A (very brief) overview of Information Extraction

NLP ML Web
Fall 2013
Andrew Rosenberg
TA/Grader: David Guy Brizan

Information Extraction

- Finding structured information from unstructured (or lightly structured) text.

Example

WASHINGTON — Senate Republicans on Sunday kept up the drumbeat of blame against President Obama for what they say is his failure to negotiate with them on the fiscal crisis that will come to a head on Thursday, when the government will run out of money to pay its bills. As the Republicans pointed fingers at the White House, Senators Harry Reid and Mitch McConnell were set to sit down again on Sunday in an effort to come up with some sort of agreement — even one that will kick the most pressing problems down the road for a few weeks or months.

- Names: “Senate Republicans”, “President Obama”, “the Republicans”, “the White House”, “Senators Harry Reid”, “Mitch McConnell”
- Entity Linking: e1={“Senate Republicans”, “the Republicans”}, e2={“President Obama”, “his”, “the White House”}
- Title: title(“President”, “Obama”), title(“Senator”, “Harry Reid”), title(“Senator”, “Mitch McConnell”)
- “Blame” Event: “X kept up the drumbeat of blame against Y”, “X pointed fingers at Y”.

http://www.nytimes.com/2013/10/14/us/politics/budget-and-debt-limit-debate.html?hp

Overview

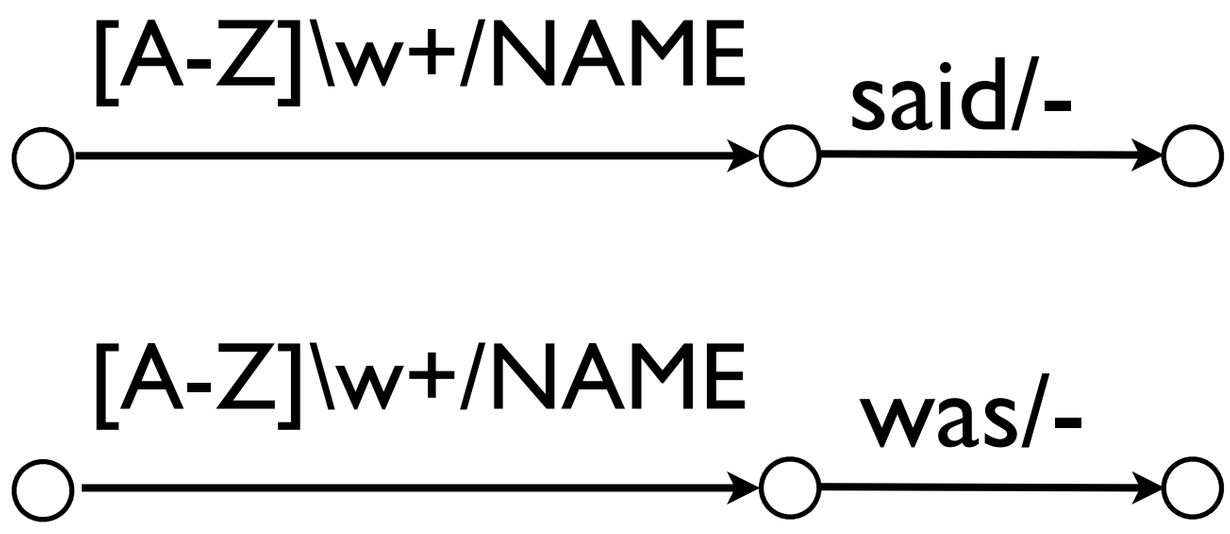
- Name Tagging
  - sequence models for Name tagging. (CRF)
  - pattern learning (wFST)
- co-reference resolution
- Slot Filling
- Bootstrapping
- Distant supervision

Name Tagging

- Identify the “Named Entities” in text.
  - Also Named Entity Recognition (NER)
- People
- Organizations including companies, teams, etc.
- Locations and/or Geo-political Entities (GPE)
- Also, Temporal Expressions, Currency

Simplest Approach

- Regular Expressions.
  - “ ([A-Z]\w+) said”
  - “ ([A-Z]\w+) was”
- FST representation



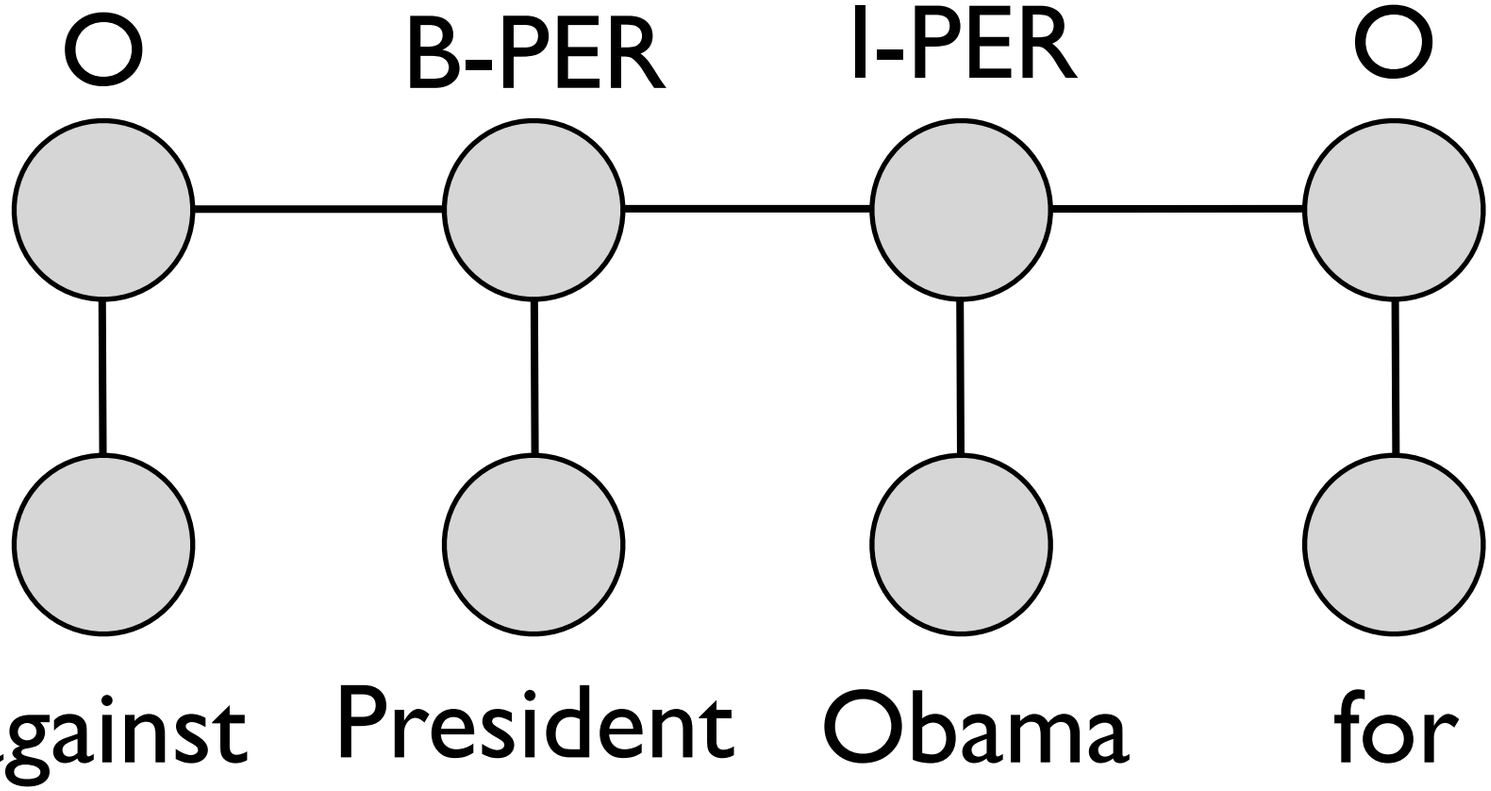
Feature Based Name Tagging

- Train a classifier based on labeled data.
- Identify features
  - Is capitalized
  - Is in a list of known names
  - Is a title (Mr. Ms. etc.)
  - Follows/Precedes a known name
  - Ends with a period
  - Precedes a verb
  - etc.

Sequence Modeling

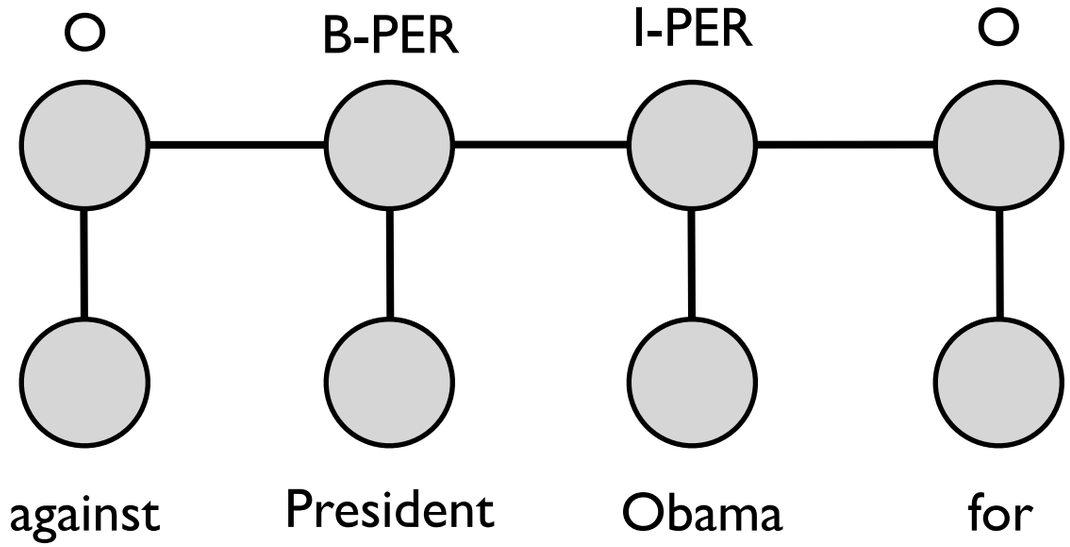
- **BIO** tagging.
  - **B**egin-Tag, **I**n-Tag, **O**ut
- There are dependencies between these tags.
  - I-PER must follow B-PER.
  - B-GPE is likely to follow O
- Want to tag the whole sequence simultaneously

Conditional Random Fields



- Conditional Model.
- Discriminatively Trained over a full sequence

Conditional Random Fields



$$p(\vec{y}|\vec{x};\theta) = \frac{1}{Z(\vec{x},\theta)} \exp\left(\sum_j \theta_j F_j(\vec{x},\vec{y})\right)$$

Feature Functions

$$= \frac{1}{Z(\vec{x},\theta)} \exp\left(\theta^T F(\vec{x},\vec{y})\right)$$

$$Z(\vec{x},\theta) = \sum_y \exp(\theta^T F(\vec{x},\vec{y}))$$

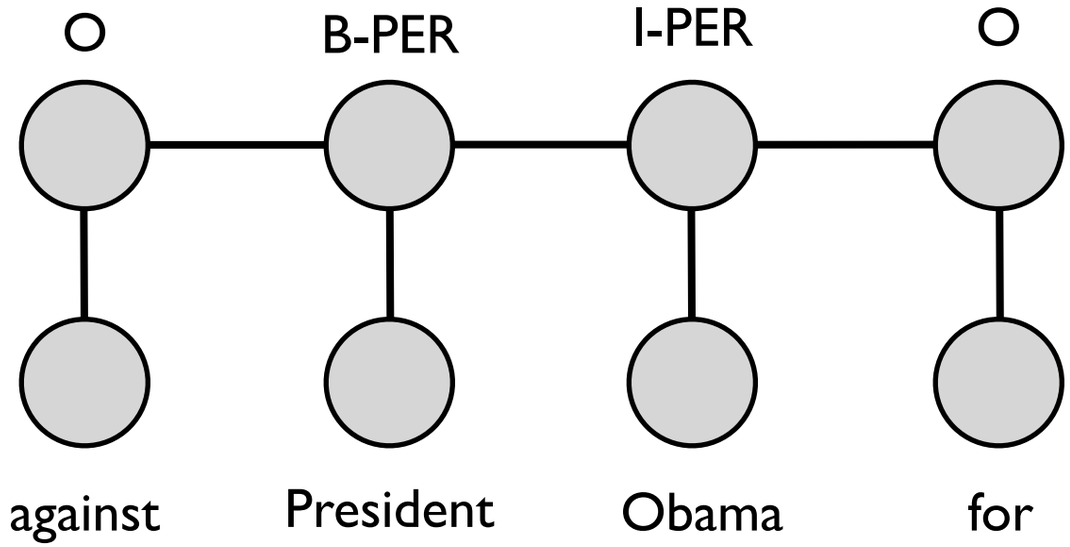
defined over pairs of outputs y_i, y_{i+1}, and any x_j

Conditional Random Fields

$$p(\vec{y}|\vec{x};\theta) = \frac{1}{Z(\vec{x},\theta)} \exp\left(\sum_j \theta_j F_j(\vec{x},\vec{y})\right)$$

$$= \frac{1}{Z(\vec{x},\theta)} \exp\left(\theta^T F(\vec{x},\vec{y})\right)$$

$$Z(\vec{x},\theta) = \sum_y \exp(\theta^T F(\vec{x},\vec{y}))$$



- Training: Gradient Descent
- Decoding: Viterbi (like HMM)

CRF vs. HMM

	HMM (generative)	CRF (discriminative
Marginal or Language Model P(sequence)	Forward algorithm linear in the length of the sequence	no.
Find optimal label sequence	Viterbi, linear in the length of the sequence	Viterbi, linear in the length of the sequence
Supervised Parameter Estimation	Bayesian Learning, Fast, Easy	Convex Optimization, Slow, harder
Unsupervised Parameter Estimation	Baum-Welch (non-convex optimization) Can be slow	Very difficult
Feature Functions	Parents and Children in the graph Very restricted	Arbitrary functions of a state and any portion of observed nodes

http://knight.cis.temple.edu/~yates/cis8538/sp11/slides/conditional-random-fields.ppt

Application of CRFs to Name Tagging

- Features in Stanford Tagger:
  - Current, Previous, Following Word
  - Current word Character n-gram
  - Current POS tag
  - Surrounding POS tag sequence
  - Current Word Shape
  - Surrounding Word Shape Sequence
  - Word in Left or Right Window
- ~86.86 F-measure on LOC, ORG, MISC, PER
- ~92.29 F-measure on Start and End time, Speaker and Location in seminar announcements.

Bootstrapping in SNOWBALL

- User provides some seed information
  - Set of names and birth dates
  - Set of companies and HQ Location

Company	HQ Location
Microsoft	Redmond
IBM	Armonk
Exxon	Irving
Intel	Santa Clara
Boeing	Seattle

Bootstrapping in SNOWBALL

- Find sentences that contain both terms.
  - Microsoft's headquarters in Redmond
  - Irving-based Exxon
  - Boeing was incorporated in Seattle

Company	HQ Location
Microsoft	Redmond
IBM	Armonk
Exxon	Irving
Intel	Santa Clara
Boeing	Seattle

Bootstrapping in SNOWBALL

- Turn these into patterns
  - <string1>'s headquarters in <string2>
  - <string2>-based <string1>
  - <string1> was incorporated in <string2>

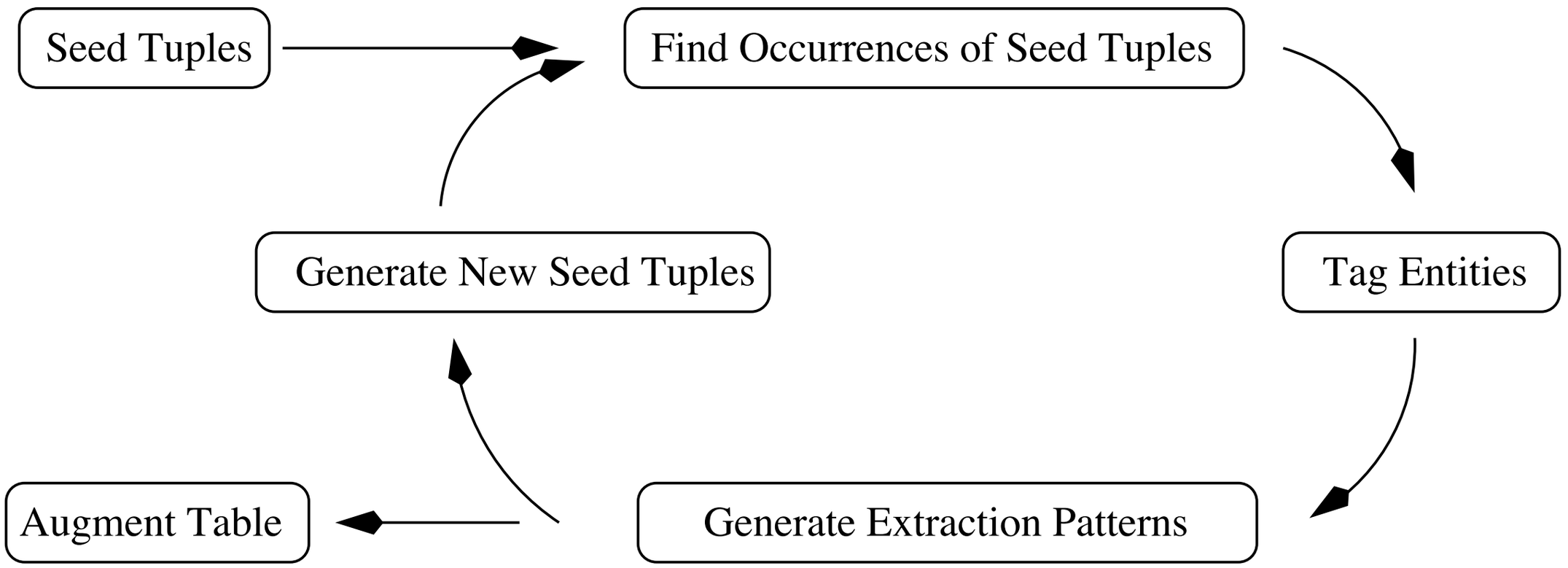
Company	HQ Location
Microsoft	Redmond
IBM	Armonk
Exxon	Irving
Intel	Santa Clara
Boeing	Seattle

Bootstrapping in SNOWBALL

- Find new sentences that match these patterns
  - <string1>'s headquarters in <string2>
  - <string2>-based <string1>
  - <string1> was incorporated in <string2>

Company	HQ Location
Microsoft	Redmond
IBM	Armonk
Exxon	Irving
Intel	Santa Clara
Boeing	Seattle
Google	Mountain View

Bootstrapping in SNOWBALL



- Not all patterns are equally good.
- Estimate the confidence of a pattern by how consistent it is with previous known information.

Co-reference Resolution

- Do two expressions refer to the same entity?
  - Who or what does a pronoun refer to?
  - President Bush, Bush, George W. Bush, Dubya, etc.
- A Machine Learning Approach to Coreference Resolution of Noun Phrases. Soon et al. 2001

Approach to Co-reference resolution

- Static Classifier (C4.5 Decision Trees)
- Features based on 2 candidate tokens i and j
  - **Distance** How many sentences apart are the mentions.
  - **Pronoun** is I or J a pronoun?
  - **String Match** are they the same string? (ignoring articles and demonstrative pronouns)
  - **Definitive?** is the NP a definitive noun phrase? “the car”
  - **Demonstrative?** is the NP a demonstrative noun phrase? “this car”
  - **Number agreement**

Approach to Co-reference resolution

- **Semantic class agreement** person, organization, date, etc.
- **Gender agreement**
- **Both Proper names?**
- **Alias?** Is this a possible alias? Date representations? possible acronym? Mr. Burns vs. Montgomery Burns
- **Appositive?** Could the two terms be in an appositive construction?
  "Andrew Rosenberg, assistant professor of computer science,..."

Paradigms for Information Extraction

- **Supervised**: Relations are hand labeled in text.
  - Pros: High quality annotations
  - Cons: Expensive to produce; domain dependent
- **Fully Unsupervised:** Analyze a large set of text
  - Pros: No annotation is necessary
  - Cons: Discovered relations may not map to desired relations
- **Bootstrapping**: Start with seed patterns, discover more relations, use these to identify more patterns
  - Pros: Low cost; flexible.
  - Cons: Semantic drift; low precision

Distant Supervision

- **Distant Supervision:** Use labeled data from an external resource as labels.
- Identify relations in Freebase.
  - 1.8M instances of 102 relations connecting 940k entities.
  - Supervision is in **relations** not **realizations**.
- Assumption: If two entities that are in a relation with each other, and they appear in the same sentence, the sentence may describe the relation.
- Examples:
  - location-contains: (Austria, Vienna)
  - film-director: (Steven Spielberg, Saving Private Ryan)

Distant Supervision

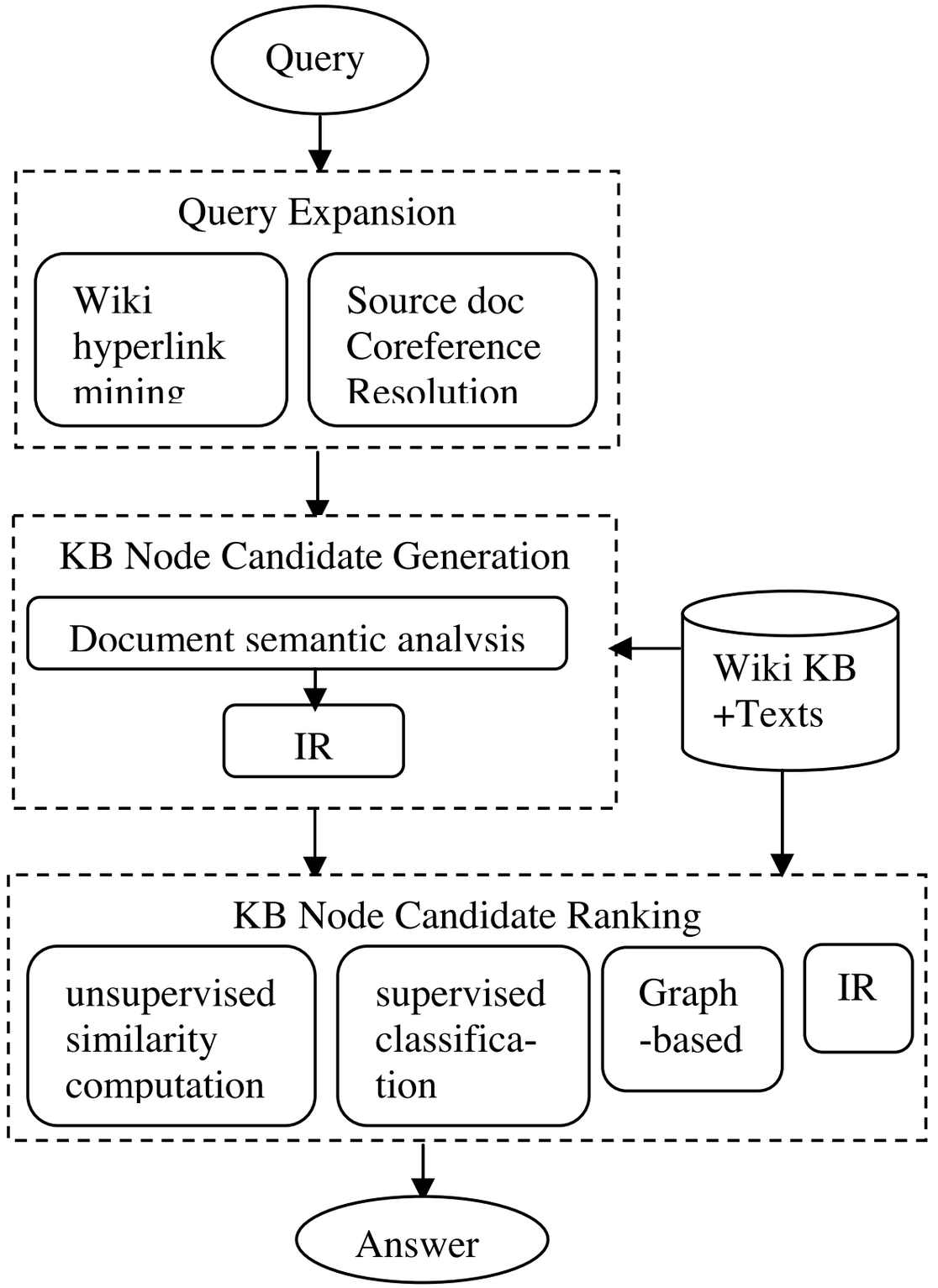
- Features are combined
  - (Steven Spielberg, Saving Private Ryan)
  - Find all sentences, extract features, collapse all features.
  - Same is done during test.
- Feature Types
  - **Lexical:** sequence of words between entities; POS tags between; which was first; surrounding words
  - **Syntactic:** Dependency path between entities
  - **Name Tagging:** PER, LOC, ORG, MISC, None for each entity
  - **Conjunction:** The features are conjunctions of features and NE tags. (very sparse, but high precision)
- Logistic regression classifier (L2-regularized, L-BGFS)

Knowledge Base Population

- Have an entity (‘query’) in a Knowledge Base
- **Entity Linking**
  - Find instances of this entity in text
- **Slot Filling**
  - Add additional attribute (slots) to the Knowledge Base for this entity.

Entity Linking

- Input: a string of text and a source document
- Output: A KB entity or NIL
- Source document is used for disambiguation



Entity Linking

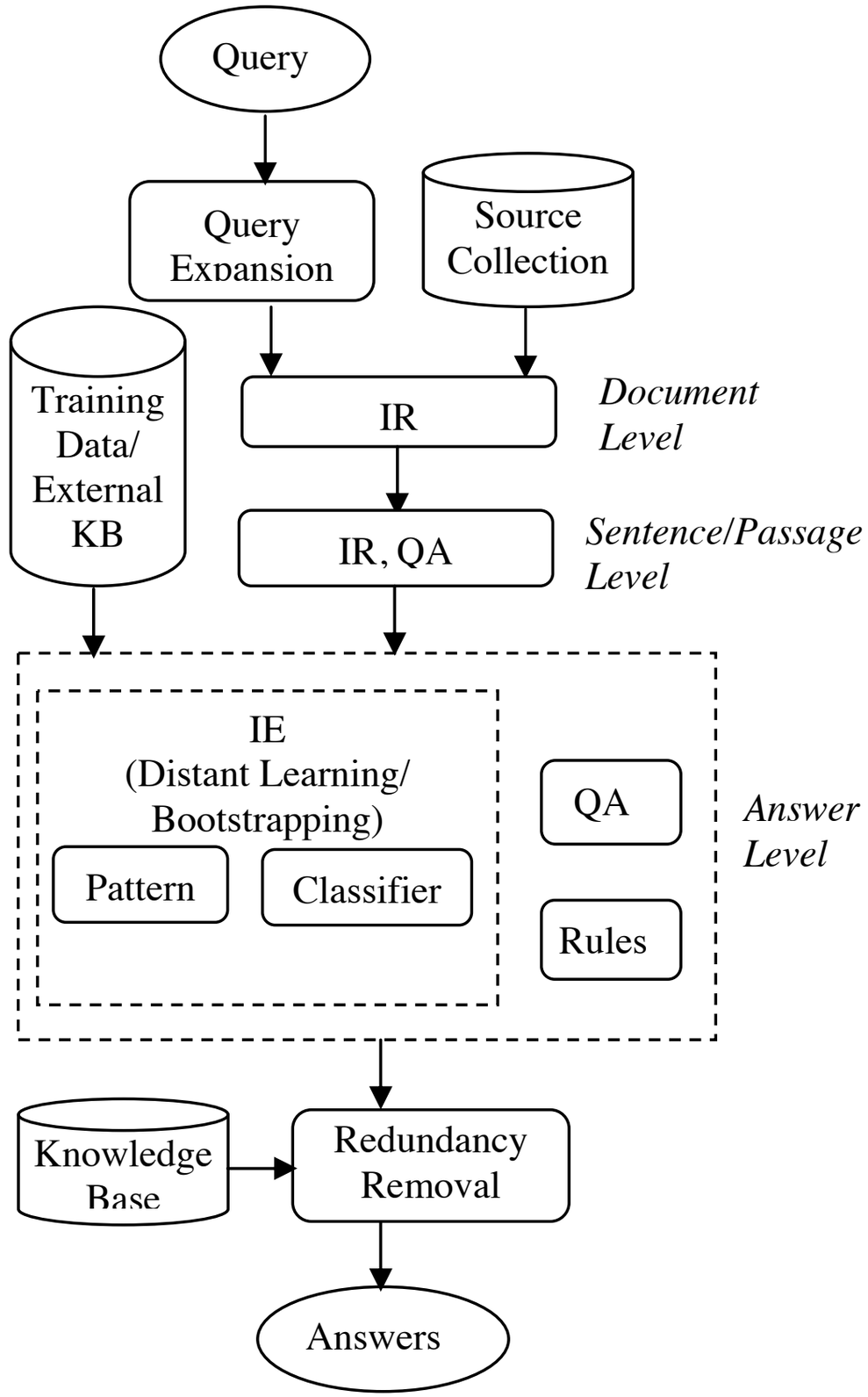
- **Query Expansion**
  - Wikipedia link structure, coreference
- **Candidate Generation**
  - Semantic analysis (name tagging, synonyms, wikipedia data etc.)
  - Standard Information Retrieval
- **Candidate Ranking**
  - Similarity based on unlabeled context of the query
  - Supervised Classification
  - Information Retrieval

Entity Linking

- Remaining Challenges
  - Documents where the query isn't the focus.
  - Analysis of hyperlinked material
    - Other links
  - Salience Ranking
    - *Georgia Tech* vs. *Georgian Technical University*

Slot Filling

- Input: entity name, type (PER, ORG, GPE), source document, KB id, known attributes
- Output: new slot fills
  - 26 types for PER: age, birthplace, spouse, employer, etc.
  - 16 for ORG: founder, HQ location, top employees, etc.



Slot Filling

- **Retrieve Documents**
  - Query Expansion
- **Answer Extraction**
  - Pattern Learning
  - Supervised Classification
  - Hand-coded Rules
  - All: Use of external resources (Freebase, DBpedia)
- **Answer Merging**
  - Identifying Redundancy

Slot Filling - Remaining Challenges

- Cross-sentence IE.
  - **Lahoud** is married to an Armenian and the couple have three children. Eldest son **Emile Emile Lahoud** was a member of parliament between 2000 and 2005
- Revertible Relations and Dependent Relations
  - **Julia Roberts** has given birth to her third child a boy named **Henry Daniel Moder**. **Henry** was born Monday in Los Angeles and weighed 8? lbs. **Roberts**, 39, and husband **Danny Moder**, 38, are already parents to...
    - ChildOf("Danny Moder", "Henry Daniel Moder")
- Challenging Co-reference
  - **Alexandra Burke** is out with the video for her second single ... taken from **the British artist**'s debut album"