A Weakly-supervised Learning Approach to Domain-specific Semantic Tagging

Ruihong Huang & Ellen Riloff
University of Utah

Domain-specific SemTagging

Goal:
is to assign a semantic class label to every noun phrase in a sentence

[A 14yo doxy]ANIMAL owned by [a reputable breeder]HUMAN is being treated for [IBD]DISEASE with [pred]DRUG.

Applications:
* coreference resolution
* word sense disambiguation
* event extraction systems
* question answering technology

State-of-the-Art:
* supervised learning, requires annotated data
* almost no domain-specific annotated dataset exists.

Our Motivation: induce semantic taggers for specific domains with no annotated corpora.

Weakly-supervised Learning

Supervised machine learning
* needs annotated dataset, e.g. text classification, POS tagging, parsing

Weakly-supervised learning
* needs minimal human supervision
* useful for resource-scarce domains

Bootstrapping
* train the initial classifiers using seeds
* bring in its most confident data points
* retrain the classifiers
* the above iterative process goes on

Our Proposed Approach

The framework
1.Train the initial classifiers using all seed words' instances for each semantic class.
2.Use bootstrapping to iteratively improve the performance.

One contradiction
* one seed word implies one semantic tag, but semantic tags are context-based.

One challenge
* exploit the corpus using bootstrapping
  * sustain momentum, avoid semantic drift

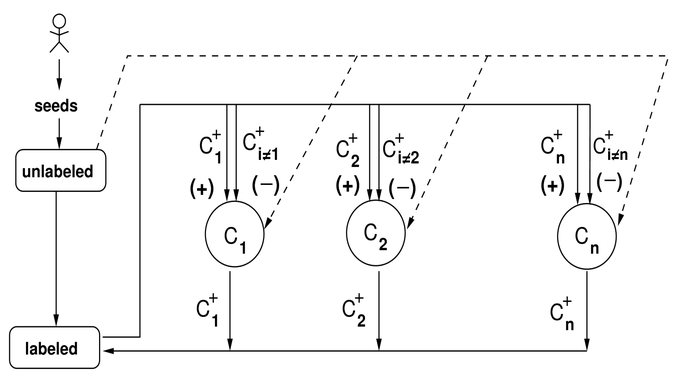

Fig 1: Cross-Category Bootstrapping

Deal with the Contradiction

* Use semantically unambiguous seeds
* Introduce an initial context-only training phase before bootstrapping begins
  * Lexical features for modifiers
  * Lexical features for window contexts

A 3yo tabby was diagnosed with [FELV] after a blood [test] showed that he tested positive.

Feature vector for [FELV] DISEASE:
was−3 diagnosed−2 with−1 after1 test2 showed3

Feature vector for [test] TEST:
with−3 FELV−2 after−1 blood_M showed1 that2 he3

Face the Challenge

* Cross-Category Bootstrapping
  * is shown in Fig 1
  * classifiers help each other to grow stronger
* One Semantic Class Per Discourse
  * bring in all word instances in one discourse
  * increase the context diversity of training set
* Dynamic Semantic Features
  * use class labels of new training instances
  * get more generalized classifiers

Experimental Results

Data set
- Message board posts from the Veterinary Information Network (VIN)
- Over half of the small animal veterinarians in the United States and Canada use VIN

Evaluation
- 22 points of Recall gain while only 4 points of Precision loss and got relatively high F-value

Method	Animal	Disease	Drug	Test	Human	Other	Avg
Seeds	38/100/55	15/99/26	22/97/36	30/94/45	81/99/89	10/93/17	37/98/53.7
Supervised	71/94/81	23/85/36	24/95/38	33/89/48	80/99/89	39/88/54	46/92/61.5
Initial Cs	59/77/67	33/84/47	42/80/55	49/77/59	82/93/87	33/80/47	53/82/64.3
After 39 Iters	86/70/77	60/81/69	69/81/75	73/69/71	86/91/89	50/81/62	75/78/76.6

Table 1: Recall/Precision/F for different settings

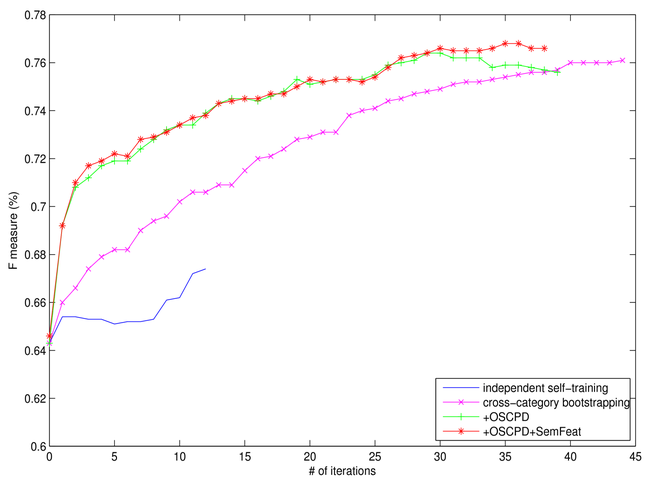

Figure 2: Average F-scores after each iteration

Ellen Riloff: riloff@cs.utah.edu
Ruihong Huang: huangrh@cs.utah.edu

This work is in submission to ACL 2010 conference.

THE UNIVERSITY OF UTAH

References

•1. M. Thelen and E. Riloff. A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts. EMNLP, 2002.
2. A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. COLT, 1998.
3. Imed Zitouni and Radu Florian. Cross-language information propagation for arabic mention detection. TALIP, 2009.
4. T. McIntosh and J. Curran. Reducing Semantic Drift with Bagging and Distributional Similarity. ACL. 2009.