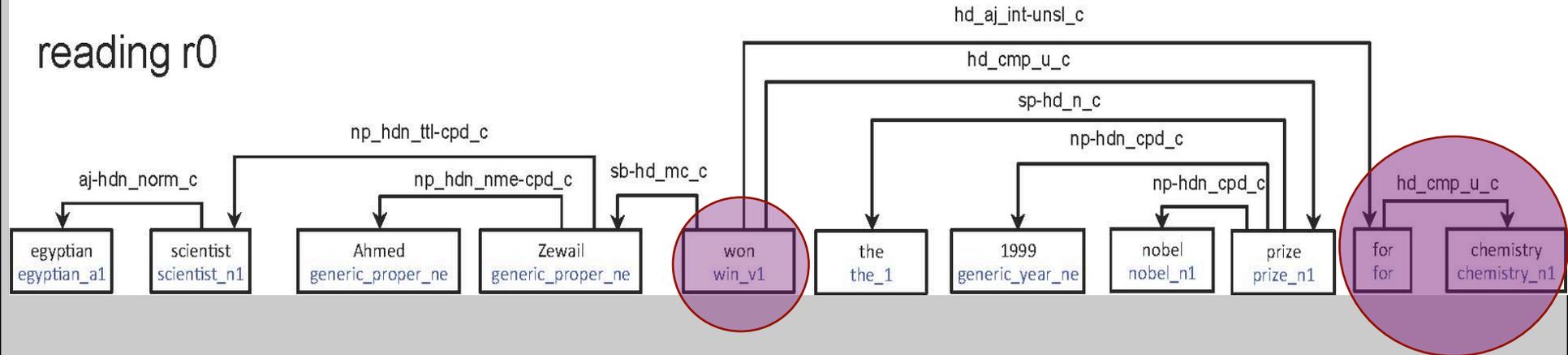


HPSG Parses: PP Attachment

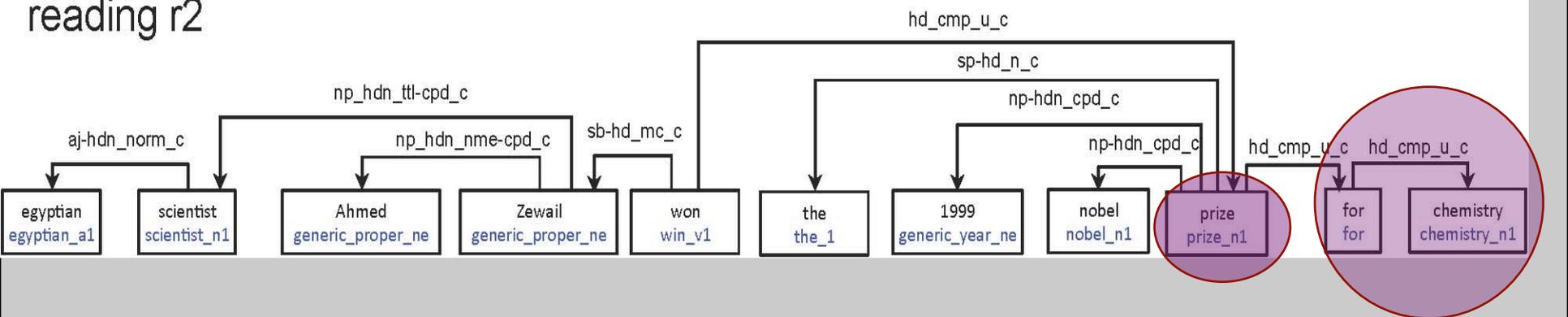


*Egyptian scientist Ahmed Zewail won the 1999 Nobel Prize **for chemistry***

reading r0



reading r2

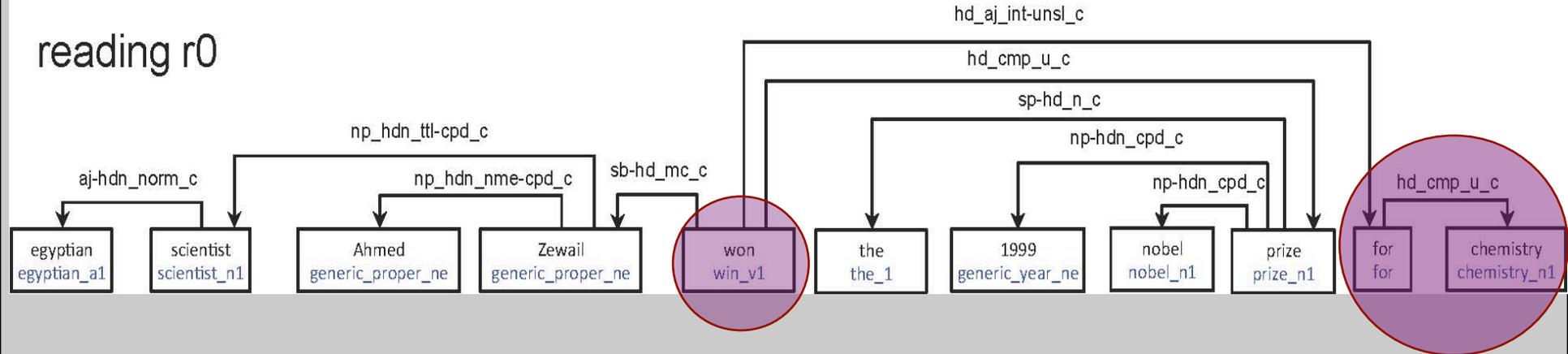


HPSG Parses: PP Attachment

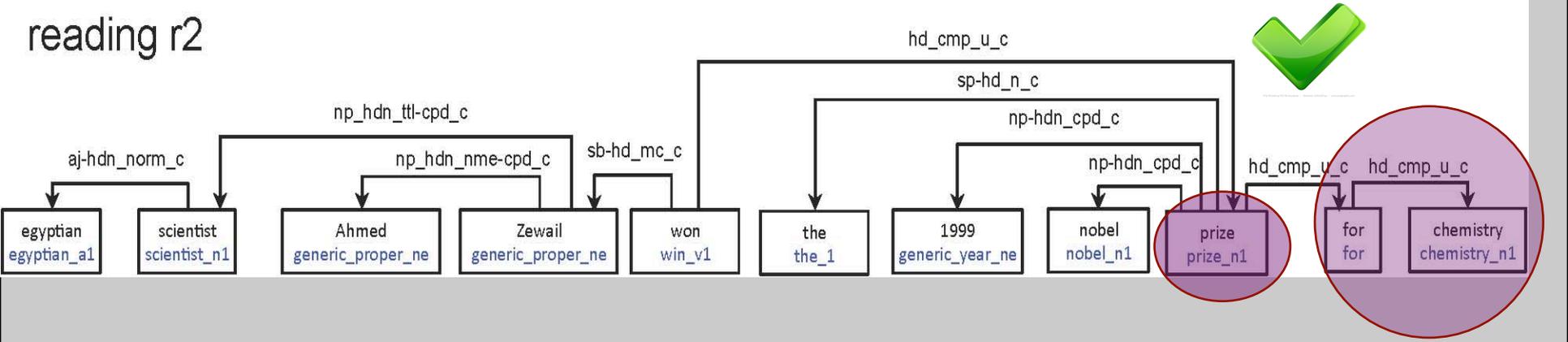


Egyptian scientist Ahmed Zewail won the 1999 Nobel Prize for chemistry

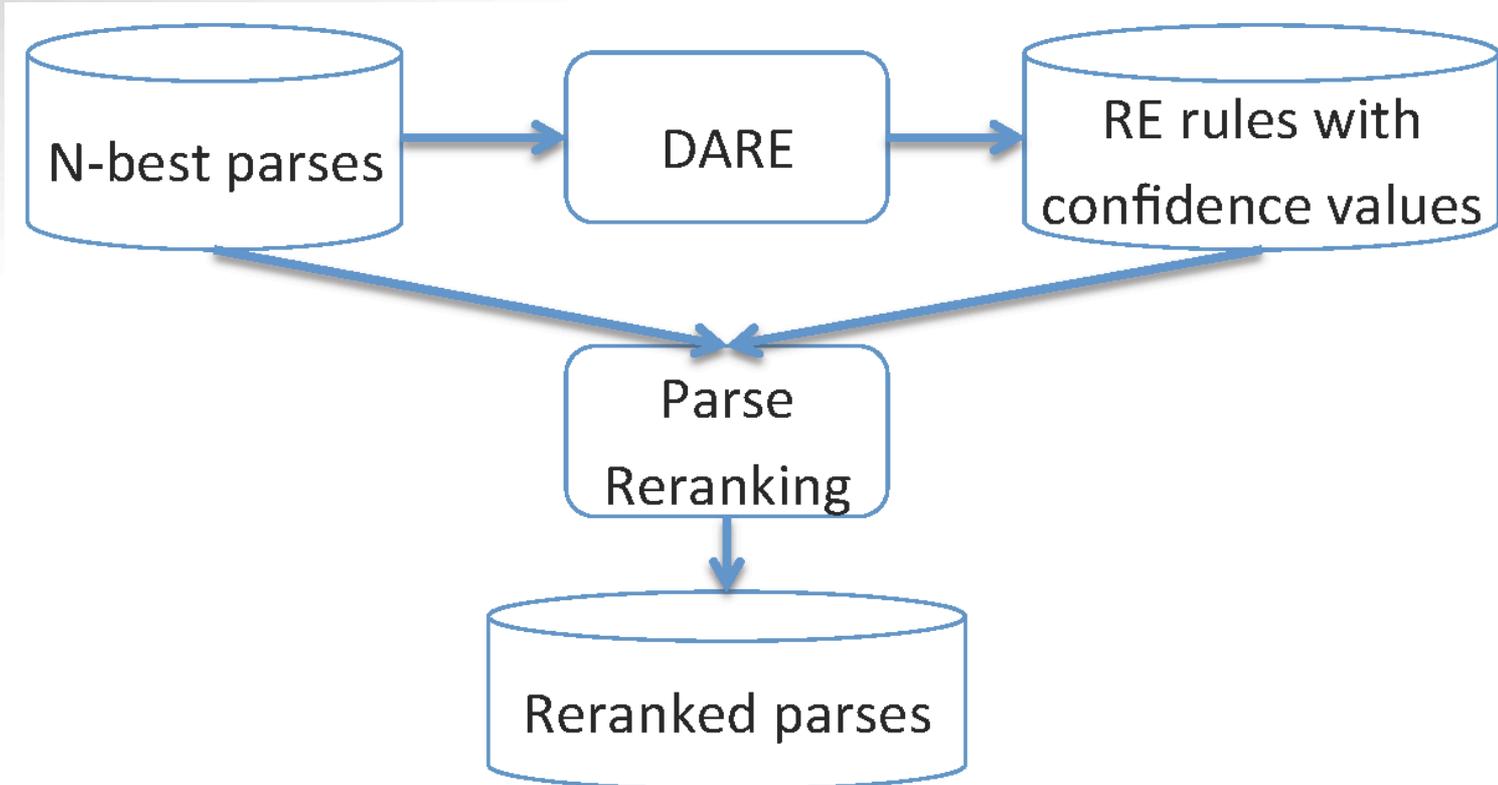
reading r0



reading r2



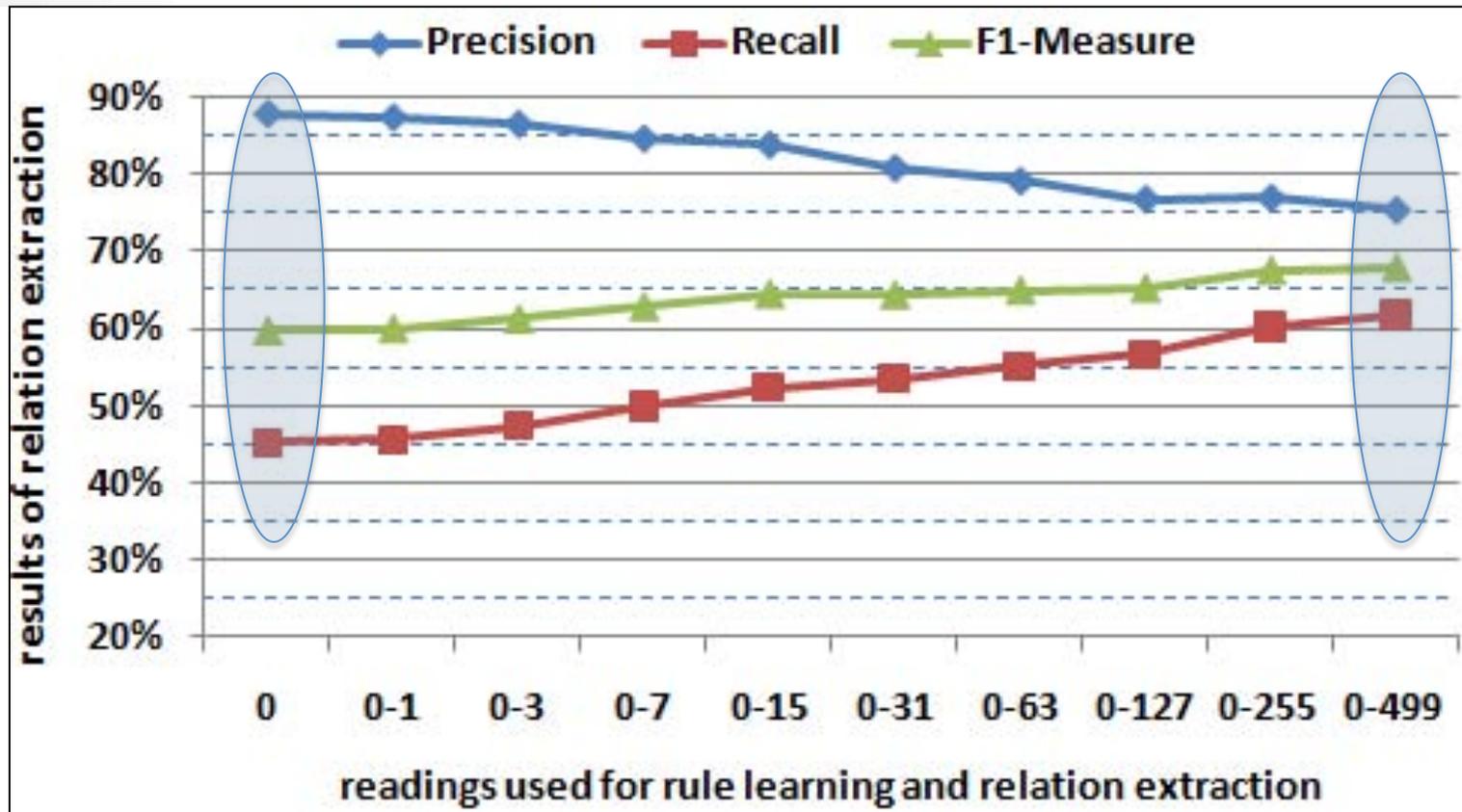
Reranking Architecture



Baseline: before Re-ranking



- **Best reading:** high precision, low recall, low F-measure
- **500 readings:** lower precision, higher recall, higher F-measure



After Re-Ranking:



- Re-ranked top readings match more sentence mentions containing RE instances
- Improvements of Recall and F-Measure



Conclusion



- The performance of large-scale RE for each application is dependent on the performance of domain-adaptation methods
- Three original contributions (among others):
 - Extension of relation extraction to n-ary relations
 - Semantic filtering with large lexical knowledge bases
 - Parser improvement for the specific RE task by reranking
- For our work we received a Google Focused Research Award



Google
Focused Research Awards

Planned Future Work



- Immediate next step of big text data analytics is to integrate the existing NLP and IE components into big data analytics platforms
- Entity linking and RE will play an essential role for semantic interoperability between structured and unstructured data
- Extension and Application of our IE technologies to the new Smart Data projects
 - Smart Data Web: Industry 4.0
 - Smart Data for Mobility: Mobility

