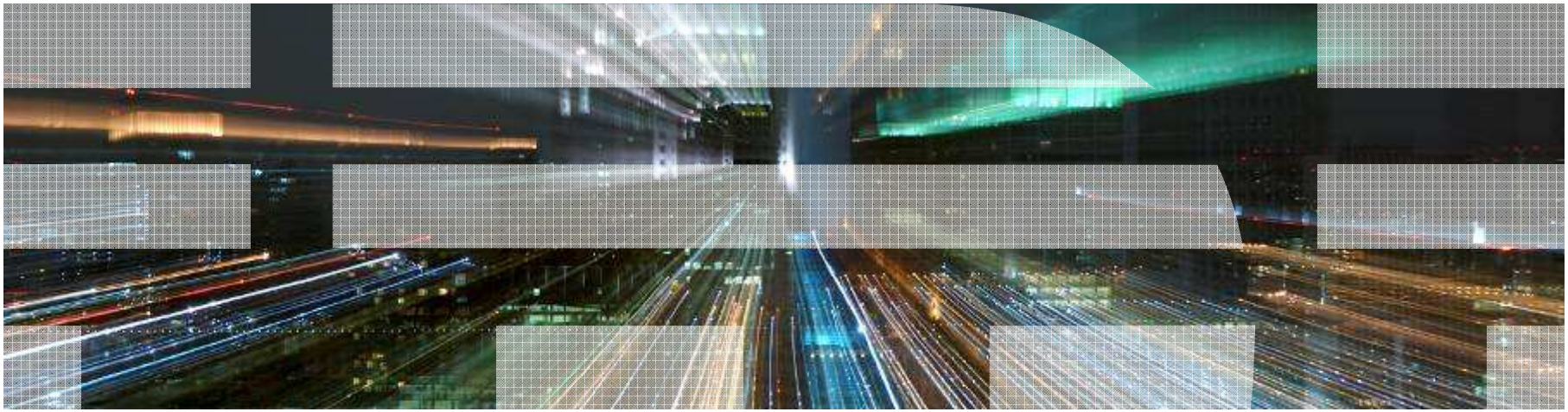
Relation Extraction with Relation Topics

- Information Fusion in Natural Language Processing

Chang Wang

DeepQA Team @ IBM Research

(Joint work with James Fan, Aditya Kalyanpur, Branimir Boguraev, David Gondek)



Information Fusion in Natural Language Processing

- DeepQA (Watson) has >100 components.

- **TWREX**: **T**opicalized **W**ide **R**elation and **E**ntity e**X**traction

Statistical Relation Extraction Module

Outline

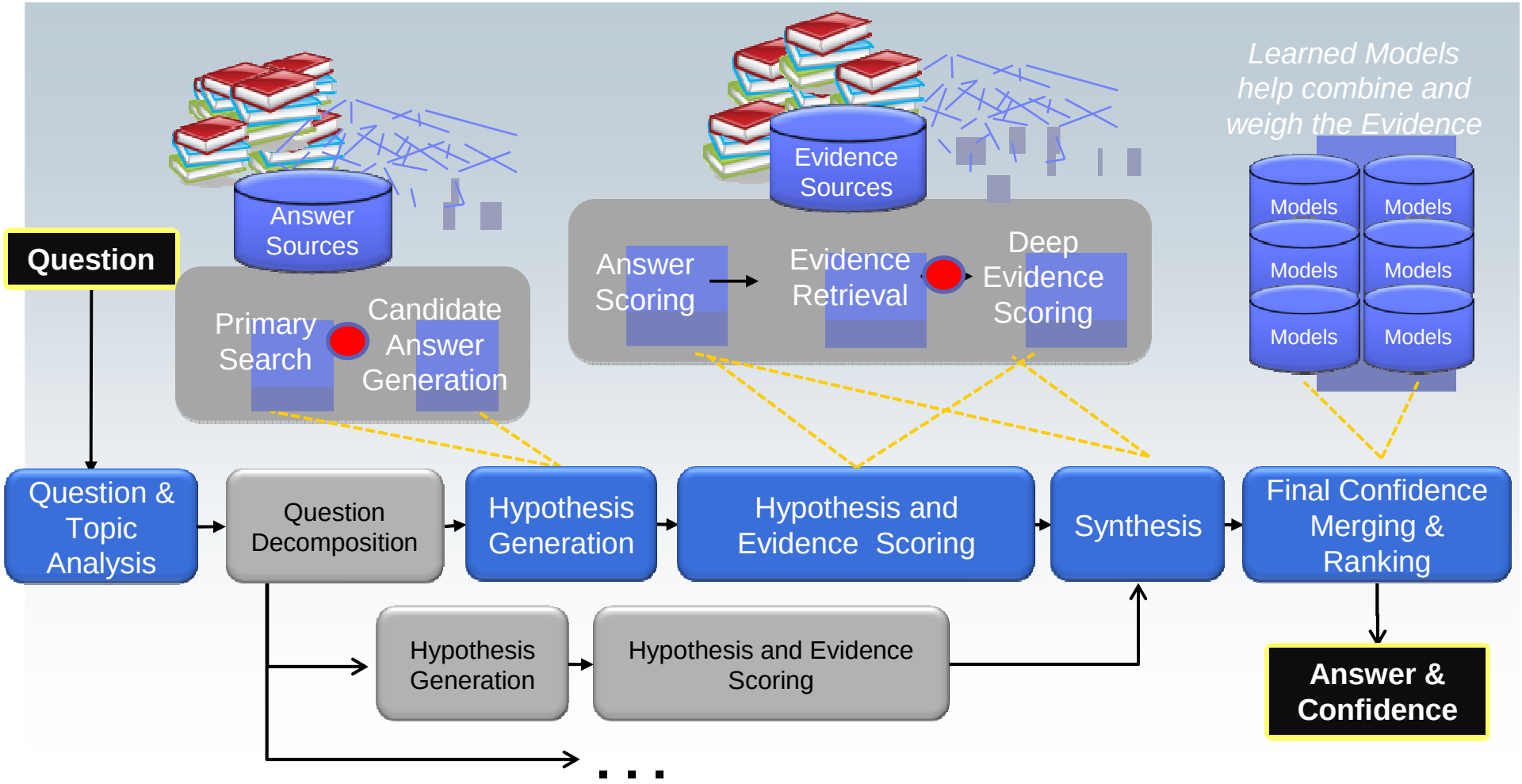
- Background:
  - Relation Detection
  - Watson Pipeline
  - Relation Detection in Watson
- TWREX Architecture
  - Challenges
  - Construction of Relation Repository
  - Relation Topics
  - Integration (Information Fusion)
- Experiments & Conclusions

Relation Extraction

- Relation extraction: to classify the relation between two entity mentions into one of predefined relation classes `locatedAt?` `customerOf?` `employedBy?`
- Example:
  - "***The New Jersey Devils*** have signed ***Adam Larsson*** to a three-year, entry-level contract"
- Applications:
  - Information extraction
  - Machine reading
  - Question answering
  - Etc … …
- Challenges
  - Expressiveness of language:
    - IBM hired James, James started at IBM, James worked for IBM, …

How Relation Extraction module is used in the DeepQA pipeline?

- On primary search results, before candidate answer generation
- On supporting evidence, before deep evidence scoring



How Relation Extraction is Used to Create Candidate Answers/ Score Passages

- Candidate Answer Generation

  The question

  "The Screwtape Letters" from a senior devil to an under devil are by this man better known for children's books.

  contains an instance of the **"authorof"** relation, whose arguments are identified as *this man* and *"the Screwtape Letters".*

  We can look up potential answers in our structured database based on the relations detected.

- Evidence Scoring

  The question:

  This hockey defenseman ended his career on June 5, 2008.

  and a supporting passage share "**activeyearsenddate**" relation:

  On June 5, 2008, Wesley announced his retirement after his 20th NHL season.

  Assign a high similarity score if the question and the passage share some semantic relations.

Outline

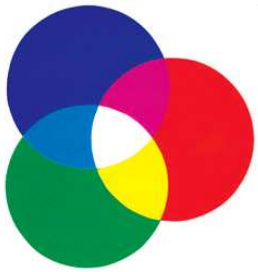
- Background:
  - Relation Detection
  - Watson Pipeline
  - Relation Detection in Watson
- TWREX Architecture
  - Challenges
  - Construction of Relation Repository
  - Relation Topics
  - Integration (Information Fusion)
- Experiments & Conclusions

TWREX: Goal and Challenges

- Goal: reuse the knowledge from the existing domains for the new domains

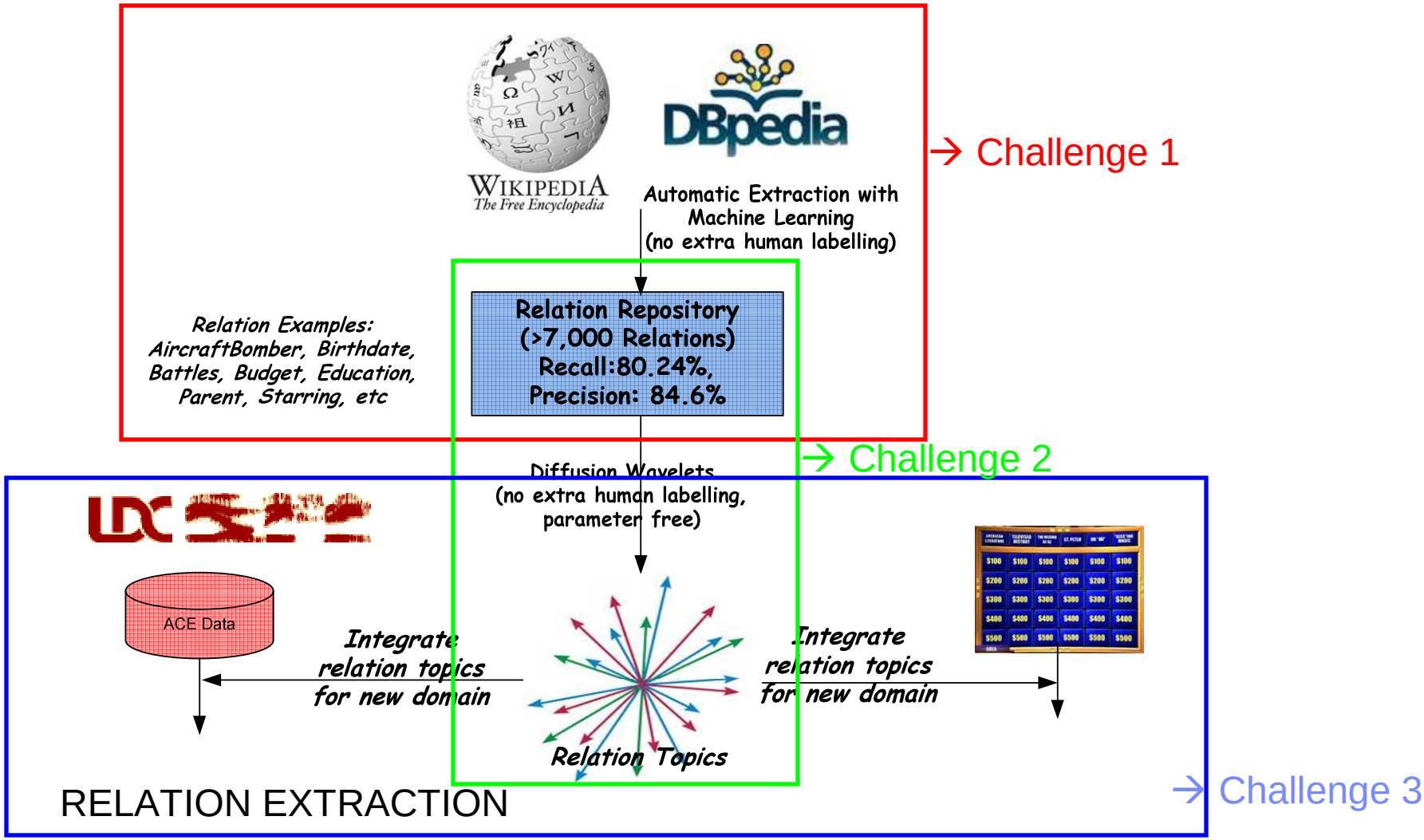
- Challenges:
  - Need to construct a relation repository that has sufficient coverage.

  - Need to redundancy, noisy information in the repository.
    For example, DBpedia has relations like “birthplace”, “placeofbirth”, “hometown”, etc

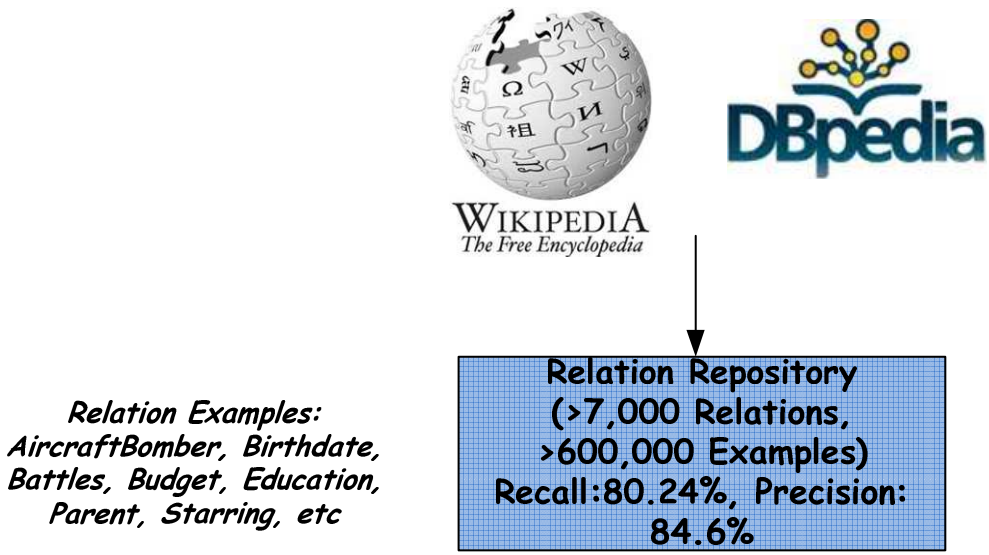


  - Need an efficient way to make use of the knowledge brought in by the other relations.

The Overall Architecture



Step 1: Building Relation Repository

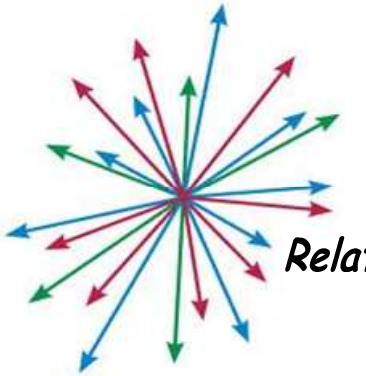


For example, the Wikipedia page for “Albert Einstein” contains an infobox property “alma mater” with value “University of Zurich”, and the first sentence mentioning the arguments is the following: “Einstein was awarded a PhD by the University of Zurich”, which expresses the relation.

Step 2: Extract Relation Topics

Relation Repository
(>7,000 Relations)
Recall:80.24%,
Precision: 84.6%

Diffusion Wavelets
(no extra human labelling,
parameter free)



Relation Topics

Existing Topic Models (LDA, LSI):

*a multinomial distribution over **words**.*

Relation Topic:

*We define a relation topic as a multinomial distribution over **relations**.*

A Relation Topic Example

doctoralstudents (0.113201),
candidate (0.014662),
academicadvisors (0.008623),
notablestudents (0.003829),
……

#=7600

where doctoraladvisor is a DBpedia relation and 0.683366 is its contribution to the topic.

Representation of Relation (dependency path)

- An Example of *ActiveYearEndDate* relation: “He retired in 1991.”
  - person100007846 | year115203791 | - | in | retire
  - person100007846 | year115203791 | retirement | - | announce

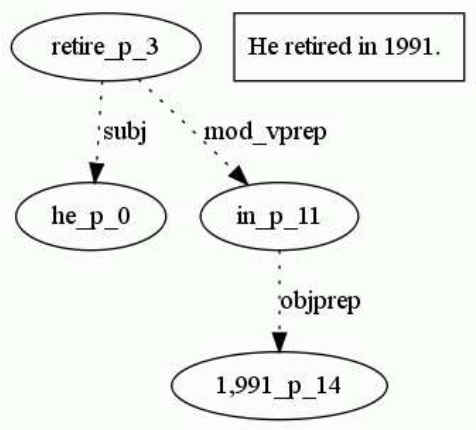
Yago Type of Argument 1

Yago Type of Argument 2

Noun

Prep

Verb

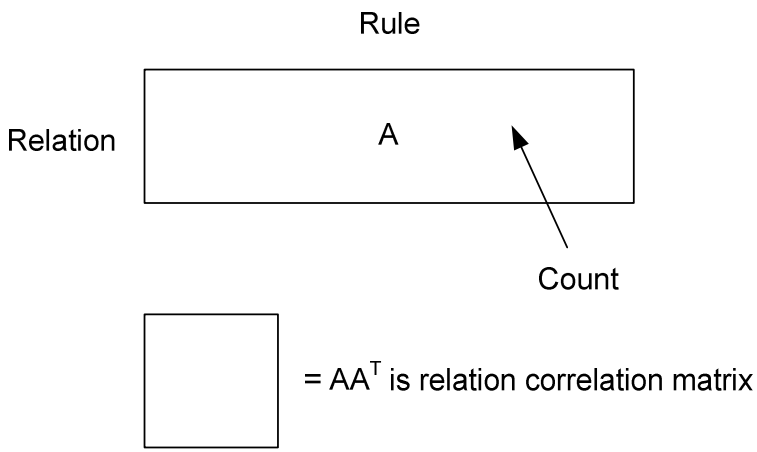


A relation is represented by a set of <rule, count> pairs.

person100007846 | year115203791 | - | in | retire (5)
person100007846 | year115203791 | retirement | - | announce (2)

Create Relation-Relation Correlation Matrix

- Some relations share some rules.
- Create the relation-relation correlation matrix.



- Apply Diffusion Wavelets [Coifman, Maggioni, 2006] to extract relation topics from the correlation matrix.

Diffusion Wavelets

- Input:
  - the correlation matrix.
- Output:
  - the number of levels of the topical hierarchy, as well as the topics at each level.

- Comparison with other topic modeling techniques (e.g. LDA, LSI)
  - No need to specify the number of topics
  - Multilevel
  - Fast

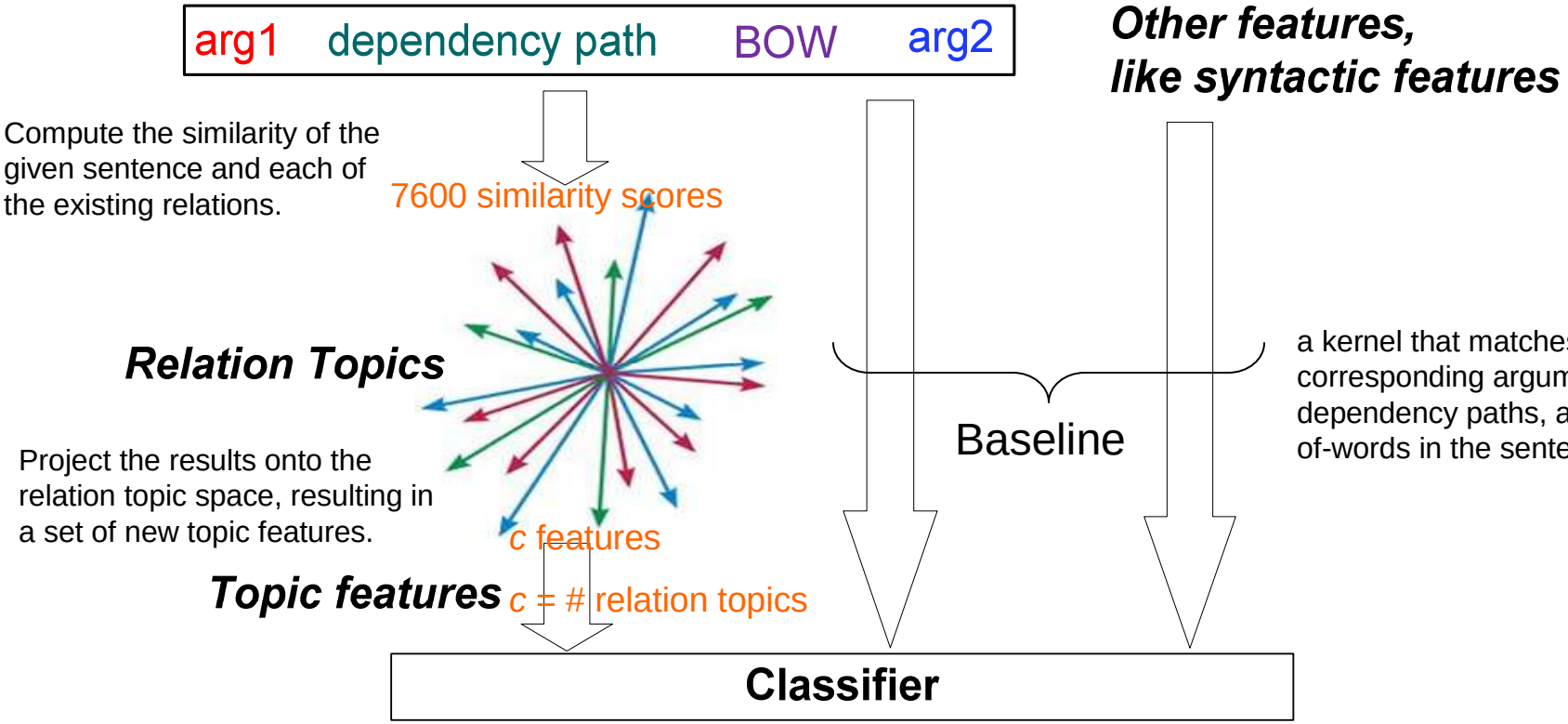
More Relation Topic Examples (Level 2)

Topic 1	founded 0.219976	built 0.036247	establisheddate 0.027167	…
Topic 2	before 0.332123	after 0.198355	predecessor 0.065609	…
Topic 3	architecture 0.895874	style 0.029560	architecturestyle 0.013483	…
Topic 4	subfamilia 0.783230	family 0.017283	genus 0.006933	…

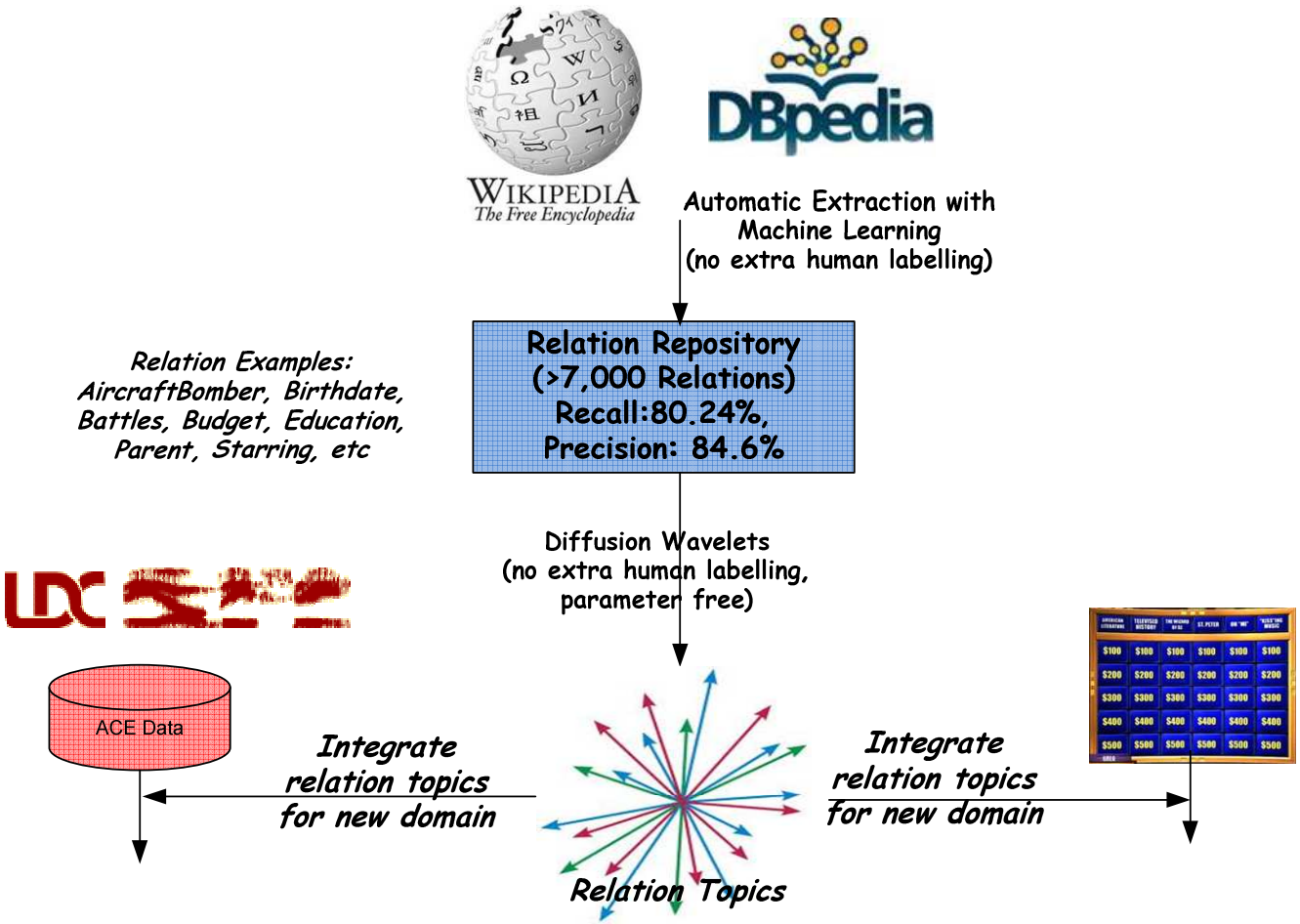
Step 3: Integration

Input: "Sauve announced his retirement from the NHL in 1989."

Sauve announced his retirement from the NHL in 1989.



The Overall TWREX Architecture



RELATION EXTRACTION

Outline

- Background:
  - Relation Detection
  - Watson Pipeline
  - Relation Detection in Watson
- TWREX Architecture
  - Challenges
  - Construction of Relation Repository
  - Relation Topics
  - Integration (Information Fusion)
- Experiments & Conclusions

Experimental Results: Relation Topics

Table 1: Number of topics at different levels (DBpedia Relations) under 5 different settings: use args, noun, preposition and verb; arg_1 only; arg_2 only; noun only and verb only.

Level	args & words
1	7628
2	269
3	32
4	7
5	3
6	2
7	1

Table 2: 3 topics at level 5 (all word types and args).

Top 4 Relations and Their Contributions
starring 86.6%, writer 3.8%, producer 3.2%, director 1.6%
birthplace 75.3%, clubs 6.1%, deathplace 5.1%, location 4.1%
clubs 55.3%, teams 9.3%, nationalteam 6.3% college 6.0%

Table 3: Some topics at level 2 (all word types and args).

Top Relations
activeyearsenddate, careerend, finalyear, retired
commands, partof, battles, notablecommanders
occupation, shortdescription, profession, dates
influenced, schooltradition, notableideas, maininterests
destinations, end, through, posttown
prizes, award, academyawards, highlights
inflow, outflow, length, maxdepth
after, successor, endingterminus
college, almamater, education

Evaluation: 2004 ACE Data

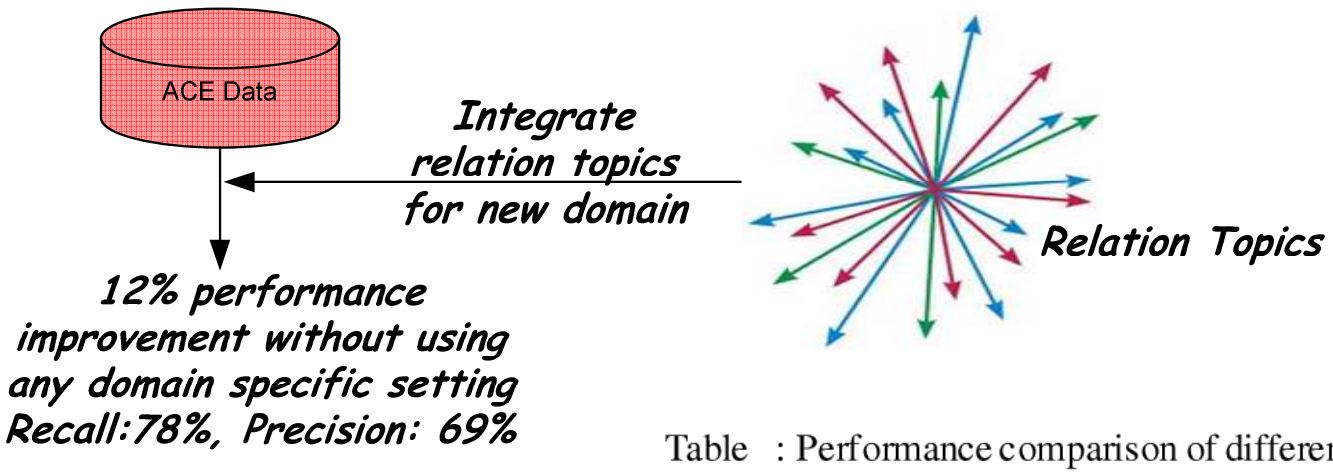


Table : Performance comparison of different approaches with SVM over the ACE 2004 data. P: Precision, R: Recall, F: F-measure.

Approaches	P(%)	R(%)	F(%)
Convolution Tree Kernel	72.5	56.7	63.6
Composite Kernel (linear)	73.50	67.00	70.10
Syntactic Kernel	69.23	70.50	70.35
Nguyen, etc (2009)	76.60	67.00	71.50
Baseline	62.00	61.19	61.15
Baseline + Topics	**69.15**	**77.88**	**73.24**

Reason for the improvement over the State of the Art Methods :

Bring in the knowledge from the existing relations.

Conclusion

- TWREX is a Watson component. It makes use of the information fusion concept.

- A novel approach to relation extraction by reusing the knowledge gained from the other domains
  - Use dbpedia and wikipedia to automatically gather instances for a large repository of relations.
  - Use diffusion wavelets to extract topics space
  - Use topics space features in addition to traditional features in relation extraction

- Results: topics provide +12% improvement on ACE 2004 relation extraction task

- Future Work