CS 4650/7650
Information Extraction

Jacob Eisenstein

November 11, 2014

Course roadmap

- Words (WSD, classification and morphology)
- Sequences (tagging)
- Trees (parsing)
- Semantics and discourse
- **Applications**
  - Today: knowledge from text
  - Next week: machine translation

Knowledge from text

- **Information extraction**
  - input: schema of desired knowledge base
  - output: populate schema from text resources
- **Question answering**
  - input: natural language questions
  - output: natural language answers
  - intermediate representation usually includes structured knowledge base

(see wolfram alpha video)

WolframAlpha computational... knowledge engine

what are the three lightest metals?

Examples Random

Input interpretation:

3 lightest metallic elements

Results:

By atomic weight: More

1	lithium	6.941 u	
2	beryllium	9.012182 u	
3	sodium	22.98976928 u	

⋮

By density: More

1	lithium	0.535 g/cm^3	
2	potassium	0.856 g/cm^3	
3	sodium	0.968 g/cm^3	

⋮

Units »

Computed by Wolfram *Mathematica* Source information » Download page

WolframAlpha computational knowledge engine

how expensive are the three lightest metals?

Examples Random

Using closest Wolfram|Alpha interpretation: **three lightest metals**

More interpretations: expensive

Input interpretation:

3 lightest metallic elements

Results:

By atomic weight: More

1	lithium	6.941 u	
2	beryllium	9.012182 u	
3	sodium	22.98976928 u	

⋮

By density: More

1	lithium	0.535 g/cm^3	
2	potassium	0.856 g/cm^3	
3	sodium	0.968 g/cm^3	

WolframAlpha computational knowledge engine

what is the lightest radioactive element?

Examples Random

Input interpretation:

lightest radioactive elements

Results:

By atomic weight: More

1	technetium	98 u	
2	promethium	145 u	
3	polonium	209 u	
4	astatine	210 u	
5	radon	222 u	

By density: More

1	radon	0.00973 g/cm^3	
2	radium	5 g/cm^3	
3	promethium	7.264 g/cm^3	
4	polonium	9.196 g/cm^3	
5	actinium	10.07 g/cm^3	

WolframAlpha computational knowledge engine

what is the lightest element which is radioactive?

Examples Random

Assuming "element" is elements | Use as a class of elements instead

Input interpretation:

lightest radioactive elements

Results:

By atomic weight: More

1	technetium	98 u
2	promethium	145 u
3	polonium	209 u
4	astatine	210 u
5	radon	222 u

By density: More

1	radon	0.00973 g/cm^3
2	radium	5 g/cm^3

WolframAlpha computational... knowledge engine

of all of the radioactive elements, which is the lightest?

Examples Random

Using closest Wolfram|Alpha interpretation: **the radioactive elements**

More interpretations: elements which

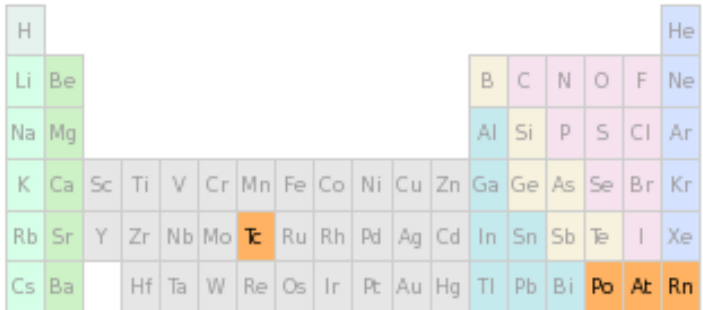
Input interpretation:

radioactive elements

Members: More

actinium | americium | astatine | berkelium | bohrium | californium | copernicium | curium | darmstadtium | dubnium | einsteinium | fermium | francium | hassium | lawrencium | meitnerium | mendelevium | neptunium | nobelium | plutonium | ... (total: 37)

Periodic table location:



Knowledge extraction in NLP

Knowledge extraction requires solving lots of NLP problems

- understand the source data (syntax, discourse, semantics)
- understand the query
- reason about how they fit together

The six Ws

- Who, what, where, when, why, how?
- IE is mostly concerned with the first four.

The six Ws

- Who, what, where, when, why, how?
- IE is mostly concerned with the first four.
  - **who/where**: named entity extraction and coreference (we've already talked about this)
  - **what**: usually defined in terms of *relations* between entities
  - **when**
    - parsing time expressions, finding the temporal order of events
    - this is a big part of IE, but I'm not going to talk about it today

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

Named entity recognition

Find and tag **mentions** of **entities** in text.

At a meeting of <ORG>the Thirteen</ORG>,
<PER>Pyat Pree</PER> tells <PER>Daenerys</PER>
that he has <OBJ>her dragons</OBJ> in the
<PER>House of the Undying</PER>.

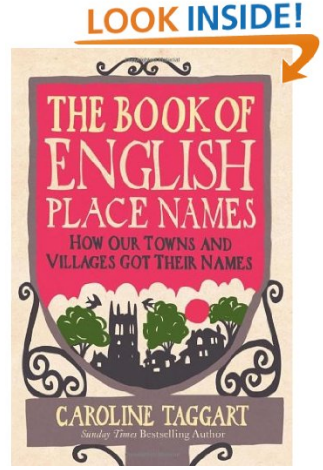
NER with rules

Entity recognition can be performed with rules.

Rule: TheGazOrganization
Priority: 50
// Matches "The <in list of company names>"
({Part of speech = DT | Part of speech = RB} {DictionaryLookup = organization})
→ Organization

Rule: LocOrganization
Priority: 50
// Matches "London Police"
({DictionaryLookup = location | DictionaryLookup = country} {DictionaryLookup
= organization} {DictionaryLookup = organization}?) → Organization

- These rules are from GATE (General Arch. for Text Engineering), `http://gate.ac.uk/`
- Rules may leverage POS tags and dictionaries.



NER with rules

Rule: INOrgXandY
Priority: 200
// Matches “in Bradford & Bingley”, or “in Bradford & Bingley Ltd”
({Token string = “in”})
({Part of speech = NNP}+ {Token string = “&”} {Orthography type = upperInitial}+ {DictionaryLookup = organization end}?):orgName → Organization=:orgName

Rule: OrgDept
Priority: 25
// Matches “Department of Pure Mathematics and Physics”
({Token.string = “Department”} {Token.string = “of”} {Orthography type = upperInitial}+ ({Token.string = ”and”} {Orthography type = upperInitial}+)?) → Organization

- Rules may overlap or disagree; the better the coverage, the more likely this is.
- Arbitrating disagreements is a complex engineering task.
- One solution: order rules by precision on training data.

NER as Sequence Labeling

Pyat/B-PER Pree/I-PER tells/O Daenerys/B-PER that/O he/O has/O her/B-OBJ dragons/I-OBJ ...

- **Tags**: B,I,O for each entity type
- **Features**: bag-of-words, word shape (characters), dictionary (list of known names), part-of-speech...
- **Method**: sequence labeling

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \sum_{i} \boldsymbol{\theta}^{\mathsf{T}} \mathbf{f}(\mathbf{w}, y_i, y_{i-1}, i)$$

  - Hidden Markov Model: $\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{w}, \mathbf{y}; \boldsymbol{\theta})$
  - Conditional Random Field: $\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} P(\mathbf{y} \mid \mathbf{w}; \boldsymbol{\theta})$
  - Structured Perceptron: $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^{(t)} + \mathbf{f}(\mathbf{w}, \mathbf{y}) - \mathbf{f}(\mathbf{w}, \hat{\mathbf{y}})$

Features on spans of text

- Dictionaries may contain **multitoken spans**
  (e.g. The House of the Undying)
- We want features that fire when a **span** matches a dictionary entry.
  $\mathbf{f}(\mathbf{w}, y_i, y_{i'}, i', i)$: set of features for the span from $i' + 1$ to i
- Can we still use Viterbi?

Features on spans of text

- Dictionaries may contain **multitoken spans** (e.g. The House of the Undying)
- We want features that fire when a **span** matches a dictionary entry. $\mathbf{f}(\mathbf{w}, y_i, y_{i'}, i', i)$: set of features for the span from $i' + 1$ to i
- Can we still use Viterbi?
- Can we still use dynamic programming?

Features on spans of text

- Dictionaries may contain **multitoken spans** (e.g. The House of the Undying)
- We want features that fire when a **span** matches a dictionary entry. $\mathbf{f}(\mathbf{w}, y_i, y_{i'}, i', i)$: set of features for the span from $i' + 1$ to i
- Can we still use Viterbi?
- Can we still use dynamic programming?

$$V(i, y) = \begin{cases} \max_{y'} \max_{i' \in i-L,\ldots,i-1} V(i', y') + \boldsymbol{\theta}^{\mathsf{T}} \mathbf{f}(\mathbf{w}, y_i, y_{i'}, i', i), & i > 0 \\ 0, & i = 0 \\ -\infty, & i < 0 \end{cases}$$

Features on spans of text

- Dictionaries may contain **multitoken spans** (e.g. The House of the Undying)
- We want features that fire when a **span** matches a dictionary entry. $\mathbf{f}(\mathbf{w}, y_i, y_{i'}, i', i)$: set of features for the span from $i' + 1$ to i
- Can we still use Viterbi?
- Can we still use dynamic programming?

$$V(i, y) = \begin{cases} \max_{y'} \max_{i' \in i-L,\ldots,i-1} V(i', y') + \boldsymbol{\theta}^\top \mathbf{f}(\mathbf{w}, y_i, y_{i'}, i', i), & i > 0 \\ 0, & i = 0 \\ -\infty, & i < 0 \end{cases}$$

- Complexity: $\mathcal{O}(nLm^2)$, with $n = \#|\mathbf{x}|, m = \#|\mathcal{Y}|, L =$ max span

The Information Extraction pipeline

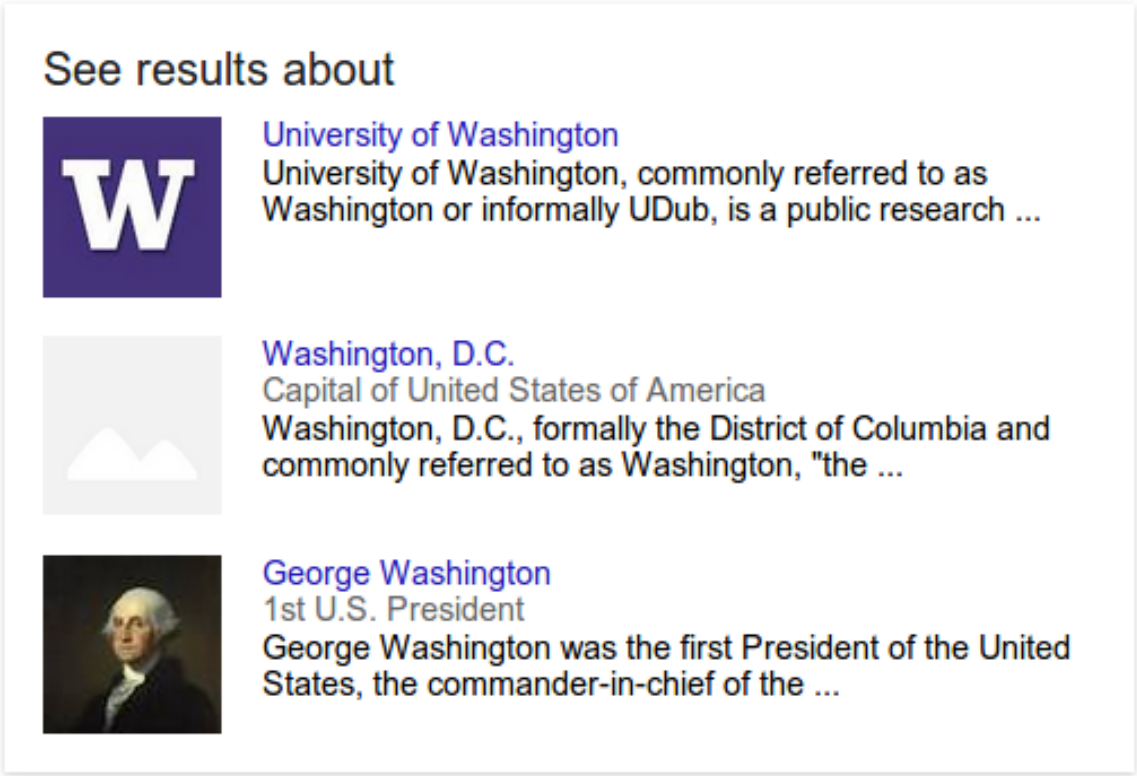
- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

Entity linking

Goal: link entity mentions to knowledge base entries.
Like multi-document coreference resolution, but must ultimately resolve to KB entry.



Entity linking: challenges

From Rao et al (2010)

1. Name variations: Boston Symphony Orchestra vs BSO, Qaddafi vs Gadaffi, etc
2. Name polysemy: Washington (person, place, football team, US Government, ...)
3. Absence: many entities do not appear in the KB.

These challenges are especially tough in combination: William Clinton is a variation of Bill Clinton, but appears in Wikipedia as two other individuals.

Entity linking: steps

Candidate identification

- Brute force: check Google Knowedge Graph for all strings that could link to an entity
- Add source document coreference resolution

Entity linking: steps

Candidate identification

- Brute force: check Google Knowedge Graph for all strings that could link to an entity
- Add source document coreference resolution

Ranking Supervised formulation (Dredze et al 2010):

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & ||\boldsymbol{\theta}||_2^2 \\ s.t. \quad & \boldsymbol{\theta}^{\mathsf{T}} f(\mathbf{w}_i, y_i) > \max_{\hat{y} \neq y_i} \boldsymbol{\theta}^{\mathsf{T}} f(\mathbf{w}_i, \hat{y}) \end{aligned}$$

Features:

- String match
- Popularity
- Local context and entity type
- Document context (similar entities)

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

Relations

A relation is a *predication* about a pair of entities.

- Davos **works for** Stannis
- King's Landing **is in** Westeros
- Joffrey's **father is** Jaime

Relations are typically permanent.

Example relations

From the Automatic Content Extraction (ACE) 2004 Task:

relation type	*subtypes*
physical	located, near, part-whole
personal-social	business, family, other
employment/membership/ subsidiary	employ-executive, employ-staff, employ-undetermined, member-of-group, partner, subsidiary, other
agent-artifact	user-or-owner, inventor-or-manufacturer, other
person-org affiliation	ethnic, ideology, other
GPE affiliation	citizen-or-resident, based-in, other
discourse	-

Freebase relations

Relation name	Size	Example
/people/person/nationality	281,107	John Dugard, South Africa
/location/location/contains	253,223	Belgium, Nijlen
/people/person/profession	208,888	Dusa McDuff, Mathematician
/people/person/place_of_birth	105,799	Edwin Hubble, Marshfield
/dining/restaurant/cuisine	86,213	MacAyo's Mexican Kitchen, Mexican
/business/business_chain/location	66,529	Apple Inc., Apple Inc., South Park, NC
/biology/organism_classification_rank	42,806	Scorpaeniformes, Order
/film/film/genre	40,658	Where the Sidewalk Ends, Film noir
/film/film/language	31,103	Enter the Phoenix, Cantonese
/biology/organism_higher_classification	30,052	Calopteryx, Calopterygidae
/film/film/country	27,217	Turtle Diary, United States
/film/writer/film	23,856	Irving Shulman, Rebel Without a Cause
/film/director/film	23,539	Michael Mann, Collateral
/film/producer/film	22,079	Diane Eskenazi, Aladdin
/people/deceased_person/place_of_death	18,814	John W. Kern, Asheville
/music/artist/origin	18,619	The Octopus Project, Austin
/people/person/religion	17,582	Joseph Chartrand, Catholicism
/book/author/works_written	17,278	Paul Auster, Travels in the Scriptorium
/soccer/football_position/players	17,244	Midfielder, Chen Tao
/people/deceased_person/cause_of_death	16,709	Richard Daintree, Tuberculosis
/book/book/genre	16,431	Pony Soldiers, Science fiction
/film/film/music	14,070	Stavisky, Stephen Sondheim
/business/company/industry	13,805	ATS Medical, Health care

Relations in text

- Typically we focus on cases in which the relation and the two entities are all mentioned in the same sentence. (Exception: Robert and Cersei were married. A son was born the next year)

Relations in text

- Typically we focus on cases in which the relation and the two entities are all mentioned in the same sentence. (Exception: Robert and Cersei were married. A son was born the next year)
- Relation extraction requires coreference resolution.
  - He has her dragons:
    Pyat Pree <HAS> Daenerys' dragons
  - Eddard died. His daughter, Sansa, said the Eulogy.
    - Must resolve his to Eddard and his daughter to Sansa
    - Then we can recover Sansa <Daughter-of> Eddard

Relations in text

- Typically we focus on cases in which the relation and the two entities are all mentioned in the same sentence. (Exception: Robert and Cersei were married. A son was born the next year)
- Relation extraction requires coreference resolution.
  - He has her dragons:
    Pyat Pree <HAS> Daenerys' dragons
  - Eddard died. His daughter, Sansa, said the Eulogy.
    - Must resolve his to Eddard and his daughter to Sansa
    - Then we can recover Sansa <Daughter-of> Eddard
- **Micro-reading**: correctly identify every relation *mention*
- **Macro-reading**: correctly identify every relation in the text

Knowledge-base population (KBP)

Extract attributes for each named person or organization:

entity	house	father	mother	position
Arya	Stark	Eddard	Catelyn	
Daenerys	Targaryen	Aerys		Mother-of-dragons
Qhorin	commoner			Knight-of-the-Watch

KBP is similar to relation extraction.

- Columns define relation types
- Rows define the left entity
- Cells define the right entity

Relations from patterns

- Early idea: lexical patterns, like regular expressions
- Possible patterns for PERSON lives-in LOCATION:
  - PERSON lives in LOCATION
  - PERSON lived in LOCATION
  - PERSON has lived in LOCATION
  - PERSON resides in LOCATION

Relations from patterns

- Early idea: lexical patterns, like regular expressions
- Possible patterns for PERSON lives-in LOCATION:
  - PERSON lives in LOCATION
  - PERSON lived in LOCATION
  - PERSON has lived in LOCATION
  - PERSON resides in LOCATION
- Can we generalize beyond lexical patterns?

Relations from patterns

- Early idea: lexical patterns, like regular expressions
- Possible patterns for PERSON LIVES-IN LOCATION:
  - PERSON <V BASE=LIVE> in LOCATION
  - PERSON has lived in LOCATION
  - PERSON resides in LOCATION
- Can we generalize beyond lexical patterns?
  - morphological analysis

Relations from patterns

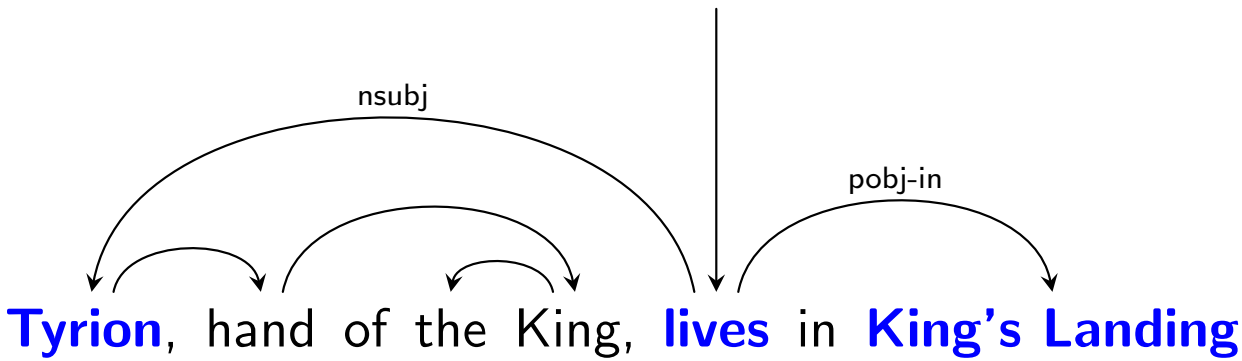
- Early idea: lexical patterns, like regular expressions
- Possible patterns for PERSON LIVES-IN LOCATION:
  - PERSON <VGROUP BASE=LIVE> in LOCATION
  - PERSON resides in LOCATION
- Can we generalize beyond lexical patterns?
  - morphological analysis
  - phrase chunking

Relations from patterns

- Early idea: lexical patterns, like regular expressions
- Possible patterns for PERSON LIVES-IN LOCATION:
  - PERSON <VGROUP SYNSET=LIVE#1> in LOCATION
- Can we generalize beyond lexical patterns?
  - morphological analysis
  - phrase chunking
  - lexical semantics

Syntactic patterns

Given a dependency parse, we can define more flexible patterns:

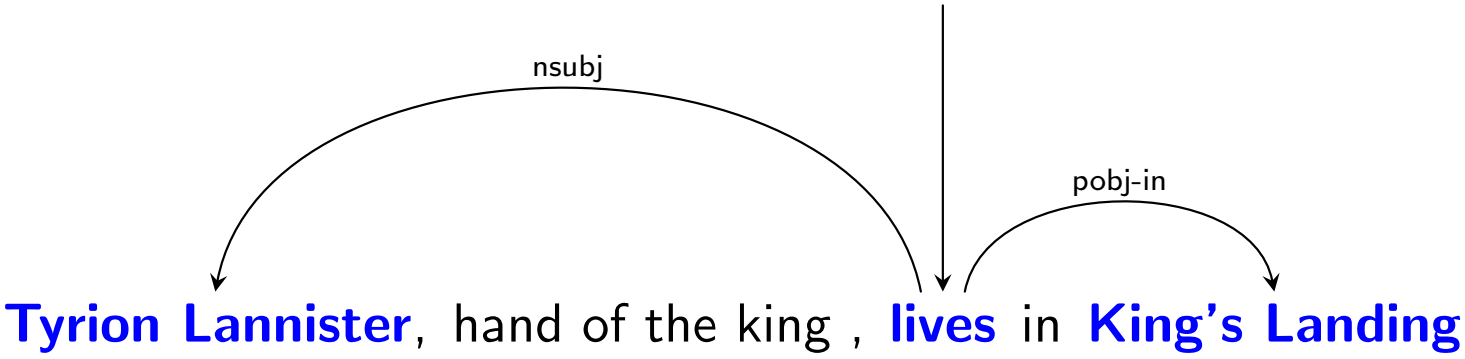


Supervised relation extraction

We can develop a classifier for each relation type, or a general classifier for detecting relations of any type.

- Feature-based classification
  - Compute features of each proposed relation
  - Learn weights from labeled data
- Kernel-based classification
  - Kind of like K-nearest-neighbors classification
  - The label for a test instance should be based on similar training instances

Feature-based relation extraction



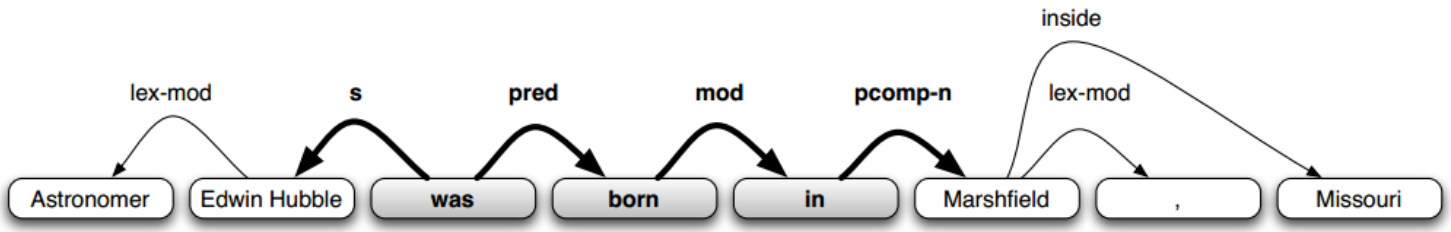
- **Heads**: Lannister, Landing, lives
- **POS**: NNP, VBZ, NNP
- **Types**: PER, LOC
- **Distance**: six words, zero entities
- **Words between entities**: hand; of; the; King; lives; in
- **Path**: NSUB↑-POBJ-IN↓
- **Path-words**: lives-in

Feature-based relation extraction

Mintz et al (2009) lexico-syntactic features:

Feature type	Left window	NE1	Middle	NE2	Right window
Lexical	[]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[]
Lexical	[Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[,]
Lexical	[#PAD#, Astronomer]	PER	[was/VERB born/VERB in/CLOSED]	LOC	[, Missouri]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{lex-mod}$,]
Syntactic	[]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Edwin Hubble $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]
Syntactic	[Astronomer $\Downarrow_{lex-mod}$]	PER	[$\Uparrow_s$ was $\Downarrow_{pred}$ born $\Downarrow_{mod}$ in $\Downarrow_{pcomp-n}$]	LOC	[$\Downarrow_{inside}$ Missouri]

Table 3: Features for 'Astronomer Edwin Hubble was born in Marshfield, Missouri'.



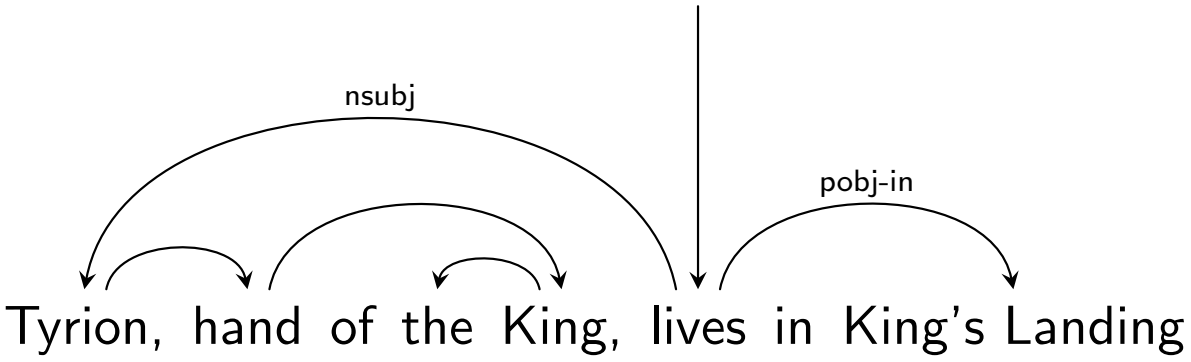
Whirlwind tour of kernel-based classification

A **kernel function** maps from pairs of instances to a non-negative real value.

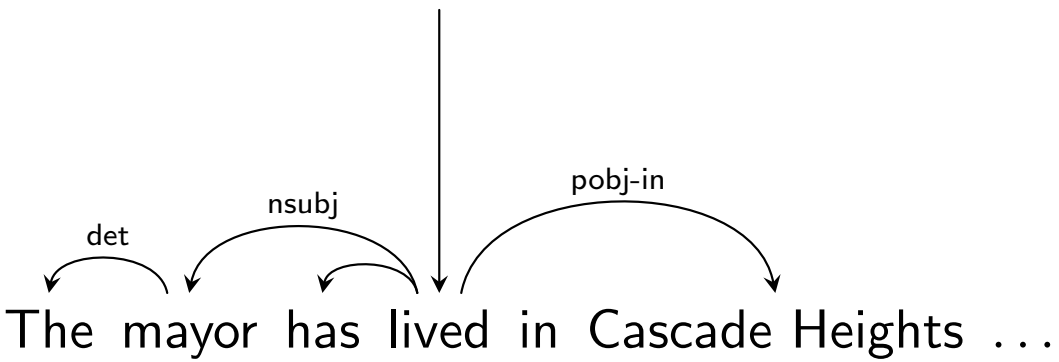
$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$$

- $K(x_1, x_2)$ is large if x_1, x_2 are similar, small if they are different
- K can count number of shared words, etc.
- K must be positive semi-definite.

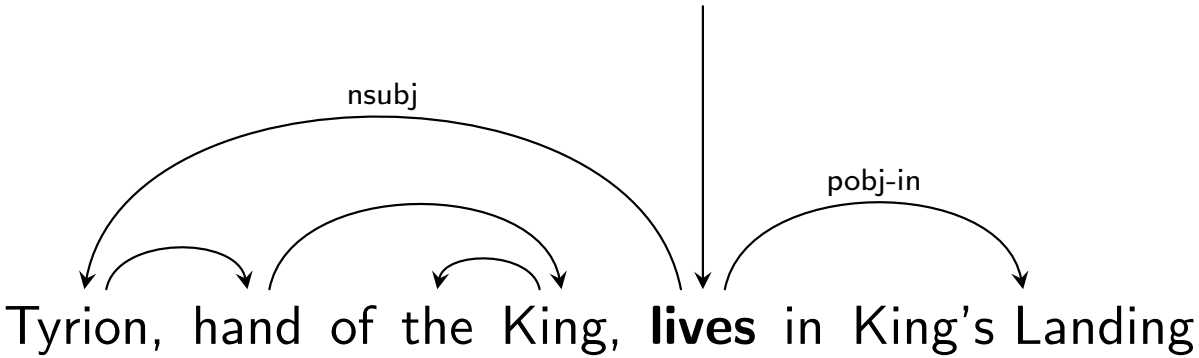
Dependency kernel example



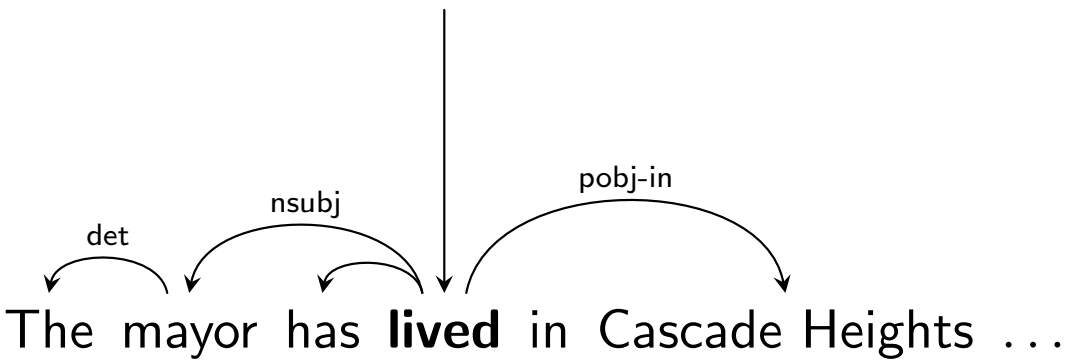
$$K(x_1, x_2) =$$



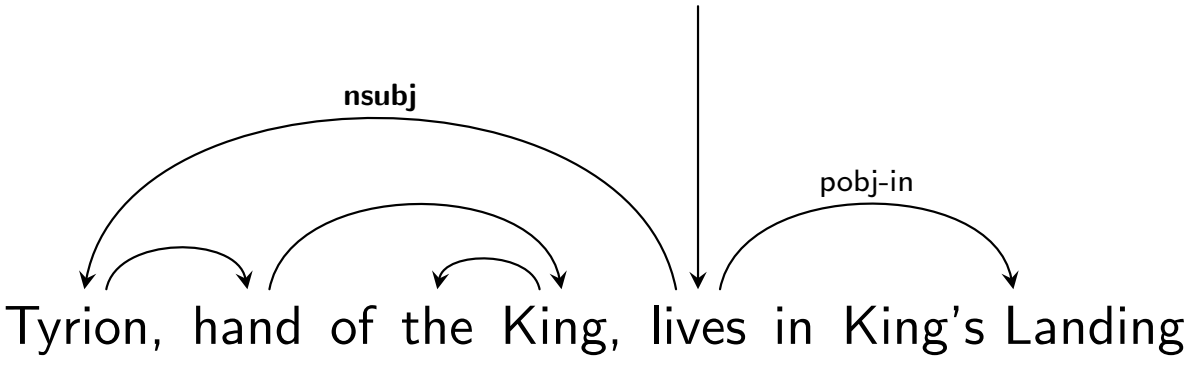
Dependency kernel example



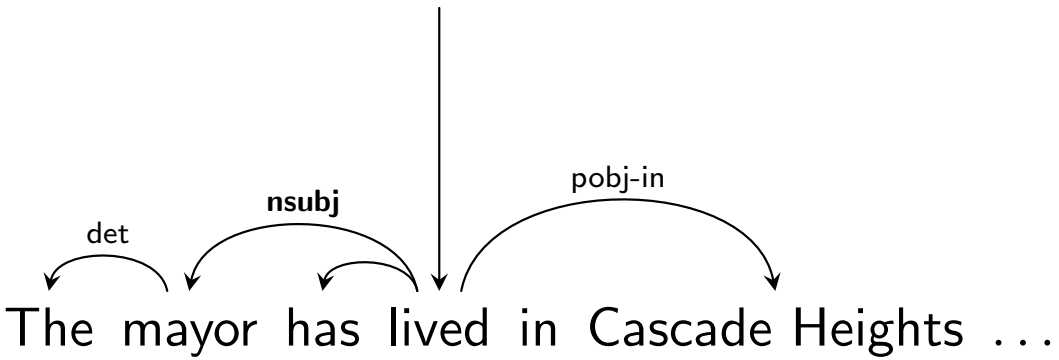
$$K(x_1, x_2) = \quad 1$$



Dependency kernel example

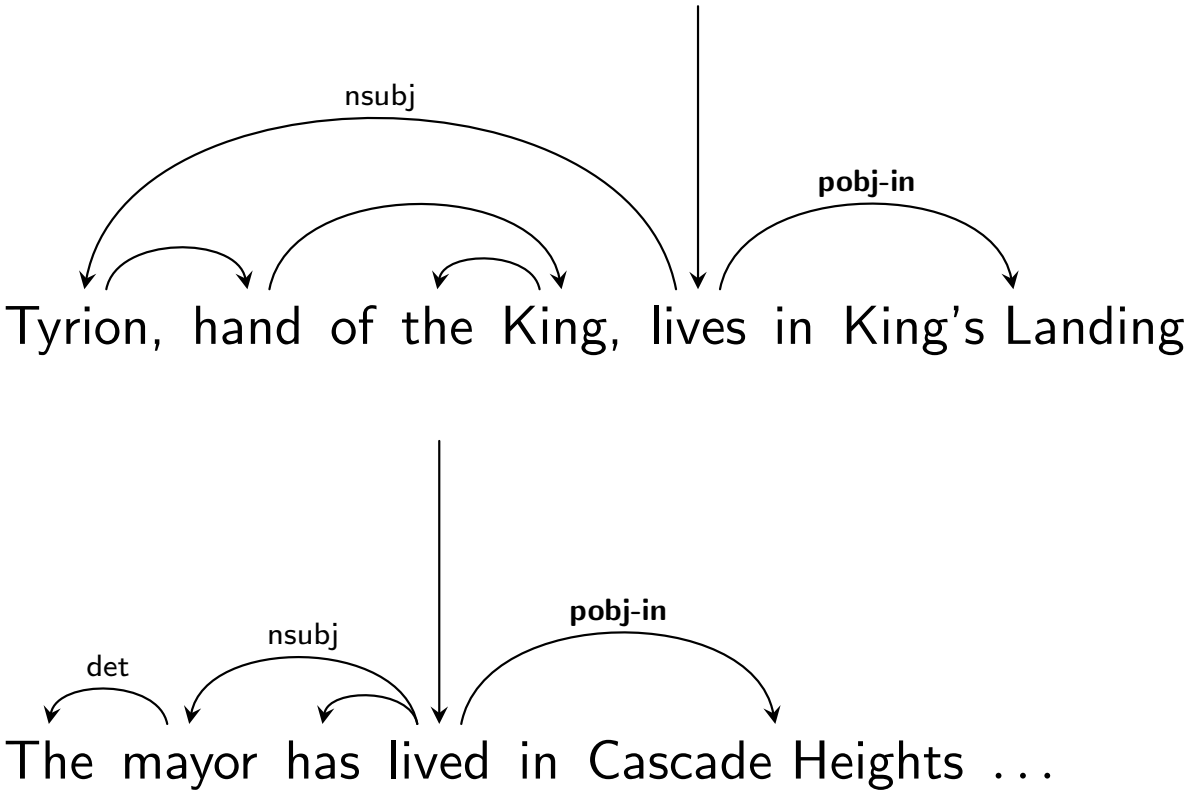


$$K(x_1, x_2) = 1$$



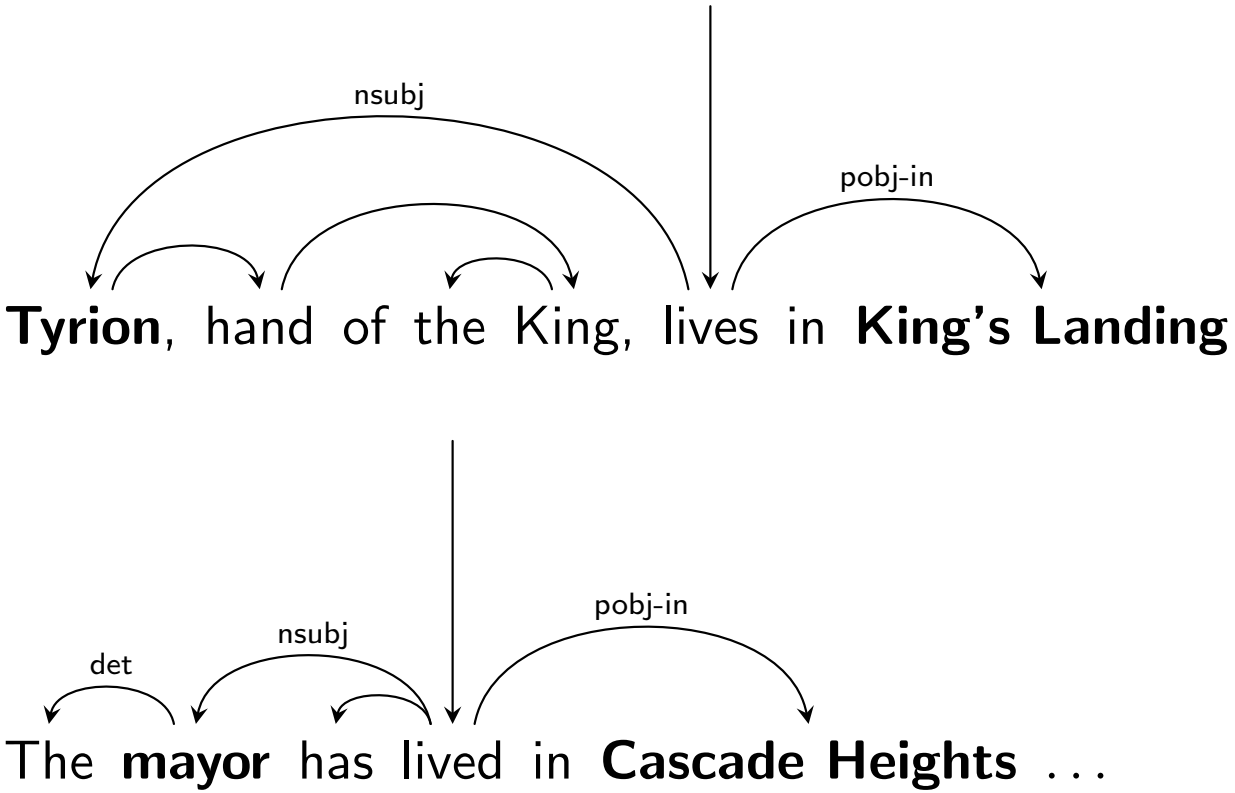
$$+ 1$$

Dependency kernel example



$$K(x_1, x_2) = 1 + 1 + 1$$

Dependency kernel example



$$
\begin{aligned}
K(x_1, x_2) = \; & 1 \\
& + 1 \\
& + 1 \\
& + 0 \\
= & 3
\end{aligned}
$$

Kernel-based classification

Binary classification rule, for $y_i \in \{-1, 1\}$

$$\hat{y}(\mathbf{x}) = \text{sign}(\sum_{i}^{N} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b)$$

Each $\alpha_i \geq 0$ is a parameter, which must be learned.

Kernel-based classification

Binary classification rule, for $y_i \in \{-1, 1\}$

$$\hat{y}(\mathbf{x}) = \text{sign}(\sum_{i}^{N} y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b)$$

Each $\alpha_i \geq 0$ is a parameter, which must be learned.

$$\max_{\alpha} L(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_j y_i, y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$s.t. \quad \alpha_i \geq 0, \forall i$$

$$\sum_i \alpha_i y_i = 0$$

Learning typically involves inverting the kernel matrix K.

Other training paradigms: bootstrapping

- Start with a few seed patterns
- Extract some high-confidence relations
- Induce more patterns
- Extract more relations
- ...

DIPRE (Brin, 1998)

- Relation of interest : (author, book)
- DIPRE's algorithm:
  - Given a small seed set of (author, book) pairs

1. Use the seed examples to label some data.
2. Induces patterns from the labeled data.
3. Apply the patterns to unlabeled data to get new set of (author,book) pairs, and add to the seed set.
4. Return to step 1, and iterate until convergence criteria is reached

Seed: (Arthur Conan Doyle, The Adventures of Sherlock Holmes)

A Web crawler finds all documents contain the pair.

...

Read The Adventures of Sherlock Holmes by Arthur Conan Doyle online or in you email

...

Extract **tuple**:

[0, Arthur Conan Doyle, The Adventures of Sherlock Holmes, Read, online or, by]

A tuple of 6 elements: [order, author, book, prefix, suffix, middle]

order = 1 if the author string occurs before the book string, = 0 otherwise

prefix and *suffix* are strings contain the 10 characters occurring to the left/right of the match

middle is the string occurring between the author and book

…

know that Sir Arthur Conan Doyle wrote The Adventures of Sherlock Holmes, in 1892

…

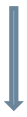
Extract tuple:

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, now that Sir, in 1892, wrote]

…

When Sir Arthur Conan Doyle wrote the adventures of Sherlock Holmes in 1892 he was high

…



Extract tuple:

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, When Sir, in 1892 he, wrote]

Extracted list of tuples:

[0, Arthur Conan Doyle, The Adventures of Sherlock Holmes, Read, online or, by]

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, now that Sir, in 1892, wrote]

[1, Arthur Conan Doyle, The Adventures of Sherlock Holmes, When Sir, in 1892 he, wrote]

…

Group tuples by matching *order* and *middle* and induce patterns

Induce patterns from group of tuples:

[longest-common-suffix of prefix strings, author, middle, book, longest-common-prefix of suffix strings]

Pattern:

[Sir, Arthur Conan Doyle, wrote, The Adventures of Sherlock Holmes, in 1892]

Pattern with wild card expression:

[Sir, .*?, wrote, .*?, in 1892]

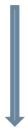
Use the wild card patterns **[Sir, .*?, wrote, .*?, in 1892]**

search the Web to find more documents

...

Sir Arthur Conan Doyle **wrote** Speckled Band **in 1892**, that is around 62 years apart which would make the stories

...



Extract new relations:

(Arthur Conan Doyle, Speckled Band)

Repeat the algorithm with the new relation.

Other training paradigms: distant supervision

Problems with bootstrapping (Mintz et al, 2009)

- [Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brother's story
  (Spielberg could be producer, actor)
- Allison co-produced the award-winning [Saving Private Ryan], directed by [Steven Spielberg]
  (Saving Private Ryan might not be a film)

Other training paradigms: distant supervision

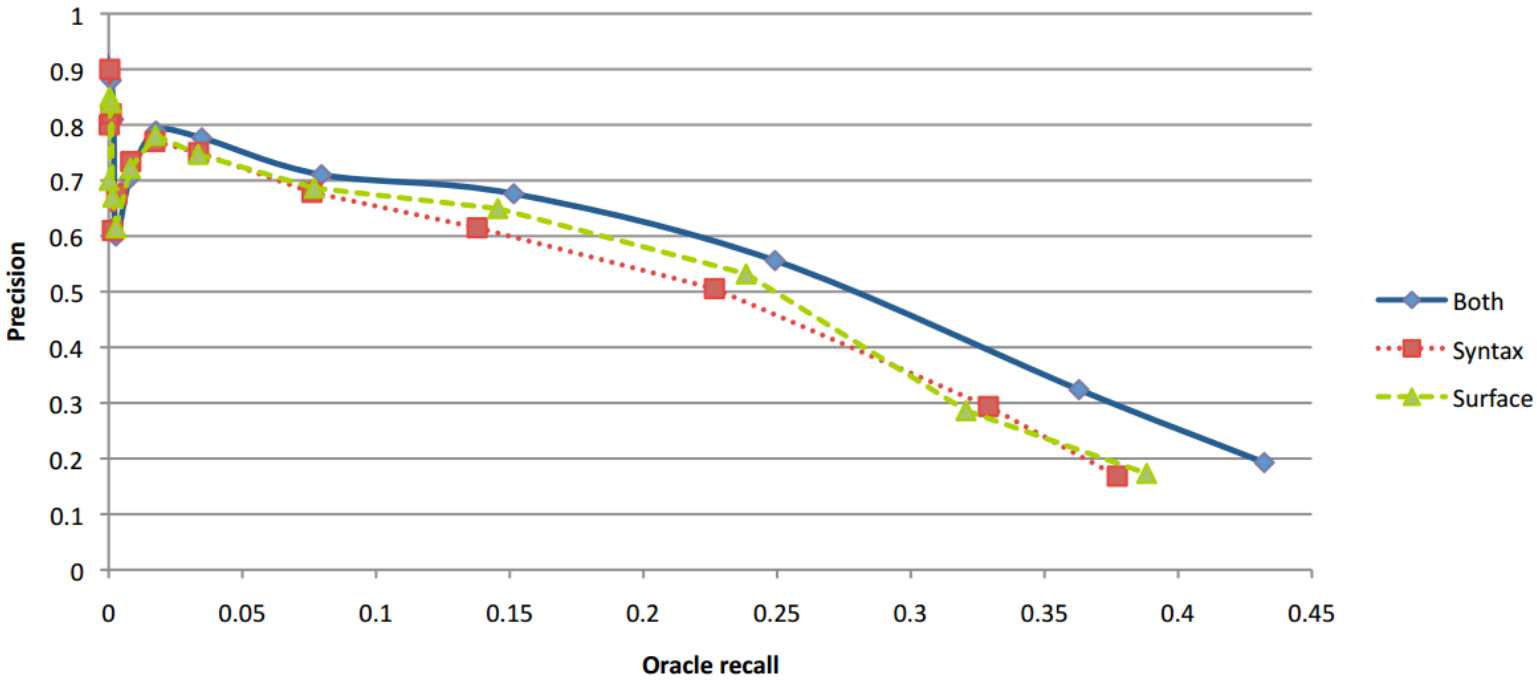
Problems with bootstrapping (Mintz et al, 2009)

- [Steven Spielberg]'s film [Saving Private Ryan] is loosely based on the brother's story
  (Spielberg could be producer, actor)
- Allison co-produced the award-winning [Saving Private Ryan], directed by [Steven Spielberg]
  (Saving Private Ryan might not be a film)

Distant supervision

- Start with a large set of known relations (e.g. from Freebase)
- Collect all sentences that include both entities in the relation. These are positive training instances.
- Sample negative training instances (for example, sentences that contain one entity in a relation but not both).

Distant supervision performance



The Information Extraction pipeline

- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

The Information Extraction pipeline

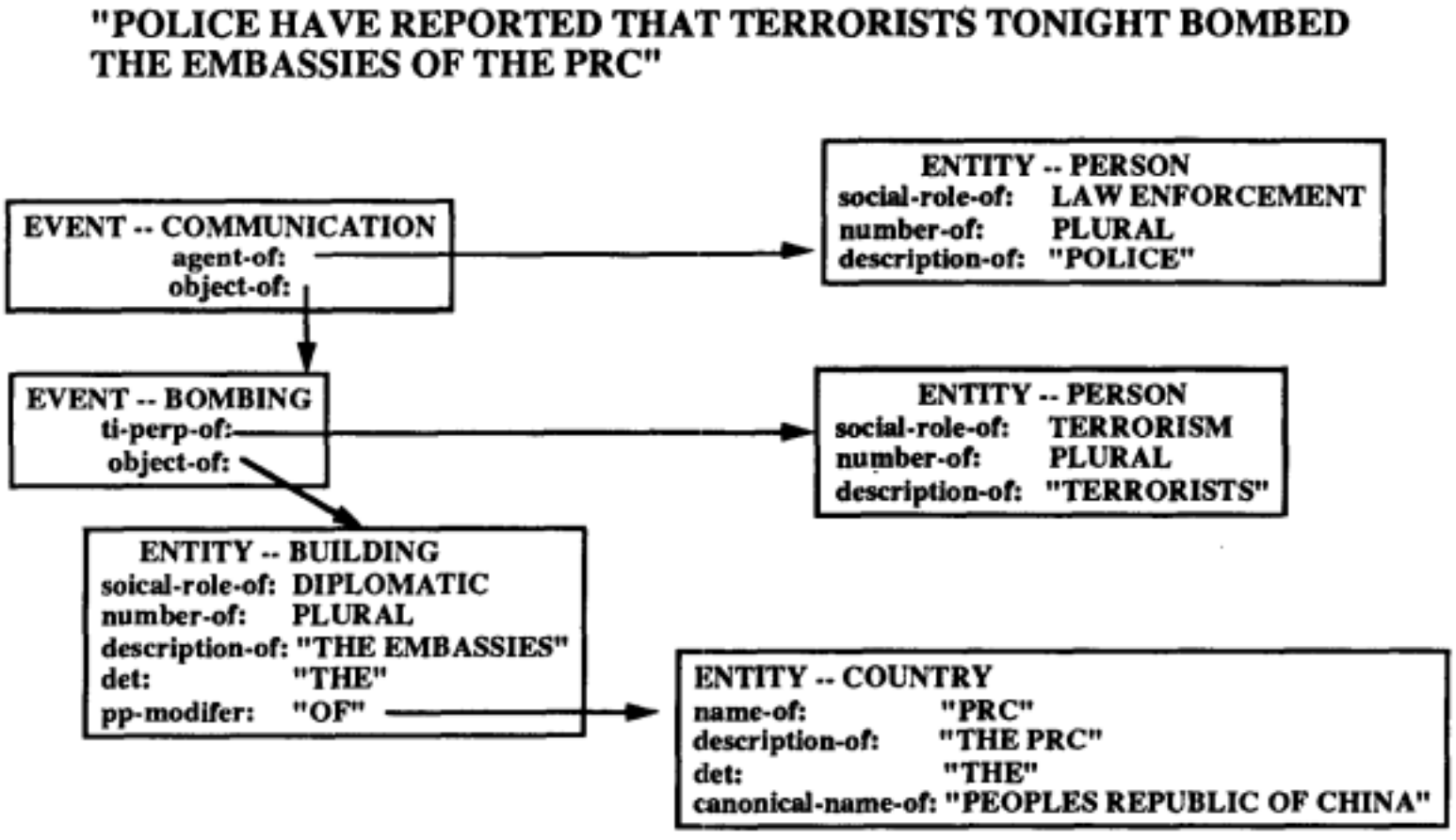
- Unstructured source: At a meeting of the Thirteen, Pyat Pree tells Daenerys he has her dragons in the House of the Undying.
- Annotated entities: At a meeting of <ORG>the Thirteen</ORG>, <PER>Pyat Pree</PER> tells <PER>Daenerys</PER> that he has <OBJ>her dragons</OBJ> in the <PER>House of the Undying</PER>.
- Linked entities:
  - <PER>Pyat Pree</PER> → Pyat Pree
  - <PER>Daenerys</PER> → Daenerys Targaryen
- Relations:
  - Pyat Pree <HAS> Dragons
  - Dragons <LOCATED-IN> House of the Undying
- Events:
  Possession: [Object: dragons; Location: House of the Undying; Possessor: Pyat Pree]

Event extraction

- **Relations** are predications involving two arguments.
- **Events** are predications involving arbitrary numbers of arguments.

Event type	Subtypes
Life	Be-born, Marry, Divorce, Injure, Die
Movement	Transport
Transaction	Transfer-ownership, Transfer-money
Business	Start-org, Merge-org, Declare-bankruptcy, End-org
Conflict	Attack, Demonstrate
Personnel	Start-position, End-position, Nominate, Elect
Justice	Arrest-jail, Release-parole, Trial-hearing Charge-indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

Representing events



Event templates

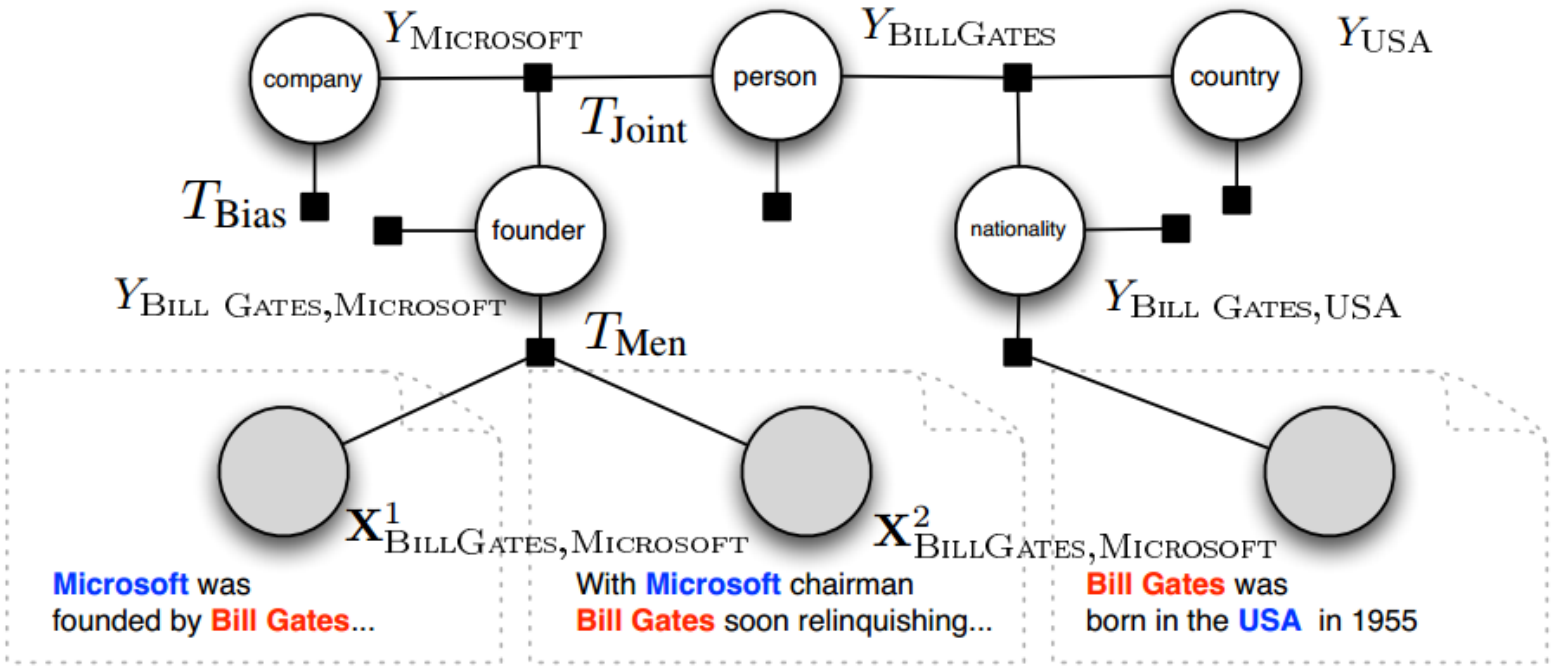
In supervised event extraction, each event type has a template of relevant attributes.

0.	MESSAGE ID	TST1-MUC3-0099
1.	TEMPLATE ID	1
2.	DATE OF INCIDENT	- 25 OCT 89
3.	TYPE OF INCIDENT	BOMBING
4.	CATEGORY OF INCIDENT	TERRORIST ACT
5.	PERPETRATOR: ID OF INDIV(S)	"TERRORISTS"
6.	PERPETRATOR: ID OR ORG(S)	-
7.	PERPETRATOR CONFIDENCE	-
8.	PHYSICAL TARGET: ID(S)	"THE EMBASSIES"
9.	PHYSICAL TARGET: TOTAL NUM	PLURAL
10.	PHYSICAL TARGET: TYPE(S)	DIPLOMAT OFFICE OR RESIDENCE: "THE EMBASSIES"
11.	HUMAN TARGET: ID(S)	-
12.	HUMAN TARGET: TOTAL NUM	-
13.	HUMAN TARGET: TYPE(S)	-
14.	TARGET: FOREIGN NATIONS	-
15.	INSTRUMENT: TYPE(S)	*
16.	LOCATION OF INCIDENT	EL SALVADOR: SAN ISIDRO (TOWN)
17.	EFFECT ON PHYSICAL TARGET	SOME DAMAGE: "THE EMBASSIES"
18.	EFFECT ON HUMAN TARGET	NO INJURY: "-"

Typical approach: train classifiers for each slot in the template

Collective relation extraction

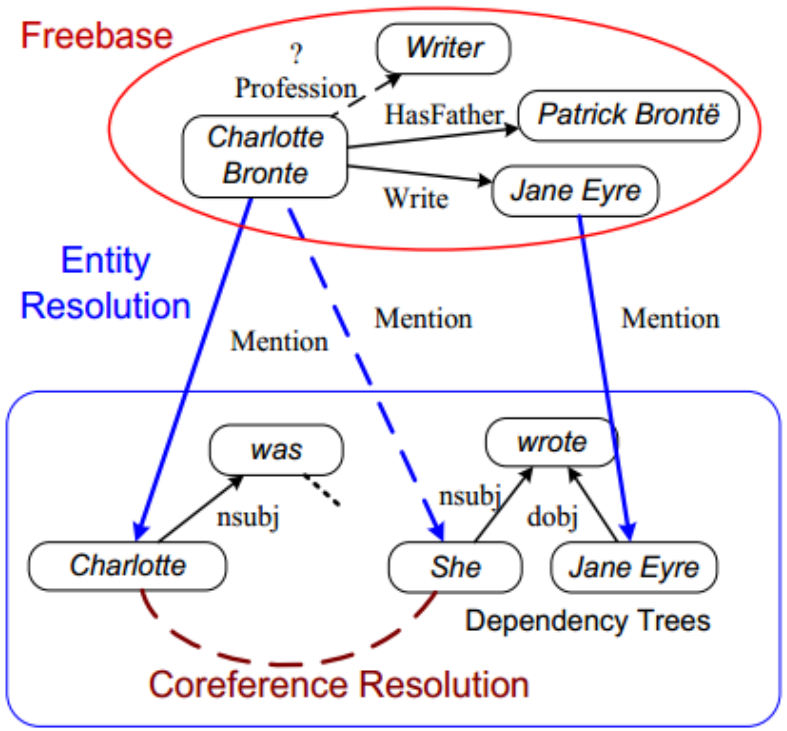
Joint reasoning about both language understanding and the underlying semantics.



(Yao, Riedel, and McCallum, 2010)

Collective relation extraction

Joint reasoning about both language understanding and the underlying semantics.



(Lao, Subramanya, Pereira, and Cohen, 2012)

Next steps: Processes

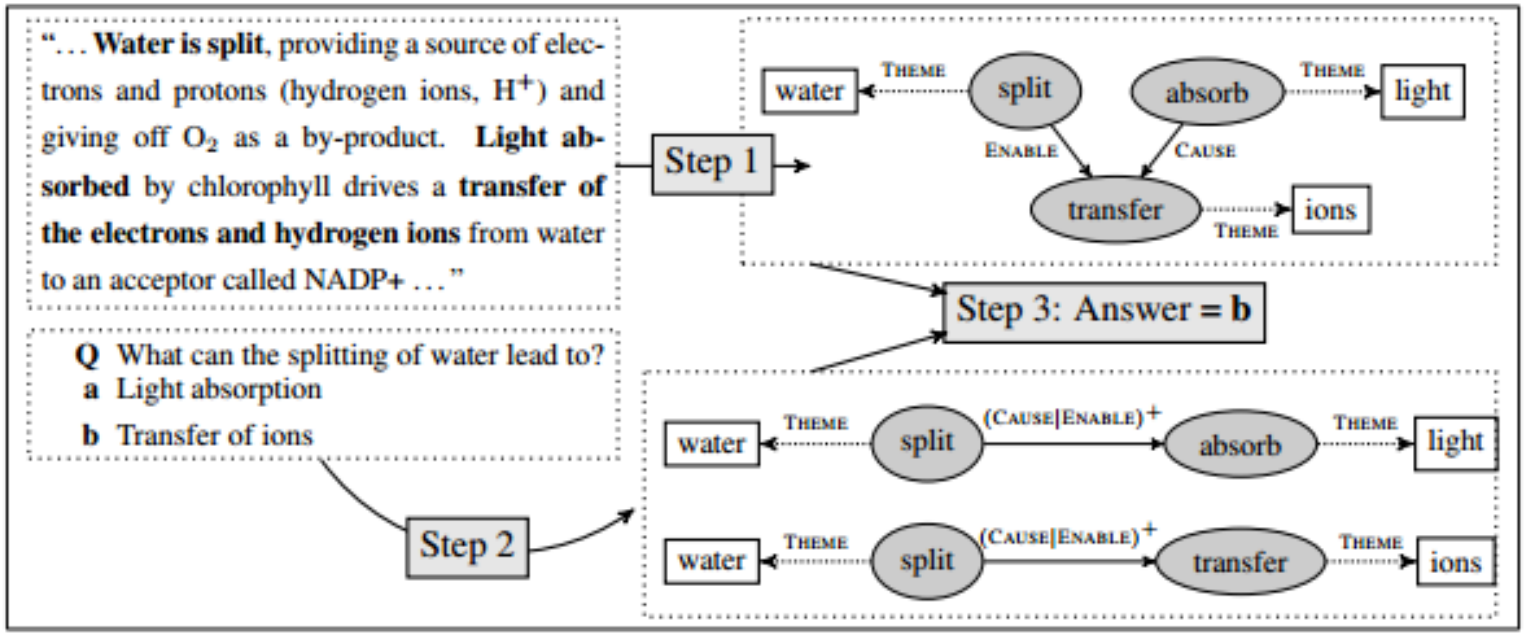


Figure 1: An overview of our reading comprehension system. First, we predict a structure from the input paragraph (the top right portion shows a partial structure skipping some arguments for brevity). Circles denote events, squares denote arguments, solid arrows represent event-event relations, and dashed arrows represent event-argument relations. Second, we map the question paired with each answer into a query that will be answered using the structure. The bottom right shows the query representation. Last, the two queries are executed against the structure, and a final answer is returned.

(Berant et al, 2014)

Processes

Berant et al (2014): a process is a directed graph, involving

Event triggers introduce events, e.g. split. They are nodes in the graph.

Arguments are entities that participate in events. They are nodes, connected to event triggers by edges labeled by semantic role.

Event-event relations are edges between event trigger nodes, including

- **cause**, **enable**, **prevent**, and their disjunctions and conjunctions
- **super**: one event is part of another

Process graph induction is formulated as an integer linear program.

Special cases: time

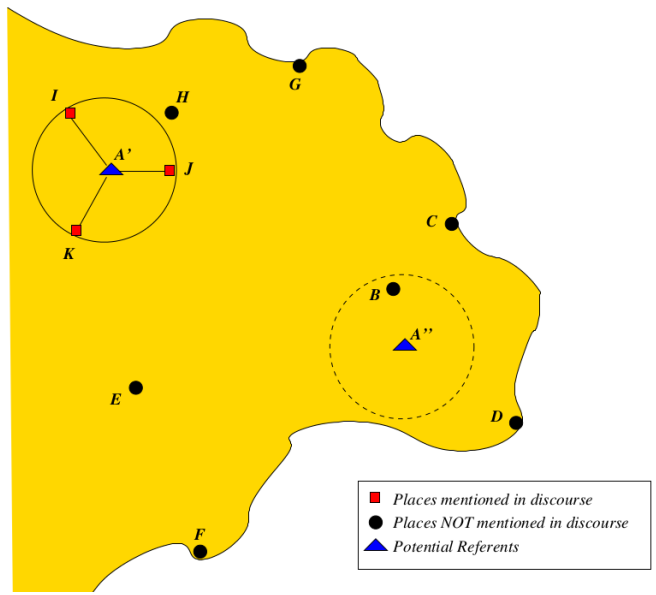
Expression	*Type*	*Value*
October of 1963	DATE	1963-10
October	DATE	2011-10
last Friday	DATE	2011-09-16
next weekend	DATE	2011-W39-WE
the day after tomorrow	DATE	2011-09-21
the nineties	DATE	199X
winter of 2000	DATE	2000-WI
5th century B.C.	DATE	-05XX
now	DATE	PRESENT_REF
Saturday morning	TIME	2011-09-24TMO
4 p.m. Tuesday	TIME	2011-09-20T16:00

- SUTIME is a rule-based system for parsing time expressions.
- Recent work (Agneli et al) has focused on statistical parsing.

Special cases: space

Location descriptions also have a structure that is hierarchical and complex, yet arguably tractable.

- the first gas station after you cross under the 85 on Peachtree Road in Vinings
- on the left after you come up stairs in TSRB
- in the bottom of a drawer in the cabinet opposite the green bookcase



The problem of resolving place names is **toponym resolution**.

Next steps: beliefs and evidence

Possibly factual	United States **may** extend its naval quarantine to Jordans Red Sea port of Aqaba.
Possibly counter-factual	They **may not** have enthused him for their particular brand of political idealism.
Source-specific faculity	Izvestiya **said** that the G-7 leaders **pretended** everything was OK in Russia's economy.
Epistemic marking	He saw the gunman, The editorialist speculated ...

FactBank

FactBank is a corpus of factuality annotations (Saurí and Pustejovsky 2009).

Table 1
FactBank annotation scheme. CT = certain; PR = probable; PS = possible; U = underspecified; + = positive; − = negative; u = unknown.

Value	Definition	Count
CT+	According to the source, it is certainly the case that X	7,749 (57.6%)
PR+	According to the source, it is probably the case that X	363 (2.7%)
PS+	According to the source, it is possibly the case that X	226 (1.7%)
CT−	According to the source, it is certainly not the case that X	433 (3.2%)
PR−	According to the source it is probably not the case that X	56 (0.4%)
PS−	According to the source it is possibly not the case that X	14 (0.1%)
CTu	The source knows whether it is the case that X or that not X	12 (0.1%)
Uu	The source does not know what the factual status of the event is, or does not commit to it	4,607 (34.2%)
		13,460

FactBank Annotations and Modeling

Magna International Inc.'s chief financial officer, James McAlpine, **resigned** and its chairman, Frank Stronach, is stepping in to help turn the automotive-parts manufacturer around, the company said.

Normalization: James McAlpine resigned

Annotations: CT+: 10

In the air, U.S. Air Force fliers say they have **engaged** in "a little cat and mouse" with Iraqi warplanes.

Normalization: U.S. Air Force fliers have engaged in "a little cat and mouse" with Iraqi warplanes

Annotations: CT+: 9, PS+: 1

FactBank Annotations and Modeling

(a) If the heavy outflows **continue**, fund managers will face increasing pressure to sell off some of their junk to pay departing investors in the weeks ahead.

Normalization: the heavy outflows will continue

Annotations: Uu: 7, PS+: 2, CT+: 1

(b) A unit of DPC Acquisition Partners said it would seek to liquidate the computer-printer maker "as soon as possible," even if a merger isn't **consummated**.

Normalization: a merger will be consummated

Annotations: Uu: 8, PS+: 2