On Extracting IS-A pairs

Presented by
-Sujatha Das

Outline for this talk

- Introduction/Problem Statement
  - What are IS-A pairs?
- Previous Approaches
  - Learning syntactic patterns
- Our Method
  - Corpus-based features
- Preliminary Results with Sub-topic Extraction
- Conclusions

Hypernym Classification

- (p,t) is an IS-A pair if *p is-a t*

WordNet terminology: p is the hyponym of t and t is the hypernym of p

- Examples
  - *apple is-a fruit ,U2 is-a "rock band"*

    "support-vector-machines" is-a "machine learning technique"
  - *Negative pairs*
    - *(apple, orange)*
    - *(Europe, country)*
    - *(feature dimensions, machine learning)*

Where is IS-A information used?

- Automated fact extraction systems (Ex. KnowItAll [1])
  - Database of (instance, class) pairs
  - Example classes such as actor, country, capital cities etc.
- Question Answering Systems
  - Verifying if the retrieved passage is a potential answer

    Ex: Bono is the lead vocalist for which Irish rock band?

Previous and Related Work

Hearst [2] Patterns

- such as, including, and other, etc.

"...works by such authors as Herrick, Goldsmith, and Shakespeare..."

- *Herrick is-a author,*
- *Goldsmith is-a author*
- *Shakespeare is-a author*

- Bootstrap with known pairs to learn patterns

Hearst patterns used in KnowItAll

1. NP1 {“,”} “such as” NPList2
2. NP1 {“,”} “and other” NP2
3. NP1 {“,”} “including” NPList2
4. NP1 “is a” NP2
5. NP1 “is the” NP2 “of” NP3
6. “the” NP1 “of NP2 “is” NP3

Related Work (contd.)

- KnowItAll [1]
  - Generalize to any relation (class)
  - Bootstrap with known pairs
    - Learn discriminative phrases per class
    - Also apply noun phrase and orthographic constraints depending on the class
  - Example:
    - China is a country in Asia
    - Garth Brooks is a country singer

    Noun phrase on the left should be capitalized and country should be the head of NP on the right

KnowItAll (contd.)

- Learn discriminative phrases based on co-occurrence statistic for each class
- Validate extracted instances with discriminative phrases
  - Example, For the class “actor”, “starred in” is a good discriminative phrase

  Assign a confidence score to the extracted instance “Cuba Gooding” using hit counts on the Web for “Cuba Gooding starred in”

Learning Extraction Patterns

- Snow, et al. [3]
- Use known pairs to extract sentences containing both the phrases from a large collection
- Parse the sentences (Dependency parse)
  - Use parse features to train a classifier (feature lexicon: ~69500 patterns)
  - Usually good patterns have dependency paths of length <=4 in the parse tree
  - Example: the Hearst pattern NP_x and other NP_y

$$(and, U : PUNC : N), -N : CONJ : N, (other, A : MOD : N)$$

Learning Extraction Patterns

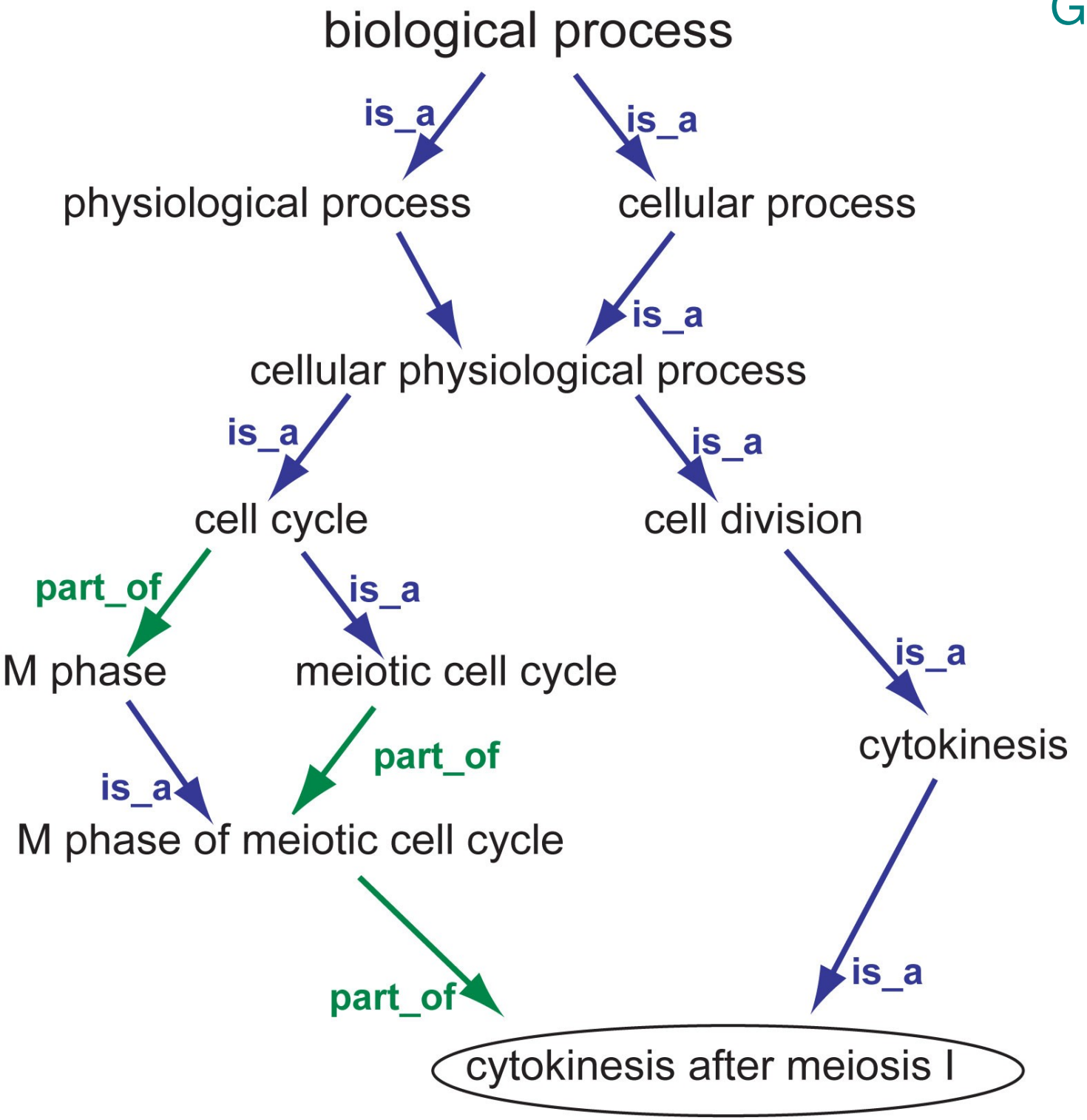
- Ritter, et al. [4]
- Low recall with Hearst patterns due to their local nature
  - Example: “...urban birds in cities such as pigeons...” (pigeon, city)?
- Most errors on right patterns (p to the left/right of t)
  - Number of right pattern matches
  - Fraction of matches which follow determiners...

How good are syntactic patterns?

- Snow, et al. [3]
  - Newswire corpus of 6 million sentences, ~69000 features
  - ~5400 manually annotated noun pairs

Interannotator Agreement:	0.8318
TREC+Wikipedia Hypernym-only Classifier (Logistic Regression):	0.3592
TREC Hybrid Linear Interpolation Hypernym/Coordinate Model:	0.3268
TREC Hypernym-only Classifier (Logistic Regression):	0.2714
Best WordNet Classifier:	0.2339
Hearst Patterns Classifier:	0.1417
"And/Or Other" Pattern Classifier:	0.1386

Gene Ontology [6]



Sense 1 (WordNet Hypernym Tree for Mark Twain)
Clemens, Samuel Langhorne Clemens, Mark Twain
=> writer, author
=> communicator
=> person, individual, someone, somebody, mortal, human, soul
=> organism, being
=> living thing, animate thing
=> object, physical object
=> entity
=> causal agent, cause, causal agency
=> entity
=> humorist, humourist
=> entertainer
=> person, individual, someone, somebody, mortal, human, soul
=> organism, being
=> living thing, animate thing
=> object, physical object
=> entity
=> causal agent, cause, causal agency
=> entity

Other sources of *is-a* evidence?

- Our focus and extension
  - *is-a* in the open-domain vs *is-a-subtopic* in a given domain
  - Examples for CS: (machine learning, artificial intelligence), (routing algorithms, computer networks)
- Motivating application: Query Expansion for Expert Search
  - Identify experts in machine learning
  - Expertise in “support vector machines” counts towards expertise in “machine learning”

Intuition

- Given a representative corpus in a domain, can phrase characteristics reveal the is-a-subtopic connection?
- *p is-a-subtopic t,* p is more specific than t. Let D_w be the set of all documents containing w
  - |D_p| <= |D_t| (D_p \subseteq D_t)
  - idf (p) vs. idf(t)
  - Sim(D_p, D_t)
  - AvgSim(D_p) vs. AvgSim(D_t)
  - PMI-IR(p,t) co-occurrence statistic
- WebHits with P={“such as”, “including”, “and other”, “or other”, “like”, “especially”}

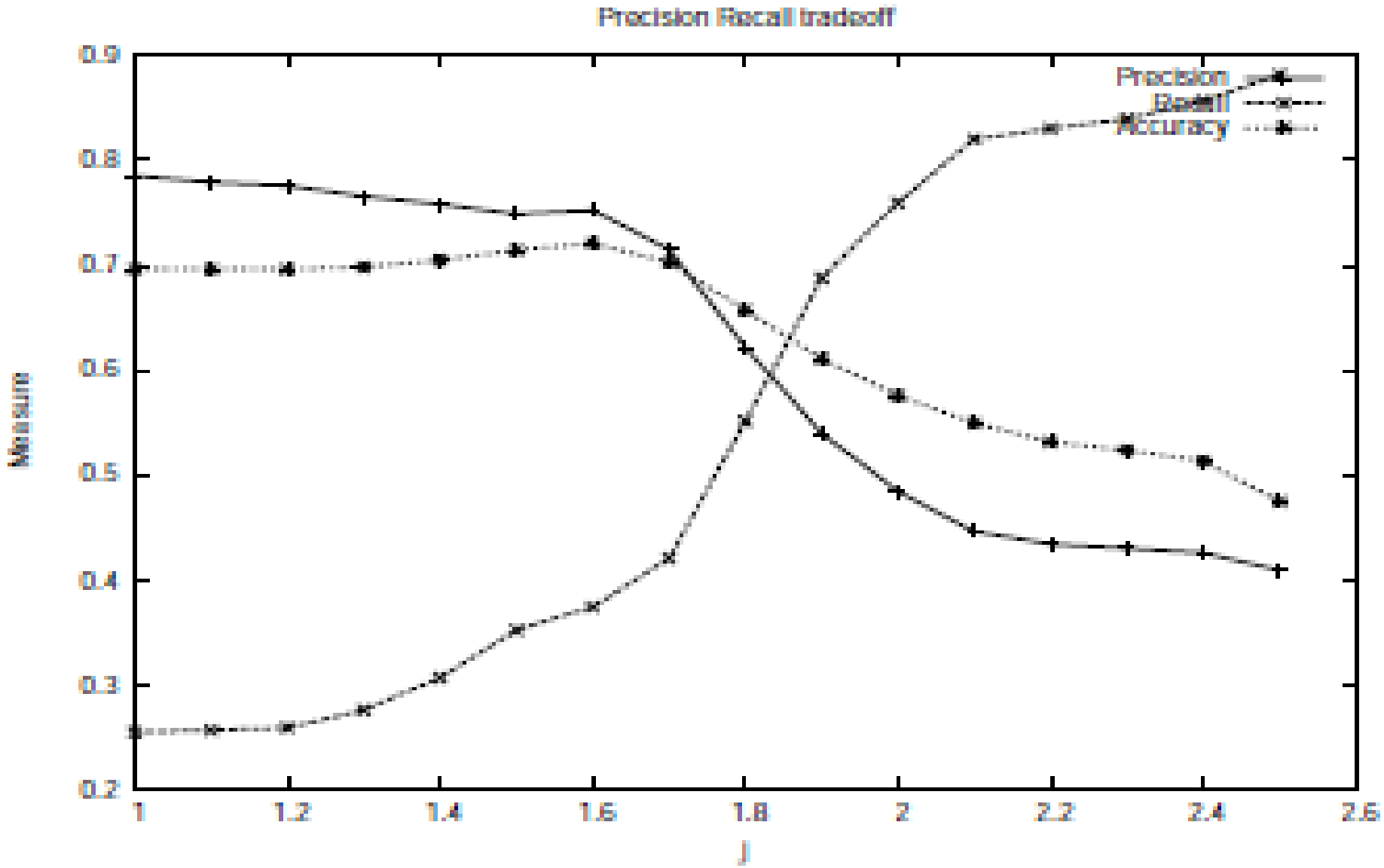
List of features

	Feature description
1	Inverse Document Frequency of p
2	Inverse Document Frequency of t
3	$\|D_p\|/\|D_t\|$
4	Ratio of IDFs of p and t
5	$PMI - IR(p,t)$
6	$Similarity(C_{Topk(D_p)}, C_{Topk(D_t)})$
7	$AverageSimilarity(Topk(D_p))$
8	$AverageSimilarity(Topk(D_t))$
9	$Similarity(First(Topk(D_p)), First(Topk(D_t)))$
10	$Similarity(Last(Topk(D_p)), Last(Topk(D_t)))$
11-16	Are NumWebHits with pattern p > 0 for $p \in P$?
17	Are NumWebHits(p) > 0 for at least one $p \in P$?
18	Are NumWebHits(p) > 0 for at least two $p \in P$?
19	Are NumWebHits(p) > 0 for at least three $p \in P$?
20	Is $\|D_t\| > \|D_p\|$?
21	Is $IDF(p) > IDF(t)$?
22	Is $AvgSim(D_p) > AvgSim(D_t)$?

Experiments

- Datasets
  - ~800 manually tagged (subtopic, topic) pairs in Computer Science domain with CiteSeerX as the representative corpus
  - ~800 extracted pairs from WordNet with abstracts from Wikipedia as the representative corpus
- SVM^{light} implementation from Joachims
  - -j options to trade-off precision and recall
  - Syntactic (11-19) vs. Corpus-based features

Precision Recall trade-off during training



Cross-validation Results

CS Dataset					
Features	j	Precision	Recall	Acc.	F1
All	1.8	0.6218	0.5513	0.6569	0.5844
Syntactic	2.0	0.6073	0.5327	0.6136	0.5676
Corpus-based	1.8	0.4358	0.8326	0.5307	0.5697
WN Dataset					
All	1.5	0.6095	0.8026	0.6442	0.6928
Syntactic	1.4	0.7450	0.6469	0.5849	0.6925
Corpus-based	1.2	0.6184	0.7108	0.6421	0.6614

How to combine features?

Phrase Pair	True	A	S	C
(machine learning, training examples)	N	N	N	Y
(pervasive computing, web services)	N	N	Y	N
(data warehousing, view maintenance)	Y	N	N	Y
(security and privacy, access control)	Y	Y	Y	N

Query Expansion

- We now have an is-a-subtopic classifier for tagging (p, t) pairs
- For query expansion, given a topic query t, we need to find phrases to feed the classifier
- Heuristic algorithm for extracting potential expansion phrases

Relevance Feedback (style)

- Use input query, q to obtain top D_q (first round of retrieval)
- Run our algorithm on D_q to obtain potential phrases P_q
- Tag (p, q) pairs in P_q with the is-a-subtopic classifier
- Use positive pairs from previous step to form q' = q \union P'_q
- Use query q' for second round of retrieval

Extracting phrases

Algorithm 1 Obtaining potential pairs for classification

Input: t, top-k documents using the query D_t,
for $d \in D_t$ **do**
  Obtain the set of n-grams, P from D_t.
  Apply stopword, numeric, special character and phrase similarity filters to remove undesirable n-grams in P.
  for $p \in P$ **do**
    output (p,t) as potential pair to the 'is-a-subtopic' classifier.
  end for
end for

Some examples

Original Query	Expansion Terms
Information Extraction	{ wrapper induction, tree automata, seed tuples, extraction rules, scalable information extraction} { extract, inform, system, text, document, process, web, pattern, gener}
Support Vector Machines	{ combination of kernels, procrustes kernel, kernel functions, feature selection} { vector, machin, support, neural, classif, method, network, featur}
Machine Learning	{ data mining, decision tree, statistical machine learning, artificial intelligence, support vector} { learn, machin, data, research, analysi, gener, system, statist, process }

Expert Search (Deng et al. [5])

- Outline of scoring process
  - Use query to retrieve top-k documents
  - Extract authors from these documents
  - Score each author using P(a,q) values

$$P(a,q) = \sum_{d \in D} P(a,q|d)P(d) = \sum_{d \in D} P(a|d)P(q|d)P(d)$$

Query Expansion Results

Dataset of 7 queries used by Deng, et al.

and CiteSeerX corpus

	Pr@10	Rec@10	MAP@10	MRR@10
No Expsn.	0.1286	0.0406	0.2089	0.4571
RELFB-10	0.1143	0.0374	0.1545	0.3452
RELFB-20	0.1286	0.0409	0.1623	0.3571
RELFB-50	0.1429	0.0454	0.2271	0.4524
ISA-TOP5	**0.2000**	**0.0662**	**0.3233**	**0.7429**

Table 5: Expert Retrieval results with different expansion strategies ($|D| = 50$).

Conclusions and to-dos

- Improving is-a classification or how to combine features?
- Can we populate a topic hierarchy?
- Efficiency issues during query expansion
  - N-gram extraction takes time
  - Other means to obtain potential phrases?
- Preliminary results on small set of queries (set of 7 rather general queries)
- Target-aware query expansion

References

- KnowItAll (WWW '04) http://portal.acm.org/citation.cfm?id=988687
- Hearst (ACL '92) http://portal.acm.org/citation.cfm?id=992154
- Snow, et al. (NIPS '05) http://www.stanford.edu/~jurafsky/paper887.pdf
- Ritter, et al. (AAAI '09 symposium) http://turing.cs.washington.edu/papers/ritter_aaai_ss09.pdf
- Deng, et al. (ICDM '08) http://portal.acm.org/citation.cfm?id=1511330
- Gene Ontology

  http://www.yeastgenome.org/help/GO.html

Thank you!

- Questions and suggestions?