

## Applications

November 26, 2009  
CS 486/686  
University of Waterloo

## Outline

- Alchemy applications
- Readings:
  - Marc Summer and Pedro Domingos (2007), *The Alchemy Tutorial*, Department of Computer Science and Engineering, University of Washington

CS486/686 Lecture Slides (c) 2009 P. Pasupat

2

## Information Retrieval

InQuery(word)  
HasWord(page, word)  
Relevant(page)  
Links(page, page)

$\text{InQuery}(+w) \wedge \text{HasWord}(p,+w) \Rightarrow \text{Relevant}(p)$   
 $\text{Relevant}(p) \wedge \text{Links}(p,p') \Rightarrow \text{Relevant}(p')$

Cf. L. Page, S. Brin, R. Motwani & T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Tech. Rept., Stanford University, 1998.

CS486/686 Lecture Slides (c) 2009 P. Pasupat

3

## Record deduplication

**Problem:** Given database, find duplicate records

HasToken(token, field, record)  
SameField(field, record, record)  
SameRecord(record, record)

$\text{HasToken}(+t,+f,r) \wedge \text{HasToken}(+t,+f,r')$   
 $\Rightarrow \text{SameField}(f,r,r')$   
 $\text{SameField}(+f,r,r') \Rightarrow \text{SameRecord}(r,r')$   
 $\text{SameRecord}(r,r') \wedge \text{SameRecord}(r',r'')$   
 $\Rightarrow \text{SameRecord}(r,r'')$

Cf. A. McCallum & B. Wellner, "Conditional Models of Identity Uncertainty with Application to Noun Coreference," in *Adv. NIPS 17*, 2005.

CS486/686 Lecture Slides (c) 2009 P. Pasupat

4

## Record resolution

Can also resolve fields:

HasToken(token, field, record)  
SameField(field, record, record)  
SameRecord(record, record)

$\text{HasToken}(+t,+f,r) \wedge \text{HasToken}(+t,+f,r')$   
 $\Rightarrow \text{SameField}(f,r,r')$   
 $\text{SameField}(+f,r,r') \Leftrightarrow \text{SameRecord}(r,r')$   
 $\text{SameRecord}(r,r') \wedge \text{SameRecord}(r',r'')$   
 $\Rightarrow \text{SameRecord}(r,r'')$   
 $\text{SameField}(f,r,r') \wedge \text{SameField}(f,r',r'')$   
 $\Rightarrow \text{SameField}(f,r,r'')$

More: P. Singla & P. Domingos, "Entity Resolution with Markov Logic", in *Proc. ICDM-2006*.

CS486/686 Lecture Slides (c) 2009 P. Pasupat

5

## Information Extraction

- **Problem:** Extract database from text or semi-structured sources
- **Example:** Extract database of publications from citation list(s) (the "CiteSeer problem")
- **Two steps:**
  - **Segmentation:**  
Use HMM to assign tokens to fields
  - **Record resolution:**  
Use logistic regression and transitivity

CS486/686 Lecture Slides (c) 2009 P. Pasupat

6

## Information Extraction

```
Token(token, position, citation)
InField(position, field, citation)
SameField(field, citation, citation)
SameCit(citation, citation)

Token(+t,i,c) => InField(i,+f,c)
InField(i,+f,c) <=> InField(i+1,+f,c)
f != f' => (!InField(i,+f,c) v !InField(i,+f',c))

Token(+t,i,c) ^ InField(i,+f,c) ^ Token(+t,i',c')
  ^ InField(i',+f,c') => SameField(+f,c,c')
SameField(+f,c,c') <=> SameCit(c,c')
SameField(f,c,c') ^ SameField(f,c',c'') => SameField(f,c,c'')
SameCit(c,c') ^ SameCit(c',c'') => SameCit(c,c'')
```

CS4386/686 Lecture Slides (c) 2009 P. Poupart

7

## Information Extraction

```
Token(token, position, citation)
InField(position, field, citation)
SameField(field, citation, citation)
SameCit(citation, citation)

Token(+t,i,c) => InField(i,+f,c)
InField(i,+f,c) ^ !Token(" ",i,c) <=> InField(i+1,+f,c)
f != f' => (!InField(i,+f,c) v !InField(i,+f',c))

Token(+t,i,c) ^ InField(i,+f,c) ^ Token(+t,i',c')
  ^ InField(i',+f,c') => SameField(+f,c,c')
SameField(+f,c,c') <=> SameCit(c,c')
SameField(f,c,c') ^ SameField(f,c',c'') => SameField(f,c,c'')
SameCit(c,c') ^ SameCit(c',c'') => SameCit(c,c'')
```

More: H. Poon & P. Domingos, "Joint Inference in Information Extraction", in *Proc. AAAI-2007*.

CS4386/686 Lecture Slides (c) 2009 P. Poupart

8

## Next Class

- Lifted inference

CS4386/686 Lecture Slides (c) 2009 P. Poupart

9