Scaling to Very Very Large Corpora for Natural Language Disambiguation

Michele Banko and Eric Brill

Microsoft Research

Background of Issue

- Compare algorithms
  - Conclusions need common test corpora
    - Fixes the size of training/test sets
- Meanwhile available data growing
- Large cost of annotating data hinders development of new corpora

Confusion Set Disambiguation

- Example sets:
  - {principle, principal}
  - {to, two, too}
  - {weather, whether}
- Key Property #1: disambiguate from a small set of potential values
- Key Property #2: labeled data is available and free

Better Results at Low Cost?

- How to get better performance?
  - Get a Ph.D. and invent a new algorithm
  - Tune parameters and optimize old ways
    - it's easy to fix a bad implementation
  - Why not just train on more data, esp. if it's available?

Details of the Paper

- Learning Methods
  - Perceptron
  - Winnow
  - naïve Bayes
  - Memory (remembers previous and next words)
- Corpus Size
  - 1 Million → 1 Billions words

Learning Curves

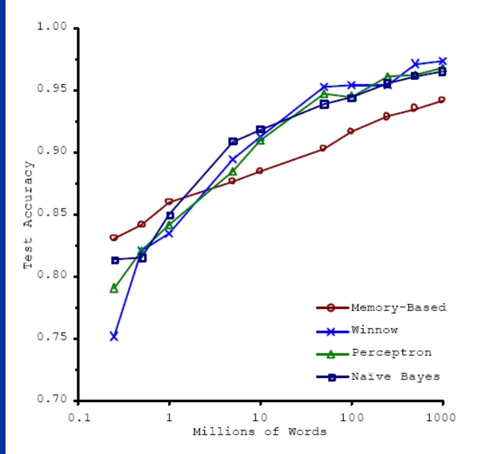

Figure 1. Learning Curves for Confusion Set Disambiguation

Cost of Larger Corpus

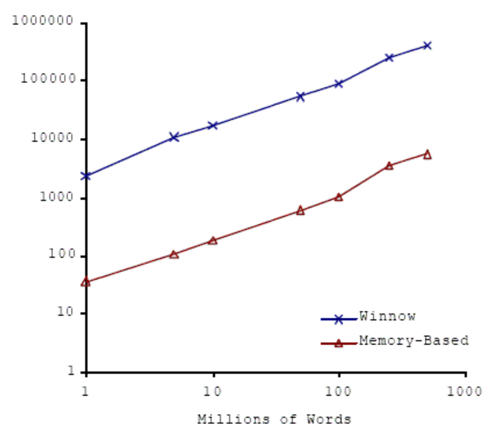

Figure 2. Representation Size vs. Training Corpus Size

Optimizations for Less Data Voting

- What is Voting?
  - Train a set of classifiers on the same corpus, then for a test classification use democracy
- Complementarity (how often they agree)
  - Direct relationship with training corpus size

Efficacy of Voting

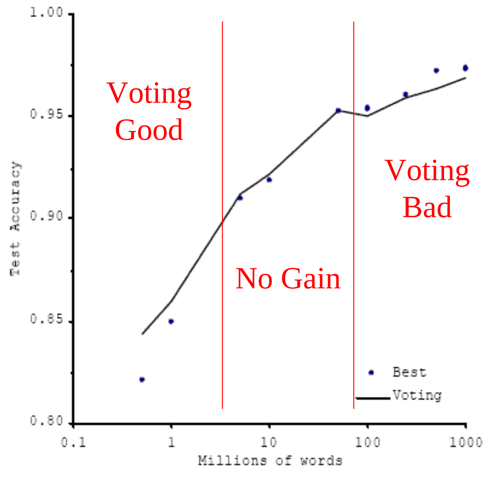

Figure 3. Voting Among Classifiers

Not So Fast…

- Although this supports a conclusion to use more data, how realistic would that be?
- Remember the "Key Properties" from earlier?
- It is only for a few problems that access to large amounts of labeled data exists.
- Manual annotation is seemingly impractical
- Let's try to take advantage of it anyway…

Active Learning

- "involves intelligently selecting a portion of samples for annotation from a pool of as-yet unannotated training samples."
- Essentially, maximizing the utility of any fixed amount of manual effort

Active Learning Examples

- Run a seed learner over the test data, and use confidence ratings as indicators of usefulness
- Alternatively, run a set of seed learners and use their agreement as an indicator

Bagging

- Generates many classifiers
- To measure uncertainty of a classification

- Select, with replacement, random sentences from the original corpus
- Generate N training sets this way, all of size equal to the original corpus

Active Learning

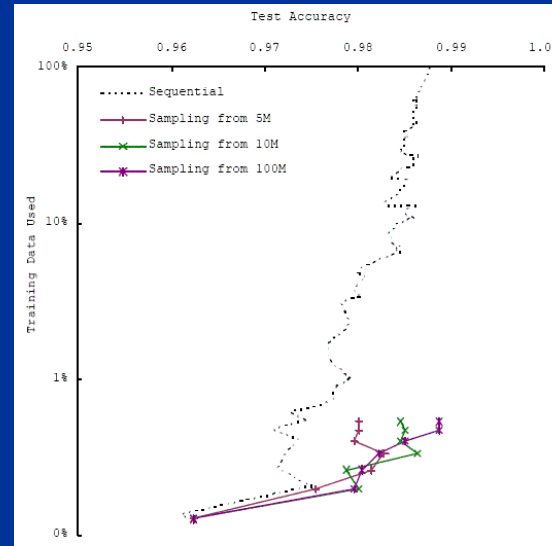

Figure 4. Active Learning with Large Corpora

Co-training and Bootstrapping

- Start with a training set of high confidence examples (perhaps manually annotated)
- Iterate:
  - Train and run your classifier over the test set
  - Add those samples of highest confidence from the test set into the training set

Weakly Supervised Learning

Classifiers In Agreement	Test Accuracy
10	0.8734
9	0.6892
8	0.6286
7	0.6027
6	0.5497
5	0.5000

Table 2. Committee Agreement vs. Accuracy

	{then, than}		{among, between}	
	Test Accuracy	% Total Training Data	Test Accuracy	% Total Training Data
10^6-wd labeled seed corpus	0.9624	0.1	0.8183	0.1
seed+5×10^6 wds, unsupervised	0.9588	0.6	0.8313	0.5
seed+10^7 wds, unsupervised	0.9620	1.2	0.8335	1.0
seed+10^8 wds, unsupervised	0.9715	12.2	0.8270	9.2
seed+5×10^8 wds, unsupervised	0.9588	61.1	0.8248	42.9
10^9 wds, supervised	0.9878	100	0.9021	100

Table 3. Committee-Based Unsupervised Learning

Weakly Supervised Learning

	Unsupervised: All Labels	Unsupervised: Most Certain Labels
{then, than}		
10^7 words	0.9524	0.9620
10^8 words	0.9588	0.9715
$5x10^8$ words	0.7604	0.9588
{among, between}		
10^7 words	0.8259	0.8335
10^8 words	0.8259	0.8270
$5x10^8$ words	0.5321	0.8248

Table 4. Comparison of Unsupervised Learning Methods

Summary

- Often more data is available than researchers are using for experimentation
- This data helps to varying degrees
  - If it's labeled, can make a big difference without requiring extra work (ex. confusion sets)
  - If it's available and some annotation can occur, active learning can help
  - If it's available but no extra work is possible, benefit can still be found (ex. bootstrapping)
- Authors suggest moving "towards increasing the size of annotated training collections"