



Semantic Analyses over Sparse Data

Students: Kristiyan Haralambiev, Jerry Liu, Cassia Martin

- Corpus Statistics and Baselines
- Challenges: Feature Selection and Clustering
- Overview: Bootstrapping
- Kris: Wordnet and Accurate Evaluation



Corpus Stats

- Inter-annotator agreement is 94%, so that is the upper limit of our task. (Inclusion of an additional “None of the above” category, which we have deemed unreliable, lowers the agreement to 83%.)
- 214,446 total annotated noun phrases (262,683 with “None of the Above”)
- 29,071 unique vocabulary items (Unlemmatized)
- 25 semantic categories
- 162 subject codes
- 127,569 with semantic category T (Abstract, 59 %)




Baselines on a random held out

- Yasmet: (87.4%)
- Weka, Naïve Bayes: (87%)
- Weka, Voting Feature Interval: (84%)

Yasmet-- Yet Another Small MaxEnt Toolkit
Developed by Franz Och

Weka-- Waikato Environment for Knowledge Analysis
Contains a whole host of machine learning algorithms

- 
- Baselines looked good, right?
 - However, the correctness rating for unseen, ambiguous words with this methods is only 23% (NB: This happens to be a small subset of the training set in question, so the statistical relevance of this figure is low.)
 - Also, our high baseline is partly due to the preponderance of 'T' (Abstract) in our dataset.
 - We wanted to experiment with tagging unseen and ambiguous words, using contextual features.




A new experiment to measure performance on unseen ambiguous words

■ Loose supervision

The testing data is taken to be all words that are unambiguous in the dictionary (that is, they only have one sense marked by the lexicographers.) Those words that were marked as '?' in the training data are thrown away, as though the automatic dictionary annotation was rejected by an annotator.

■ No supervision

The testing data is all unambiguous words. Any question marks in the training data is replaced with the original, unambiguous annotation.



New baselines (using Yasmet)

- Using one previous context word

 - Loose supervision 45%

 - No supervision 44.3%

- Using one word prior and one word after

 - Loose supervision 46.8%

 - No supervision 44.5%

- Much lower!

- NB: automatically tagging all nouns with 'T' (the most common semantic category) yields 52%.

- Therefore, we needed to find more intelligent contextual clues.



Possible non-lexeme features

- Parsing information (CHAOS parser)

 - Preposition (in prepositional phrase)

 - Sub_verb, Obj_verb

 - Modifiers

 - Etc.

- Bag of words on surrounding context

- Dictionary information

 - Possible semCat, subjAreas

- European Wordnet generalizations



Massive data scarcity and overtraining problems

- Features taken from individual noun phrases are too infrequent to generalize effectively to the testing set.

E.g. preliminary Yasmet results using all available parsing data from CHAOS:

- Annotated data 68%
 - Loose supervision 52%
 - No supervision 41%
- We need some clustering!



Clustering

- Clustering adjectives by the category of nouns they tend to modify

e.g. Common adjectives that occur within a noun phrases with Liquid head words are:

- freshwater, acidic, Baltic, territorial, natural, alcoholic

On the other hand, common words like 'hard', 'other', 'great', 'warm' are also quite common

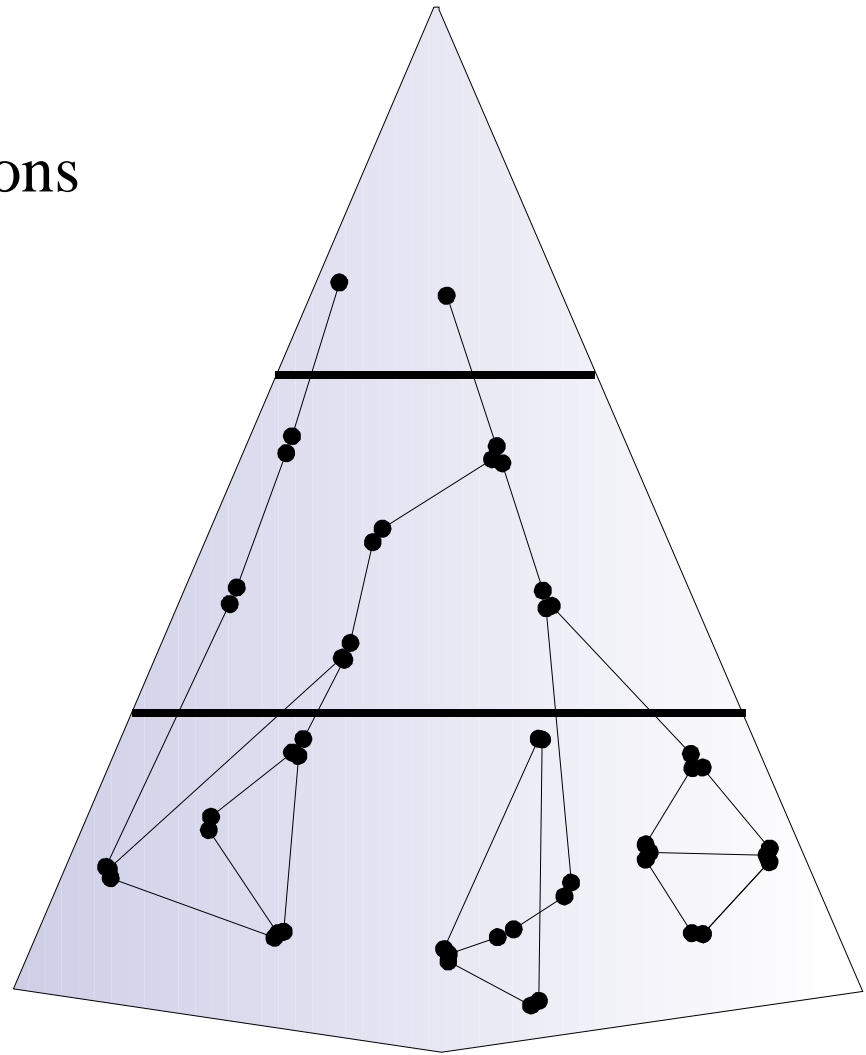
- We are currently working on creating classes which minimize entropy over the semantic categories and measuring mutual information gain over the selected sets.

Clustering by Wordnet

Nouns – 3 top level generalizations
Verbs – about 1000

Mid level – strict hierarchy

Bottom level – synonym
sets (network of words)



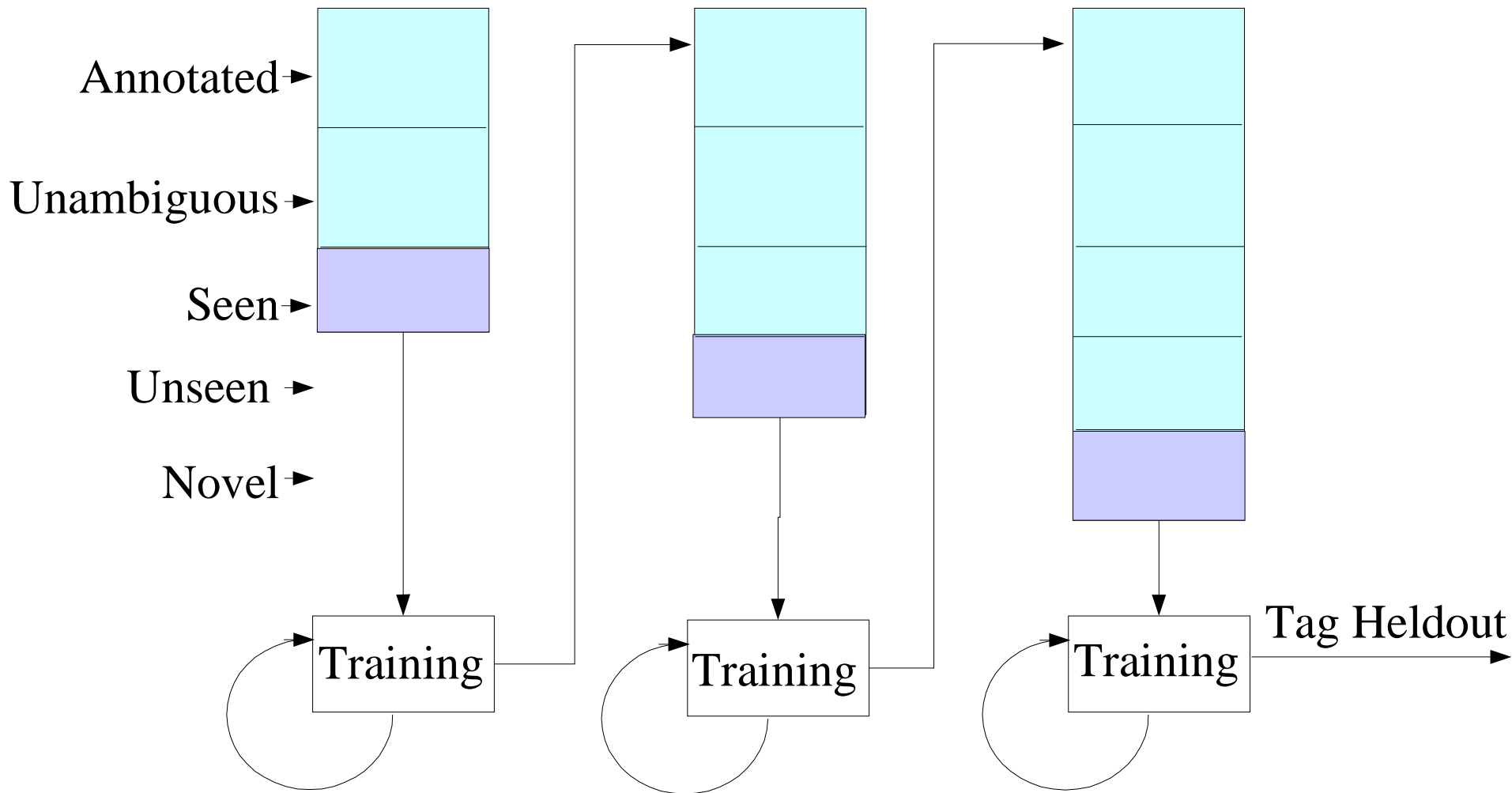


Long term goal/ Bootstrapping

- Once we have a clear idea of what features and clustering techniques work on our data, we wish to extend the approach to a large subset of the BNC (26 million word)

Annotated (Sheffield) corpus

- Seen words in all of BNC
- unseen words in all BNC (in dictionary)
- novel words in BNC
- tag held out data



Unambiguous

Marked as $\{0, 0, \dots, 0, 1\}$

Ambiguous

Marked with appropriate probabilities.
Eg, Novel is $\{1/25, \dots, 1/25\}$