Language Models for Automatic Hypernym Discovery

September 28, 2007

Introduction

- Task:
  - Automatically discover hypernyms from text corpora (the web)
  - Find one or more hypernym for an arbitrary entity
- Problem:
  - WordNet combined with existing methods produces high precision, but low recall
  - At 90% precision:
    - Proper Nouns: 40%
    - Common Nouns: 55%
- Goal:
  - Improve recall
- **This is still work in progress**

Examples

- **Microsoft Chairman Bill Gates** *is a* **Executive**
- **QVC** *is a* **Home Shopping Channel**
- **Death Star** *is a* **Enron Trading Strategy**

Why are Hypernyms important?

- Inference [3]
    - Every **fish** swims $\Rightarrow$ Every **shark** swims
- Web Search
    - "[**U.S. Company**] reported increased profits"
- Lots of other uses in NLP
    - Document clustering
    - Topic Identification[2]

Outline

Hearst patterns

- NP_X and/or other NP_Y
- NP_X is a NP_Y
- NP_Y such as NP_X
- NP_Y including NP_X
- NP_Y, especially NP_X

Limitations of Hearst patterns

- Not 100% accurate

entity	*class*	*sentence*
London	world	...all over the **world** including **London**...
neck	body	...the entire **body** including the **neck**...
patient	candidate	...**candidates** for other treatments, such as **patients** who ...

- Not every hypernym pair will occur in one of these patterns
- We get high precision, and low recall

WordNet contains lots of hypernym relations

- Electronic dictionary
- Contains relations between "synsets"
- Good coverage of common nouns
- General Purpose

Some problems with WordNet

- Manually Constructed
- Lacks coverage of proper nouns
- Lacks domain-specific information

Examples:

- Don't occur in any Hearst patterns or WordNet.
  - 117 Million web page corpus
- Pope Gregory XIII
  - When **Pope Gregory XIII** implemented the Gregorian calendar in 1582, the New Year's celebration was switched to January 1.
- Buckner Bay
  - On October 9, **Buckner Bay** was filled with ships at anchor.
- King Edmund
  - **King Edmund** was called Ironside for his valor.

Efforts to Extend WordNet Using Corpora

- Snow, Jurafsky, Ng [7]
  - Added 10,000 new synsets to WordNet at 84% precision
- Caraballo [1]
  - Built a hypernym hierarchy from scratch without using WordNet
  - Low precision but high recall (39% precision, 60% recall)
- YAGO [8]
  - Used Wikipedia categories to add entities to WordNet

Outline

Coordinate Terms

- Coordinate terms are a pair of words which share a hypernym
  - Example: *car*, *bike*
- If two terms are semantically similar, then it is likely that they share a hypernym.
- If W_1 is similar to W_2, then $H(W_1, C) \Rightarrow H(W_2, C)$
- **Using coordinate terms we can find hypernyms without relying on lexio syntactic patterns.**

Previous sources of coordinate term evidence

1. Coordination patterns (Roark, Charniak [5])
   - *planes, trains and automobiles*
2. Context vectors/distributional similarity [7]

Outline

REALM[6]

- Assesses the correctness of sparse extractions using unsupervised language models
- HMM-T
  - Type checking using Hidden Markov Models
  - Example:
    - `Headquartered(Intel, Santa Clara)`
- REL-GRAMS

HMM-T Features

- Train an HMM over the corpus
- Compute hidden state distributions
- Use these hidden state distributions as features for classification, or a distance measure
    - Example: $P(\vec{s}|mandolin) =< 0.1, 0.5, 0.2, 0.2 >$
- HMM-T data:
    - 216,073 noun phrases
    - Vector of 20 probabilities for each NP

Can HMM-T features separate two classes?

- Cities and People
  - Label and select *city* and *person* NPs using WordNet
  - 685 cities
  - 303 people (includes common nouns)
- Experiments:
  - Singular Value Decomposition + Visualization
  - Hierarchical Clustering

Singular Value Decomposition

- Perform SVD on the HMM-T data (all 216,073 NPs)
- Plot Cities and People using the first two singular vectors and see how well they are separated

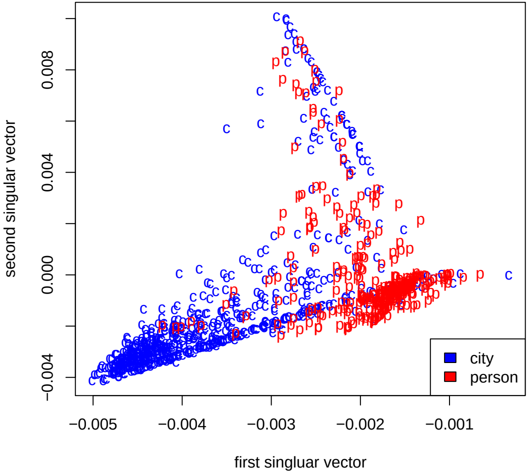

Figure: Singular Value Decomposition

Hierarchical Clustering

- Randomly select 10 cities and 10 people
- Perform Hierarchical clustering on this subset

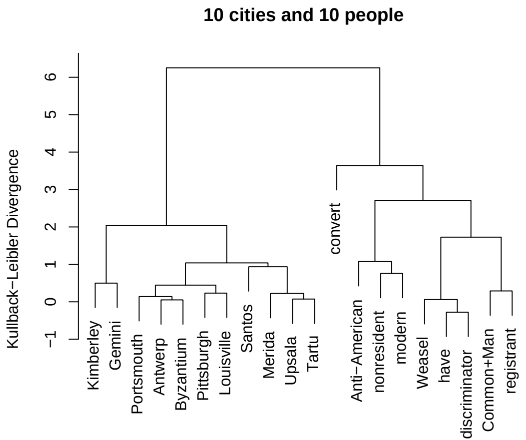

Figure: Hierarchical Clustering

Outline

Semi-Supervised Learning

- Many classes will have few labeled examples
- SSL may be applicable
- Label Propagation
- Does unlabeled data improve classification performance?
- Experiment
  - 685 cities, 303 people
  - 10 labeled examples
  - Label Propagation vs. Supervised techniques

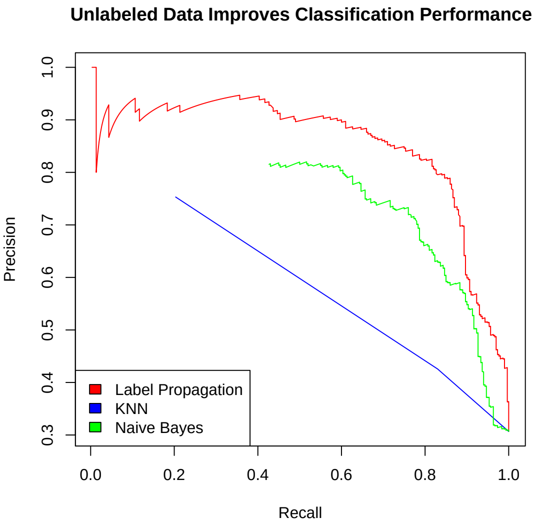

Figure: Label Propagation (10 labeled examples)

Outline

Hearst Pattern Evaluation

- 2 hand labeled data sets:
  - 299 common nouns
  - 396 proper nouns
- Find the 5 best hypernyms for each entity using Hearst patterns.
- Goal: find one or more hypernym for each entity
  - Recall is the fraction of entities for which we find one or more correct hypernym
  - Precision is the fraction of hypernyms which are correctly classified at a given cutoff

Methods

1. Rule Based
    - At least one left, and one right pattern
        - *Cities such as Seattle*
        - *Seattle and other Cities*
    - Existentially quantified in less than 50% of extractions
2. SVM Classifier
    - Trained using WordNet
    - Features:
        - Total number of times the pair appears in any Hearst pattern
        - Total number of left/right patterns
        - Number of times it's existentially quantified
        - Total number of *is a* extractions

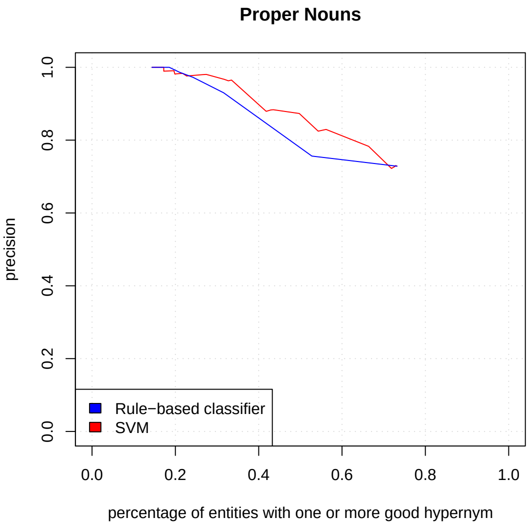

Figure: Precision/Recall on Proper Nouns (Including WordNet)

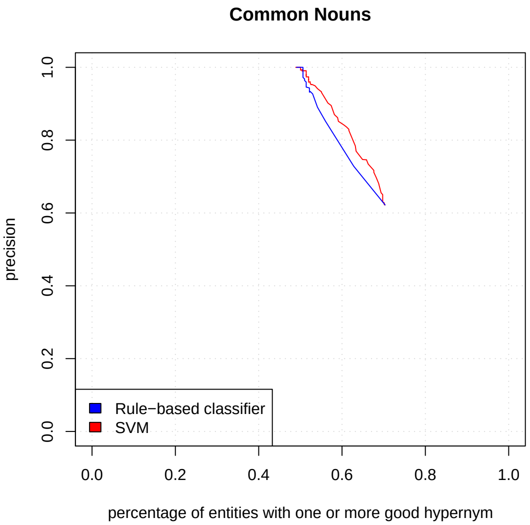

Figure: Precision/Recall on Common Nouns (Including WordNet)

Outline

Conclusions

- Using WordNet and Hearst patterns we can find at least one good hypernym for 40% of proper nouns and 55% of common nouns.
- By discovering coordinate relations we can improve recall
  - Find hypernyms for entities which don't appear in any Hearst patterns or WordNet
- HMM-T features work well for features to a distance metric or classifier
- Semi-supervised learning appears applicable

Future Work

- Augment baseline with results from coordination patterns
- Compare precision/recall between:
  - WN + Hearst Patterns + Coordination Patterns
  - WN + Hearst Patterns + HMM-T
  - WN + HP + CP + HMM-T
- Semi Supervised Learning

Sharon A. Caraballo.
Automatic construction of a hypernym-labeled noun hierarchy from text.
In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

Chris Clifton and Robert Cooley.
Topcat: Data mining for topic identification in a text corpus.
In *Principles of Data Mining and Knowledge Discovery*, pages 174–183, 1999.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning.
Learning to recognize features of valid textual entailments.
In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*,

pages 41–48, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

- Simone Paolo Ponzetto and Michael Strube.
  Deriving a large-scale taxonomy from wikipedia.
  In *AAAI*, pages 1440–1445, 2007.

- Brian Roark and Eugene Charniak.
  Noun-phrase co-occurence statistics for semi-automatic semantic lexicon construction.
  In *COLING-ACL*, pages 1110–1116, 1998.

- Doug Downey Stefan Schoenmackers and Oren Etzioni.
  Sparse information extraction: Unsupervised language models to the rescue.
  In *ACL '07*, 2007.

- Rion Snow, Daniel Jurafsky, and Andrew Y. Ng.
  Semantic taxonomy induction from heterogenous evidence.
  In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th*

annual meeting of the ACL, pages 801–808, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

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.

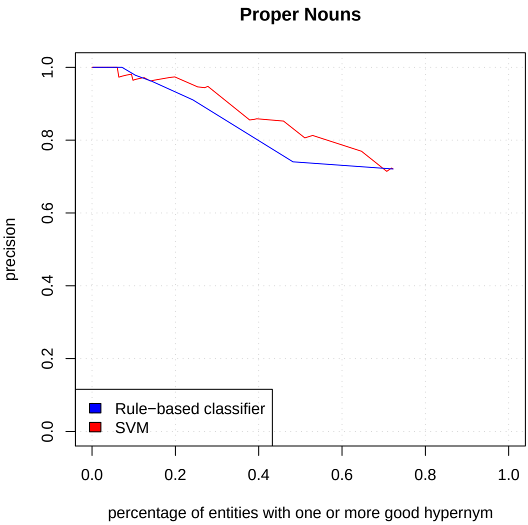

Figure: Precision/Recall on Proper Nouns

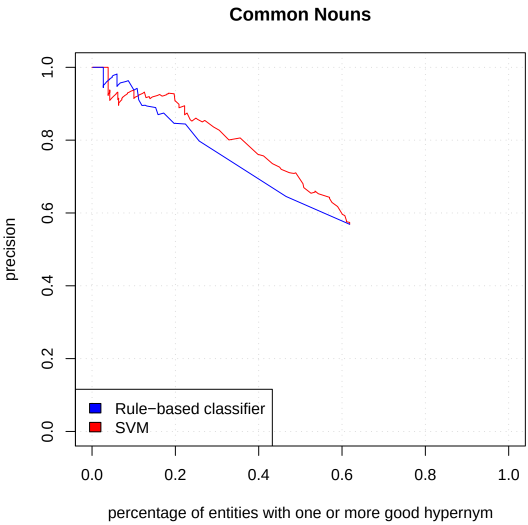

Figure: Precision/Recall on Common Nouns