Unsupervised Discrimination of Person Names in Web Contexts

Ted Pedersen[1] and Anagha Kulkarni[2]

[1] University of Minnesota, Duluth, MN 55812, USA
[2] Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. Ambiguous person names are a problem in many forms of written text, including that which is found on the Web. In this paper we explore the use of unsupervised clustering techniques to discriminate among entities named in Web pages. We examine three main issues via an extensive experimental study. First, the effect of using a held–out set of training data for feature selection versus using the data in which the ambiguous names occur. Second, the impact of using different measures of association for identifying lexical features. Third, the success of different cluster stopping measures that automatically determine the number of clusters in the data.

1 Introduction

As the Web increases in coverage, there is a growing problem of ambiguity, since different people or organizations can share the same name. In this paper we evaluate the effectiveness of unsupervised methodologies that cluster short contexts based on their similarity. We apply these techniques to the problem of discriminating among named entities as found in Web pages.

These techniques are based on the Distributional Hypothesis (e.g., [3], [6]) which holds that words that occur in similar contexts will tend to have similar meanings. Our approach is to cluster Web contexts that contain an ambiguous name such that each resulting cluster represents a particular entity. These contexts are approximately 100 word–long passages of text taken from Web pages, where an ambiguous name is located in the middle of the context.

These methods have previously been applied to discriminating among the meanings of ambiguous names and words, or grouping short contexts based on their topic. Specific examples where these methods have been applied include word sense discrimination (e.g., [11], [12]), email clustering (e.g., [4]), and named entity discrimination (e.g., [10]).

The techniques we will describe are language independent (c.f., [9]) and as such only rely on lexical features that can be identified in raw corpora or Web pages. They do not incorporate any syntactic or linguistic information, nor do they utilize any manually created or maintained knowledge sources. As such they are ideal for Web contexts, which are often not well formed and include many strings that are not typically a part of knowledge bases or dictionaries. While our

A. Gelbukh (Ed.): CICLing 2007, LNCS 4394, pp. 299–310, 2007.

evaluation is done with English language texts, these methods can be applied to Web contexts in any other language.

This paper reviews our methods of feature selection, paying particular attention to several different measures of association we evaluate. It then outlines the cluster stopping methods we use to predict the number of clusters automatically, and then describes how these clusters can be evaluated. We then discuss our experimental data and the results we obtained.

2 Lexical Features

A corpus of feature selection data is used to identify the bigram features that will represent the Web contexts to be clustered (i.e., the evaluation or test data). The feature selection data may be the evaluation data itself, or a separate corpus of held out training data that will not be clustered.

Bigrams are ordered pairs of words that occur next to each other. These are selected by identifying which of these pairs occur together more often than we would expect by chance. We compare Fisher's Exact Test[7], the Log-Likelihood Ratio[1], the Odds Ratio, and Pointwise Mutual Information (PMI).

All of these measures are based on word and bigram counts obtained from the feature selection data. Figure 1 summarizes the notation that we use to represent the bigram counts, which are stored in a 2×2 contingency table. Each bigram

	cat	¬cat	totals
big	n_{11}= 10	n_{12}= 20	n_{1+}= 30
¬big	n_{21}= 40	n_{22}= 930	n_{2+}= 970
totals	n_{+1}=50	n_{+2}=950	n_{++}=1000

Fig. 1. Representation of Bigram Counts

observed in the feature selection data is considered a candidate bigram and has a table associated with it. In Figure 1 the candidate bigram is *big cat*. The value of n_{11} shows how many times *big cat* occurs in the corpus. The value of n_{12} shows how often bigrams occur where *big* is the first word and *cat* is not the second. Likewise, n_{21} indicates how many bigrams occur where *big* is not the first word but *cat* is the second Finally, n_{22} is the count of bigrams where neither the first word is *big* nor is the second word *cat*. The counts in n_{1+} and n_{+1} indicate how often *big* and *cat* occur as the first and second words of any bigram in the corpus. The total number of bigrams in the corpus is represented by n_{++}, which is the sum of all the interior cell counts.

We make use of a stop–list to exclude bigrams made up of non-content words. We create our stop list automatically by computing the Inverse Document Frequency (IDF) for each word that occurs in the feature selection data. This is equal to the number of Web contexts in the feature selection data divided by the number of Web contexts in which the given word occurs. Any word with an IDF greater than or equal to 10 is considered a stop word since this means

that the word occurs in 10% or more of the contexts, and may be of limited value in discrimination since it occurs so widely. Any bigram consisting of one or two stop words or that does not exceed a given frequency cutoff is not used as a feature. Below we describe each of the measures that we used for identifying lexical features.

Pointwise Mutual Information is defined as shown in Equation 1.

$$PMI = \log \frac{n_{11}}{m_{11}} = \log \frac{n_{11} * n_{++}}{n_{1+} * n_{+1}} \tag{1}$$

PMI is simply the ratio of the observed number of times the candidate bigram occurs (n_{11}), divided by the number of times this bigram would be expected to occur if the words in the bigram were truly independent (m_{11}). The expected value is calculated by taking the the product of the marginal totals n_{1+} and n_{+1} and dividing by the sample size n_{++}.

If the observed value is much greater than the expected value, this means that the bigram has occurred more often than would be expected by chance, and the pair of words is strongly associated and should be selected as a feature. A bigram is used as a feature if it has a PMI score of 5 or above, which means intuitively that the bigram has occurred at a rate 5 times expected by chance.

PMI suffers from a well known bias towards bigrams that are made up of words that only occur with each other, and in fact gives the highest score to any bigram that only occurs 1 time, and where the words that make up the bigram only occur in that bigram. While this is not desirable behavior in general, when identifying significant bigrams this can actually be a positive characteristic. In many cases the distribution of identities in ambiguous Web names is very skewed, and the features associated with one name may dominate to the point where the features of the other name can not even be recognized. However, if there is very distinct bigram that occurs with a low frequency name, it can still be identified by PMI since it will rise to the top even with relatively low frequency.

The Log–Likelihood Ratio (G^2) is defined as shown in Equation 2.

$$G^2 = 2 * \sum_{i,j} n_{ij} * \log * \frac{n_{ij}}{m_{ij}} \tag{2}$$

where n_{ij} is the observed count of bigrams, where i and j are 1 or 2 and are defined as shown in Figure 1. The value of G^2 indicates the degree to which the occurrence of that bigram deviates from what would be expected by chance. Thus, the larger the G^2 value the more likely that the words in the bigram are not independent. Any bigram with a G^2 value greater than or equal to 3.84 is considered a feature. This is the value associated with a 95% probability that the words in the bigram are not independent. This value comes from the Chi–squared distribution, which approximates the distribution of the Log–Likelihood Ratio and can therefore be used as a source of critical values.

Note that PMI is in fact one term in the G^2 equation (when i and j are both equal to 1). However, rather than focusing on just the count and expected value of the candidate bigram, G^2 considers the counts of the other bigrams in the sample as well. This allows for a formal test of statistical significance, which answers the question of how likely it would be for the candidate bigram to be drawn from the given sample, if the words in the candidate bigram are truly independent.

Fisher's Exact Test([2], [7]) computes the probability that an observed bigram is statistically significant by exhaustively computing the probability of every possible contingency table that would lead to the marginal totals that are in the observed table.

When performing Fisher's Exact Test on a 2×2 contingency table the marginal totals n_{1+} and n_{+1} and the sample size n_{++} must be fixed at their observed values. Given this, the value of n_{11} determines the value of n_{12}, n_{21} and n_{22}. All of the possible 2×2 tables that adhere to the fixed marginal totals are generated and the probability of each table is computed using the hyper-geometric distribution as is shown in Equation 3.

$$P = \frac{1}{n_{11}!n_{12}!n_{21}!n_{22}!} * \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n_{++}!} \tag{3}$$

A left sided test will tell how likely it is for a bigram to occur less frequently than the one we have observed with the given marginal totals. Thus, a high value of P means that the bigram is statistical significant, since it is much more likely that bigrams would occur less frequently that we observed if they were independent. We can calculate P by adding the probabilities of all the possible 2×2 contingency tables where n_{11} is less than the observed value. Any candidate bigram with a total probability greater than or equal to 0.95 is considered a feature, which is equivalent to the threshold used in the log–likelihood ratio.

The Odds Ratio is defined as shown in Equation 4.

$$odds = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{n_{11} * n_{22}}{n_{21} * n_{12}} \tag{4}$$

The numerator is the odds of *big cat* occurring versus *X cat*, where *X* can be any word other than *big*. The denominator is the odds of *big Y* occurring, where *Y* can be any word other than *cat*, versus any bigram that does not include *big* as the first word and *cat* as the second word. This ratio can also be expressed as the cross product of the counts in the contingency table, as shown in Equation 4. The higher this ratio, the greater the odds that the candidate bigram is significant. We use a value of 1,000 as our threshold for the odds ratio.

3 Second Order Context Representation

We represent the Web contexts to be clustered using a second order representation that follows from [12] and is based directly on [11].

We create a matrix from the bigrams identified as features, where the rows represent the first word in a bigram, and the columns represent the second. The cell values are the scores found for the bigrams by whichever of the measures above were used. Each row of this matrix forms a vector that represents the words that follow that particular word in the bigrams identified as features.

Each context to be clustered is represented such that each word in the context for which a row vector exists is replaced by that vector. Recall that in our feature selection process we removed any bigrams that contained one or two stop words, so the words in the contexts that will be represented are content words. After the vectors are substituted for the words, any words for which there are no corresponding vectors are removed, and the vectors are averaged together to represent the context. Each context is represented by such a vector, and these become the input to the clustering algorithm.

4 Cluster Stopping

We use the method of Repeated Bisections for clustering. This is a hybrid method that repeatedly bisects the contexts so as to maximize a given criterion function. We have used the I2 internal criterion function, which is a measure of within–cluster (intra) similarity. This measures the distance of all the contexts in that cluster to the centroid, and the goal is to find clusters where that distance is minimized.

While there are existing approaches that carry out word sense discrimination (e.g., [5], [11], [12]), these have required that the user specify in advance the number of clusters to be discovered. This is a significant limitation, since in general a user will not know this number, and in fact discovering it might be a goal of the experiment in the first place.

Instead, we rely on three cluster stopping measures introduced in [8] to determine the number of clusters automatically. These include the Adapted Gap Statistic [13], the PK2 measure, and the PK3 measure. As such we do not need to specify ahead of time the number of clusters that we expect to find, this is determined automatically. We find a solution with 1 cluster, then 2 clusters, and so forth, up to a number of clusters where there is no further improvement in the quality of the solution. Then, we examine the trend of criterion function scores (I2) for these successive solutions, and identify the point at which adding to the number of clusters does not significantly improve upon the quality of the solution.

The PK2 measure compares the value of the criterion function for successive pairs of clusters k and $k-1$. When this ratio approaches 1, then the creation of additional clusters is not improving the quality of the solution, and should be

stopped. The PK3 measure takes the ratio of the criterion function value at k with the sum of the criterion functions at $k-1$ and $k+1$. PK3 will be close to 1 if these three values form a line, meaning that the criterion function is still improving, since the line will break at the point where a plateau exists and the scores no longer improve. When using PK2 or PK3, we select the value of k that is closest to but still greater than one standard deviation in the value of the PK2 or PK3 score.

The Gap Statistic compares the observed and expected values of the criterion function. The expected values are estimated from a randomly generated data set that maintins the same marginal totals as the observed data. Thus, this data represents the same population as that of the observed data, except that it is made up of noise. When random data is clustered the criterion function should exhibit a relatively consistent score as k increases, which will quantify the amount of noise present in the data. Selecting the number of clusters reduces to finding the point where the difference between criterion function score of the observed and expected values is greatest. This is the point at which the observed data is least like noise, and the point where the optimal number of clusters exists.

5 Experimental Data

We have manually disambiguated Web contexts obtained from the Google Search Engine API to create gold standard data for five different ambiguous names:

> Richard Alston, Sarah Connor, George Miller, Ted Pedersen, Michael Collins

Web contexts for each of these names was collected in May 2006 using the Google API, as supported by the CPAN module WebService-GoogleHack-0.15. The top 50 html (or htm) pages found when searching for each of these names were retrieved, and any links from those pages to pages in the same domain were followed and those pages retrieved. However, the links on the second level pages were not traversed.

All the pages retrieved were formatted and cleaned as follows. First, all HTML tags were stripped away using the CPAN module HTML-Format-2.04. This data was divided into contexts using the freely available NameConflate program (version 0.16)[1]. Each context contains a single ambiguous name. Note that contexts may contain variants of the names listed above, such as *M. Collins* or *Ted A. Pedersen*.

Each Web context consists of approximately 100 total words, where the ambiguous name is located in the center of the context. Table 1 shows the number of contexts associated with each name, and the distribution of identities associated with the contexts:

[1] http://www.umn.edu/home/tpederse/tools.html

Table 1. Name Data

Name: Identity	Count	%	Name: Identity	Count	%
Richard Alston:	247		Michael Collins:	359	
Choreographer	176	71.3	Irish Leader	269	74.9
Senator (Australia)	71	28.7	MIT Professor	41	11.4
Sarah Connor:	150		Wisconsin Professor	32	8.9
German Singer	109	72.7	NASA Astronaut	17	4.7
Terminator Character	41	27.3	Ted Pedersen:	333	
George Miller:	286		Minnesota Professor	255	76.6
Congressman (USA)	217	75.9	Children's Author	43	12.9
Film Director (Australia)	57	19.9	Son of Sea Captain	25	7.5
Princeton Professor	12	4.2	TV Writer	10	3.00

6 Evaluation

After the clusters have been discovered, they are aligned with the gold standard data such that the agreement between the two is maximized. Each discovered cluster is aligned to a single gold standard cluster, and it is possible that the number of discovered clusters will be more or less than the gold standard amount.

The quality of the clustering is scored using the F–measure, which is the harmonic mean of precision and recall. We define precision to be the number of contexts that are assigned to their correct class, divided by the number of contexts that are assigned a class. Recall is defined as the number of contexts assigned to their correct class, divided by the total number of contexts. Precision and recall differ because the clustering algorithm may decide not to cluster a context, and if the clustering algorithm creates more clusters than there are in the human gold standard, the extra clusters that remain after alignment with the human gold standard are discarded.

Thus, the F-measure provides an indication of how well the clustering is being carried out both in terms of discovering the number of clusters, and then in terms of the quality of the resulting clusters.

Note that in clustering if all of the Web contexts for a given name are assigned to the same cluster, the F–Measure will be equal to the percentage of the majority identity in the data. Thus, this serves as a baseline measure to which we can compare.

7 Experimental Results

For each of the five names in the evaluation data, we carried out a number of experimental variations. The feature selection data was either the contexts to be clustered themselves, or contexts (articles) from the New York Times portion of the English GigaWord Corpus. We used the first 25,000 and 75,000 contexts as our two sets of feature selection data. We also experimented with four different measures of association for feature selection, and three different methods of cluster stopping.

Table 2. ALSTON results : 2 identities, majority 71.26

	nyt-25 5		nyt-75 5		nyt-25 10		nyt-75 10		nyt-25 20		nyt-75 20		test 2		test 5	
Fisher																
gap	3	68.32	2	**88.66**	3	77.64	3	73.33	3	80.19	3	72.68	1	71.26	1	70.99
pk2	3	68.32	2	**88.66**	3	77.64	3	73.33	3	80.19	3	72.68	41	16.36	25	17.58
pk3	3	68.32	2	**88.66**	3	77.64	3	73.33	3	80.19	3	72.68	21	27.27	10	35.18
man	2	90.28	2	**88.66**	2	**99.19**	2	88.66	2	**99.10**	2	88.66	2	81.38	2	60.45
ll																
gap	4	73.60	3	91.97	4	71.83	3	85.71	4	70.83	5	67.18	1	71.26	1	70.99
pk2	3	90.83	5	72.02	4	71.83	5	67.72	4	70.83	5	67.18	8	60.06	5	47.59
pk3	4	73.60	3	91.97	4	71.83	2	**95.14**	4	70.83	2	93.12	4	71.17	7	47.59
man	2	92.31	2	88.26	2	92.71	2	**95.14**	2	91.90	2	93.12	2	79.76	2	53.55
odds																
gap	1	71.26	1	71.26	1	71.26	1	71.26	1	71.26	1	71.26	6	58.33	1	**72.08**
pk2	5	60.10	5	59.38	5	58.45	4	70.16	5	57.67	4	66.15	5	55.64	5	44.25
pk3	3	50.66	3	65.48	5	58.45	4	70.16	6	55.43	4	66.15	6	58.33	7	46.30
man	2	58.70	2	60.32	2	**90.28**	2	83.00	2	68.42	2	85.02	2	72.06	2	58.33
pmi																
gap	3	72.49	3	69.47	3	77.83	3	72.16	3	78.73	3	71.35	1	71.26	1	70.99
pk2	3	72.49	3	69.47	3	77.83	3	72.16	3	78.73	3	71.35	48	13.58	32	18.91
pk3	3	72.49	3	69.47	3	77.83	2	**89.47**	3	78.73	2	**89.47**	5	64.48	31	8.91
man	2	91.90	2	89.07	2	**99.19**	2	89.47	2	**99.19**	2	89.47	2	88.66	2	83.98

Table 3. CONNOR results : 2 identities, majority 72.67

	nyt-25 5		nyt-75 5		nyt-25 10		nyt-75 10		nyt-25 20		nyt-75 20		test 2		test 5	
Fisher																
gap	1	72.67	2	57.33	2	66.00	2	62.00	1	72.67	3	72.22	3	58.91	1	69.57
pk2	3	**79.20**	2	57.33	4	75.21	4	70.16	4	76.73	3	72.22	4	58.91	4	43.97
pk3	3	**79.20**	2	57.33	4	75.21	4	70.16	4	76.73	2	62.00	6	55.90	4	43.97
man	2	**66.00**	2	57.33	2	**66.00**	2	62.00	2	64.00	2	62.00	2	52.38	2	49.28
ll																
gap	2	50.00	2	50.00	28	40.43	7	52.94	13	48.48	2	50.00	1	**70.07**	1	69.57
pk2	3	52.55	2	50.00	3	52.01	2	50.00	3	66.92	2	50.00	4	61.72	4	37.07
pk3	2	50.00	2	50.00	2	50.00	2	50.00	2	50.00	2	50.00	9	50.73	2	49.28
man	2	50.00	2	50.00	2	50.00	2	50.00	2	50.00	2	50.00	2	**51.02**	2	49.28
odds																
gap	1	72.67	1	72.67	1	72.67	1	72.67	1	72.67	6	80.74	1	70.07	1	69.82
pk2	4	56.59	9	54.98	4	48.63	9	73.56	4	53.28	14	73.23	4	61.72	4	37.07
pk3	2	63.33	2	67.33	5	59.65	3	77.24	3	47.79	2	**90.00**	3	61.72	2	48.73
man	2	63.33	2	67.33	2	67.33	2	78.67	2	61.33	2	**90.00**	2	51.02	2	48.73
pmi																
gap	1	72.67	1	72.67	4	66.11	1	72.67	2	68.67	1	72.67	2	63.95	1	69.82
pk2	3	**80.16**	2	65.33	4	66.11	2	65.33	4	62.98	3	78.71	4	66.39	4	45.30
pk3	2	59.33	2	65.33	4	66.11	2	65.33	4	62.98	3	78.71	4	66.39	4	45.30
man	2	59.33	2	65.33	2	66.67	2	65.33	2	**68.67**	2	65.33	2	63.95	2	50.91

Table 4. MILLER results : 3 identities, majority 75.87

	nyt-25 5		nyt-75 5		nyt-25 10		nyt-75 10		nyt-25 20		nyt-75 20		test 2		test 5	
Fisher																
gap	2	63.99	1	**75.87**	2	72.88	1	**75.87**	2	60.49	1	**75.87**	6	46.25	2	60.84
pk2	4	61.79	26	27.41	4	49.90	4	49.62	5	54.12	5	47.71	6	46.25	5	43.51
pk3	5	53.23	3	62.94	3	56.99	2	59.09	2	60.49	2	57.34	6	46.25	3	60.14
man	3	**67.83**	3	62.94	3	56.99	3	61.54	3	62.24	3	59.79	3	62.59	3	60.14
ll																
gap	3	43.36	3	43.01	2	51.05	3	41.96	2	55.94	3	46.50	6	52.44	6	38.03
pk2	4	51.17	4	42.03	4	46.63	4	38.72	5	42.54	4	40.15	6	**57.37**	6	38.03
pk3	3	43.36	3	43.01	3	50.35	6	38.31	3	39.86	6	39.45	4	54.34	4	44.07
man	3	43.36	3	43.01	3	50.35	3	41.96	3	39.86	3	46.50	3	**65.03**	3	55.24
odds																
gap	1	**75.87**	10	38.71	1	**75.87**	1	**75.87**	1	**75.87**	1	**75.87**	7	42.57	1	**75.87**
pk2	6	44.30	5	37.69	5	48.18	6	40.92	4	50.39	5	49.61	5	43.12	4	41.78
pk3	4	45.60	7	38.57	3	44.41	6	40.92	4	50.39	5	49.61	7	42.57	4	41.78
man	3	**48.60**	3	46.85	3	44.41	3	44.06	3	44.41	3	45.80	3	39.51	3	43.01
pmi																
gap	2	58.74	2	58.04	1	**75.87**	2	62.24	2	63.99	1	**75.87**	7	50.44	1	**75.87**
pk2	4	50.58	5	48.57	5	52.71	7	51.76	5	54.47	5	49.04	6	50.44	5	50.51
pk3	10	36.36	2	58.04	2	63.29	7	51.76	6	55.62	6	49.48	7	50.44	4	48.58
man	3	62.59	3	60.84	3	**66.43**	3	47.55	3	**66.43**	3	61.54	3	46.50	3	59.09

Table 5. COLLINS results : 4 identities, majority 74.93

	nyt-25 5		nyt-75 5		nyt-25 10		nyt-75 10		nyt-25 20		nyt-75 20		test 2		test 5	
Fisher																
gap	5	48.40	2	73.54	5	41.82	2	71.59	3	80.19	2	72.42	1	74.93	1	74.93
pk2	4	61.28	3	44.85	5	41.82	4	64.62	3	80.19	5	60.45	3	**90.25**	5	71.02
pk3	2	62.40	2	73.54	2	59.05	2	71.59	3	80.19	2	72.42	3	**90.25**	5	71.02
man	4	61.28	4	52.92	4	54.60	4	64.62	4	42.90	4	55.99	4	**65.18**	4	62.12
ll																
gap	6	47.86	9	46.61	5	57.96	6	48.59	5	41.92	7	48.58	7	54.10	1	74.93
pk2	5	46.94	6	46.25	5	57.96	5	49.77	5	41.92	5	51.07	6	51.24	6	63.53
pk3	6	47.86	3	52.09	7	48.21	6	48.59	4	52.92	5	51.07	2	**69.92**	4	46.24
man	4	48.19	4	40.95	4	49.58	4	37.88	4	**52.92**	4	39.28	4	52.09	4	46.24
odds																
gap	1	**74.93**	1	**74.93**	32	27.90	1	**74.93**	1	**74.93**	1	**74.93**	6	49.21	1	**74.93**
pk2	5	45.20	6	57.92	6	45.16	6	47.78	6	40.87	5	47.27	6	49.21	5	55.89
pk3	4	55.99	8	52.01	4	42.62	4	44.29	9	43.69	6	35.16	6	49.21	5	55.89
man	4	**55.99**	4	45.40	4	42.62	4	44.29	4	49.30	4	44.57	4	36.49	4	51.53
pmi																
gap	3	71.03	3	45.13	3	64.35	3	43.73	5	50.38	4	52.92	1	74.93	1	74.93
pk2	5	42.94	5	53.89	5	50.84	5	57.01	5	50.38	4	52.92	5	55.95	5	59.41
pk3	3	71.03	5	53.89	5	50.84	5	57.01	5	50.38	9	57.20	4	**75.21**	4	55.99
man	4	57.38	4	52.09	4	64.35	4	50.70	4	45.96	4	52.92	4	**75.21**	4	55.99

Table 6. PEDERSEN results : 4 identities, majority 76.58

	nyt-25 5		nyt-75 5		nyt-25 10		nyt-75 10		nyt-25 20		nyt-75 20		test 2		test 5	
Fisher																
gap	3	47.75	2	47.15	2	63.66	2	55.56	3	60.96	3	42.94	1	**76.58**	1	**76.58**
pk2	5	43.69	3	42.34	5	48.15	4	35.74	3	60.96	4	35.74	5	56.12	7	53.83
pk3	5	43.69	2	47.15	2	63.66	2	55.56	3	60.96	3	42.94	3	43.84	9	51.05
man	4	40.54	4	41.44	4	**45.05**	4	35.74	4	38.74	4	35.74	4	43.24	4	37.84
ll																
gap	2	69.97	8	41.78	2	54.65	3	45.95	2	49.25	3	51.05	1	**76.58**	1	**76.58**
pk2	5	45.98	6	35.19	5	52.97	6	46.45	5	48.61	6	42.88	5	50.99	6	46.84
pk3	2	69.97	6	35.19	2	54.65	6	46.45	2	49.25	3	51.05	7	62.97	8	47.45
man	4	52.85	4	42.04	4	**60.66**	4	40.24	4	55.86	4	51.05	4	46.55	4	49.25
odds																
gap	1	**76.58**	1	**76.58**	1	**76.58**	1	**76.58**	1	**76.58**	1	**76.58**	1	**76.58**	1	**76.58**
pk2	5	32.01	5	50.00	5	32.00	5	40.20	5	40.89	5	32.41	5	45.44	5	48.40
pk3	4	43.54	5	50.00	3	58.26	3	39.94	5	40.89	5	32.41	7	45.44	5	48.40
man	4	43.54	4	**44.74**	4	42.34	4	39.94	4	38.44	4	41.74	4	42.64	4	42.34
pmi																
gap	3	46.55	3	45.05	2	46.85	2	45.35	2	63.36	2	45.95	1	**76.58**	1	**76.58**
pk2	4	42.04	4	40.84	4	49.55	4	41.14	4	38.14	3	43.54	6	62.41	6	47.54
pk3	3	46.55	2	45.05	3	49.55	4	41.14	2	63.36	4	42.94	2	64.86	5	46.05
man	4	42.04	4	40.84	4	49.55	4	41.14	4	38.14	4	42.94	4	45.05	4	**55.26**

The results of our experiments are shown in Tables 2, 3, 4, 5, and 6. Each table is organized as follows. The feature selection data is indicated in the columns: nyt-25 and nyt-75 refer to the 25,000 and 75,000 context collections from the New York Times, and *test* refers to the use of the evaluation data as feature selection data. The numbers below the feature selection data are the frequency cutoffs used. Remember that these indicate that the words that make up a bigram feature must have occurred at least that many times in the feature selection data.

The measures of association and the cluster stopping techniques are shown in the rows. Note that *man* refers to when we set the number of clusters manually to the value that we know to be correct. The integer values in the table are the number of clusters predicted by the cluster stopping method, and the F-Measure obtained with the given combination of settings.

8 Discussion and Conclusions

For each measure of association in our tables of results, we indicate the highest F–Measure attained by the cluster stopping measures (gap, pk2, pk3) and the manually set number of clusters (man). These values are shown in bold face. It would seem that the manual setting of the same number of clusters as is found in the gold standard data should be the best case scenario. However, we can see a number of cases where the discovered number of clusters results in a better

F–Measure even if the number of clusters discovered does not agree with the evaluation data. This can occur because the evaluation data is rather skewed and some of the very small classes are difficult to discover and overall results may improve simply by ignoring those classes.

Across all of the names, we observe that the results based on using the held–out set of training data tend to be somewhat better than those based on using the evaluation data for feature selection. It may be that the evaluation data is simply not large enough to provide a reasonable set of features to perform discrimination.

We can see that for the Alston, Connor, and Collins results there are combinations of settings that result in F–Measures significantly higher than the majority class. However, for the Miller and Pedersen results no combination of settings exceeds that majority class. This initially surprised us since both of these names have fairly distinct senses. However, upon examining the features we found that the contexts for the majority classes were extremely rich in text, while the minority sense were somewhat impoverished. Thus, no matter what kind of feature identification techniques were employed, it was simply not possible to identify features for any of the minority classes.

In general there is not a clearly superior measure of association for all five of the names. In the Alston data the log–likelihood ratio achieved the highest results, in the Connor data it was Pointwise Mutual Information, and in the Collins data it was Fisher's Exact Test. For the Miller and Pedersen data none of the measures of association fared particularly well.

Among the cluster stopping methods, the Adapted Gap Statistic did somewhat better with the more difficult Pedersen and Miller data since it often predicted just one cluster, which results in an F–Measure equal to the majority class. In the case of a hard discrimination decision, this is actually not a bad option, since in effect the cluster stopping algorithm is saying it is unable to make any distinctions so it leaves all the contexts in the same cluster. With the Alston, Connor, and Miller data in general PK2 and PK3 performed slightly better than the Adapted Gap Statistic.

Acknowledgments

This work was supported by a National Science Foundation Faculty Early CAREER Development Award (#0092784).

All of the experiments in this paper were carried out with version 0.95 of the SenseClusters package, freely available from http://senseclusters.sourceforge.net.

References

1. T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993.
2. R. Fisher. *The Design of Experiments.* Oliver and Boyd, London, 1935.
3. Z. Harris. *Mathematical Structures of Language.* Wiley, New York, 1968.

4. A. Kulkarni and T. Pedersen. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. In *Proceedings of the Second Indian International Conference on Artificial Intelligence*, pages 703–722, Pune, India, December 2005.
5. E. Levin, M. Sharifi, and J. Ball. Evaluation of utility of LSA for word sense discrimination. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 77–80, New York City, June 2006.
6. G.A. Miller and W.G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.
7. T. Pedersen. Fishing for exactness. In *Proceedings of the South Central SAS User's Group (SCSUG-96) Conference*, pages 188–200, Austin, TX, October 1996.
8. T. Pedersen and A. Kulkarni. Selecting the right number of senses based on clustering criterion functions. In *Proceedings of the Posters and Demo Program of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 111–114, Trento, Italy, April 2006.
9. T. Pedersen, A. Kulkarni, R. Angheluta, Z. Kozareva, and T. Solorio. An unsupervised language independent method of name discrimination using second order co-occurrence features. In *Proceedings of the Seventh International Conference on Intelligent Text Processing and Computational Linguistics*, pages 208–222, Mexico City, February 2006.
10. T. Pedersen, A. Purandare, and A. Kulkarni. Name discrimination by clustering similar contexts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 220–231, Mexico City, February 2005.
11. A. Purandare and T. Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 41–48, Boston, MA, 2004.
12. H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.
13. R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a dataset via the Gap statistic. *Journal of the Royal Statistics Society (Series B)*, pages 411–423, 2001.