A Simple Approach to Building Ensembles of Naïve Bayesian Classifiers for Word Sense Disambiguation

Ted Pedersen
Department of Computer Science
University of Minnesota Duluth

Presented by Montana Low

The Big Picture

- Create an effective way to disambiguate words, by using multiple Naïve Bayesian classifiers
  - Each classifier is given a different sized window of context
  - The meaning of the word is decided via a vote of all the classifiers

Motivation

- Simple classifiers with shallow lexical features perform as well as, or better than, more complex classifiers with deep linguistically motivated features
- Ensembles have been used with success for part-of-speech tagging, and parsing

Building the Classifiers

- The features are binary values, does $Word_i$ occur within the context window?
- Left and right windows can be different sizes of 0,1,2,3,4,5,10,25, or 50 words. Results in 81 different classifiers.

$$p(F_1, F_2, \ldots, F_n, M) = p(M) \prod_{i=1}^{n} (F_i | M)$$

The Data

Line Meanings	Count
product	2218
written or spoken text	405
telephone connection	429
formation of people or things ; queue	349
an artificial division; boundary	376
a thin, flexible object; cord	371
Total	4148

The experiments in this paper and previous work use a uniformly distributed subset of this corpus, where each meaning occurs 349 times.

Interest Meanings	Count
money paid for the use of money	1252
a share in a company or business	500
readiness to give attention	361
advantage, advancement or favor	178
activity that one gives attention to	66
causing attention to be given to	11
Total	2368

The experiments in this paper and previous work use the entire corpus, where each sense occurs the number of times shown above.

The Experiment

- 5 fold cross validation was employed
  - Divide the data into 5 randomly shuffled “folds”
  - 4 folds were used to train the classifiers
  - 1 fold was divided in half to form 2 test sets
  - The first test set was used to test the individual classifiers to see if they were good enough to be included in the final ensemble
  - The second test set was used to test the ensemble

Individual Results

Individual Line Classifier Accuracy

	0	1	2	3	4	5	10	25	50
50	0.63	0.73	0.80	0.82	0.83	0.83	0.83	0.83	0.83
25	0.63	0.74	0.80	0.82	0.84	0.83	0.83	0.83	0.83
10	0.62	0.75	0.81	0.82	0.83	0.83	0.83	0.83	0.84
5	0.61	0.75	0.80	0.81	0.82	0.82	0.82	0.82	0.83
4	0.60	0.73	0.80	0.82	0.82	0.82	0.82	0.82	0.82
3	0.58	0.73	0.79	0.82	0.83	0.83	0.82	0.81	0.82
2	0.53	0.71	0.79	0.81	0.82	0.82	0.81	0.81	0.81
1	0.42	0.68	0.78	0.79	0.80	0.79	0.80	0.81	0.81
0	0.14	0.58	0.73	0.77	0.79	0.79	0.79	0.79	0.80

Best(4,25) = 84%

Individual Interest Classifier Accuracy

	0	1	2	3	4	5	10	25	50
50	0.74	0.80	0.82	0.83	0.83	0.83	0.82	0.80	0.81
25	0.73	0.80	0.82	0.83	0.83	0.83	0.81	0.80	0.80
10	0.75	0.82	0.84	0.84	0.84	0.84	0.82	0.81	0.81
5	0.73	0.83	0.85	0.86	0.85	0.85	0.83	0.81	0.81
4	0.72	0.83	0.85	0.85	0.84	0.84	0.83	0.81	0.80
3	0.70	0.84	0.86	0.86	0.86	0.85	0.83	0.81	0.80
2	0.66	0.83	0.85	0.86	0.86	0.84	0.83	0.80	0.80
1	0.63	0.82	0.85	0.85	0.86	0.85	0.82	0.81	0.80
0	0.53	0.72	0.77	0.78	0.79	0.77	0.77	0.76	0.75

Best(4,1) = 86%

The Ensemble Results

- Line Ensemble achieved accuracy of 88% compared to 84%
- Interest Ensemble achieved 89% accuracy compared to 86%
- These are both significant improvements as judged by McNemar's test with $p = .01$

Comparison to Previous Work

- Interest
  - Bruce and Wiebe, 1994: 78%
  - Ng and Lee, 1996: 87%
  - Pedersen et al., 1997: 74%, 84%
- Line
  - Leacock et al., 1993: 71%, 72%, 76%
  - Mooney, 1996: 72%, 71%
  - Towell and Voorhees, 1998: 87%
  - Leacock et al., 1998: 84%

Future Work

- Majority vote was better than weighted vote, why?
- Giving voting rights to all 81 performed worse.
  - How to select best members for the ensemble?
- Why don't deeper features help?
- More and different window sizes
  - They are working on an algorithm to determin the optimal window size

Conclusions

- Ensembles are neat, because they work.