Coreference Resolution in Wikipedia

Alan Ritter

December 4, 2007

Introduction

- Example from the Wikipedia article on **Martha Stewart**:

arg1	*she*
predicate	*was accused of*
arg2	*insider trading*

Introduction

- Example from the Wikipedia article on **Martha Stewart**:

arg1	*she*
predicate	*was accused of*
arg2	*insider trading*

- Why focus on Wikipedia?
  - High quality
  - Semantic information (categories, infoboxes, etc. . .)
  - Large, and constantly growing.

Outline

Introduction

What to Resolve?

Wikipedia Categories

Method

Results

Conclusions

Pronoun replacement - Examples

- From the page on **Ernest Hemingway**
  - *he*
  - **suffered**
  - *significant memory loss*
- From the page on **Francis Ford Coppola**
  - *he*
  - **studied**
  - *theater*
  - **at**
  - *Hofstra University*
- From the page on **James K. Polk**
  - *he*
  - **oversaw**
  - *the opening of the U.S. Naval Academy*

"The X" replacement - Examples

- From the page on **Eureka, Missouri**
  - *the city*
  - **had**
  - *a total population of 7,676*
- From the page on **Amazon River**
  - *the river*
  - **divides into**
  - *two main streams*
- From the page on **Communist Party of China**
  - *the party*
  - **was massacred at**
  - *the hands of the Kuomintang*

Outline

Introduction

What to Resolve?

Wikipedia Categories

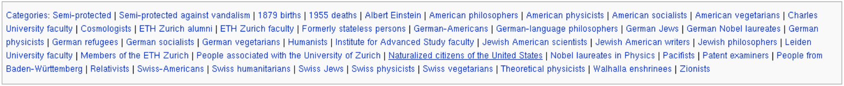
Method

Results

Conclusions

Wikipedia Categories

- Articles are tagged with ”categories”
- From the article on **Albert Einstein**:
  - American physicists
  - People associated with the University of Zurich
  - Nobel laureates in Physics



Categories: Semi-protected | Semi-protected against vandalism | 1879 births | 1955 deaths | Albert Einstein | American philosophers | American physicists | American socialists | American vegetarians | Charles University faculty | Cosmologists | ETH Zurich alumni | ETH Zurich faculty | Formerly stateless persons | German-Americans | German-language philosophers | German Jews | German Nobel laureates | German physicists | German refugees | German socialists | German vegetarians | Humanists | Institute for Advanced Study faculty | Jewish American scientists | Jewish American writers | Jewish philosophers | Leiden University faculty | Members of the ETH Zurich | People associated with the University of Zurich | Naturalized citizens of the United States | Nobel laureates in Physics | Pacifists | Patent examiners | People from Baden-Württemberg | Relativists | Swiss-Americans | Swiss humanitarians | Swiss Jews | Swiss physicists | Swiss vegetarians | Theoretical physicists | Walhalla enshrinees | Zionists

Problems with Wikipedia Categories

- Very messy
- Lots of “administrative categories”
  - Semi-protected against vandalism
  - Articles lacking sources from November 2006
  - Danish election stubs
- For coreference resolution, I just want to know whether or not a particular entity is a person, company, etc...
- Need to map Wikipedia categories to WordNet...

Problems with Wikipedia Categories

- Very messy
- Lots of "administrative categories"
    - Semi-protected against vandalism
    - Articles lacking sources from November 2006
    - Danish election stubs
- For coreference resolution, I just want to know whether or not a particular entity is a person, company, etc...
- Need to map Wikipedia categories to WordNet...
- Somebody has already done this!
    - YAGO[1]

Outline

Introduction

What to Resolve?

Wikipedia Categories

Method

Results

Conclusions

Overview

1. Gather all extractions from article T where:
   - arg1 $=$ (he$|$she) **and** IS-A(T,person)
   - arg1 $=$ "the X" **and** IS-A(T, X)
2. Estimate gender of page's title (if IS-A(T, person))
3. Train a classifier to filter out mistakes
   - Gather features
   - Hand label data

Features

- **genderAgrees**
  - When resolving he|she check if gender agrees
  - Estimate gender by counting ratio of *he* to *she*
- **salient**
  - How recently the title entity has been mentioned?
- **position**
  - Position of the sentence in the Wikipedia page

Outline

Introduction

What to Resolve?

Wikipedia Categories

Method

Results

Conclusions

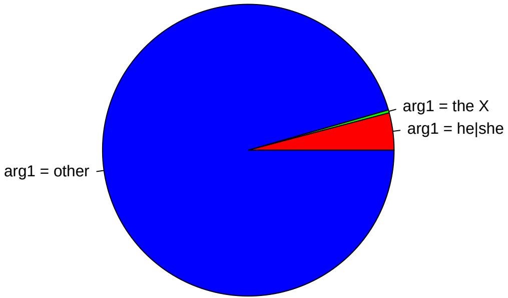
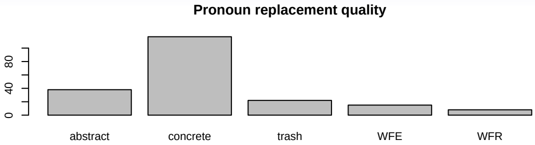
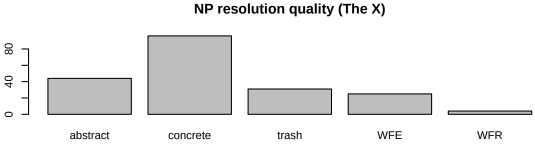


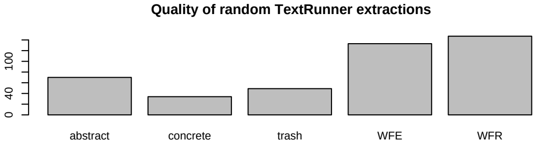
Figure: Percentage of candidate TextRunner extractions from Wikipedia



Pronoun replacement quality



NP resolution quality (The X)



Quality of random TextRunner extractions

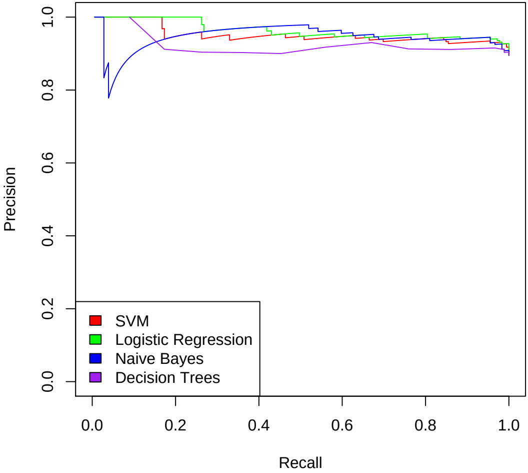


Figure: Precision/Recall for Pronouns (10 fold cross validation)

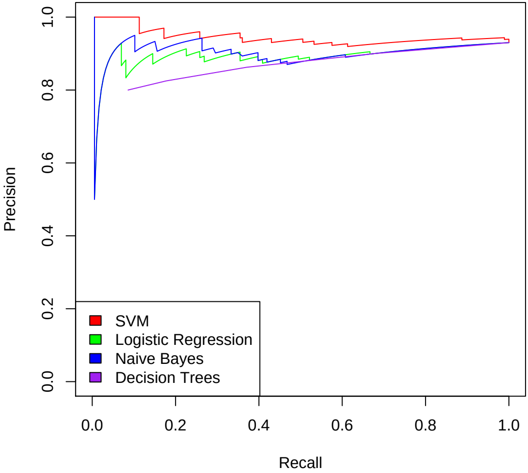


Figure: Precision/Recall for “the X” (10 fold cross validation)

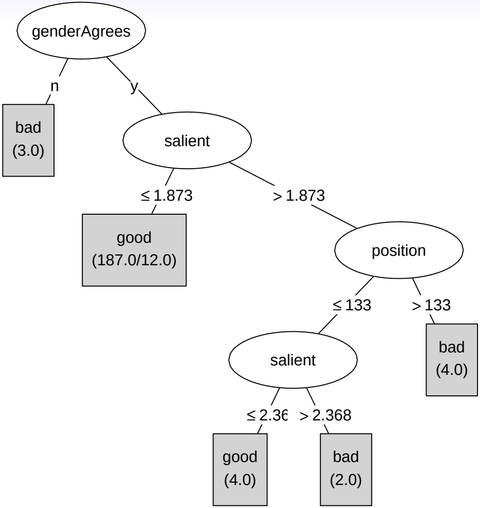


Figure: Decision Tree

Outline

Introduction

What to Resolve?

Wikipedia Categories

Method

Results

Conclusions

Conclusions

- Easier than the standard coreference resolution problem
- $\approx$ 5% improvement in recall
- Higher quality than random TextRunner extractions

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum.
Yago: a core of semantic knowledge.
In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM Press.

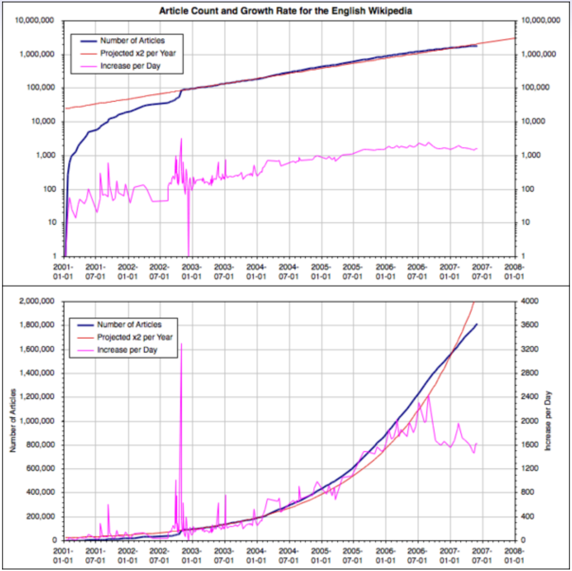


Figure: Growth of English Wikipedia

Traditional Anaphora Resolution

- Considers all noun phrases before an anaphor as antecedent.
- Typical sources of information:
  - Gender
  - Number (plural or singular)
  - High level syntactic/semantic rules (use a parser)