

Exploration of Three Cluster Stopping Rules for the task of Word Sense Discrimination

Anagha Kulkarni

Graduate Student, Department of Computer Science,
University of Minnesota, Duluth.
Summer Intern, 2005, Division of Biomedical Informatics,
Mayo Clinic, Rochester.

25th August 2005



Acknowledgements

- Dr. Guergana Savova, for her support and guidance.
- Dr. Ted Pedersen, Computer Science, U of Minnesota, Duluth, for recommending me and for his ideas and suggestions.
- Dr. Terry Therneau, Biostatistics Mayo, for his guidance and feedback.
- James Buntrock, for his support.
- Marcy, Patrick, Dana and Tanya for numerous valuable suggestions, feedbacks and support.
- Dr. Christopher Chute, for providing this opportunity.



Overview

- Motivation
- Theoretical Background
 - Word Sense Discrimination
 - Cluster Stopping Rules – In general
 - “The three” Cluster Stopping Rules
- Experiments Conducted
 - Experimental Setup
 - Experimental Results
- Discussion
- Future Work
- References



Motivation

- Improvements needed for indexing and retrieving information from clinical notes.
- Word Sense Discrimination (WSD) can help index more appropriately and retrieve relevant notes.



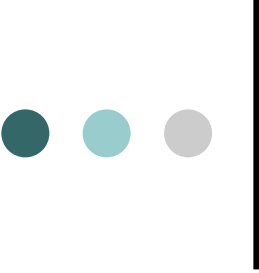
Theoretical Background

...



Word Sense Discrimination (WSD)

- One word with multiple senses/meanings.
- **Cold** –
 1. Cold Temperature
 2. Common Cold
 3. Cold Sensation
- **Culture** –
 1. Anthropological Culture
 2. Laboratory Culture
- Unsupervised WSD using the contextual similarity (Pedersen and Bruce, 1997; Schutze, 1998)
 - “We report that **cold** storage induces...”
 - “... susceptibility to **colds** appeared to be positively associated ...”
 - “... the affected limb with touch and **cold** allodynia ...”



SenseClusters (v0.69) Package

<http://senseclusters.sourceforge.net>

- Given numerous contexts, SenseClusters' groups together similar contexts.
- Open Source software for Unsupervised Clustering of Similar Contexts.
- Started by Amruta Purandare and Dr. Ted Pedersen in September 2002 and continued by myself and Dr. Pedersen from September 2004.
- What is “Unsupervised”? - Without any training/knowledge sources.



How does Unsupervised WSD work?

- Represent each context in terms of lexical features. Thus translate each context into a feature vector.
- Lexical Feature – unigrams, bigrams
- Simple example:

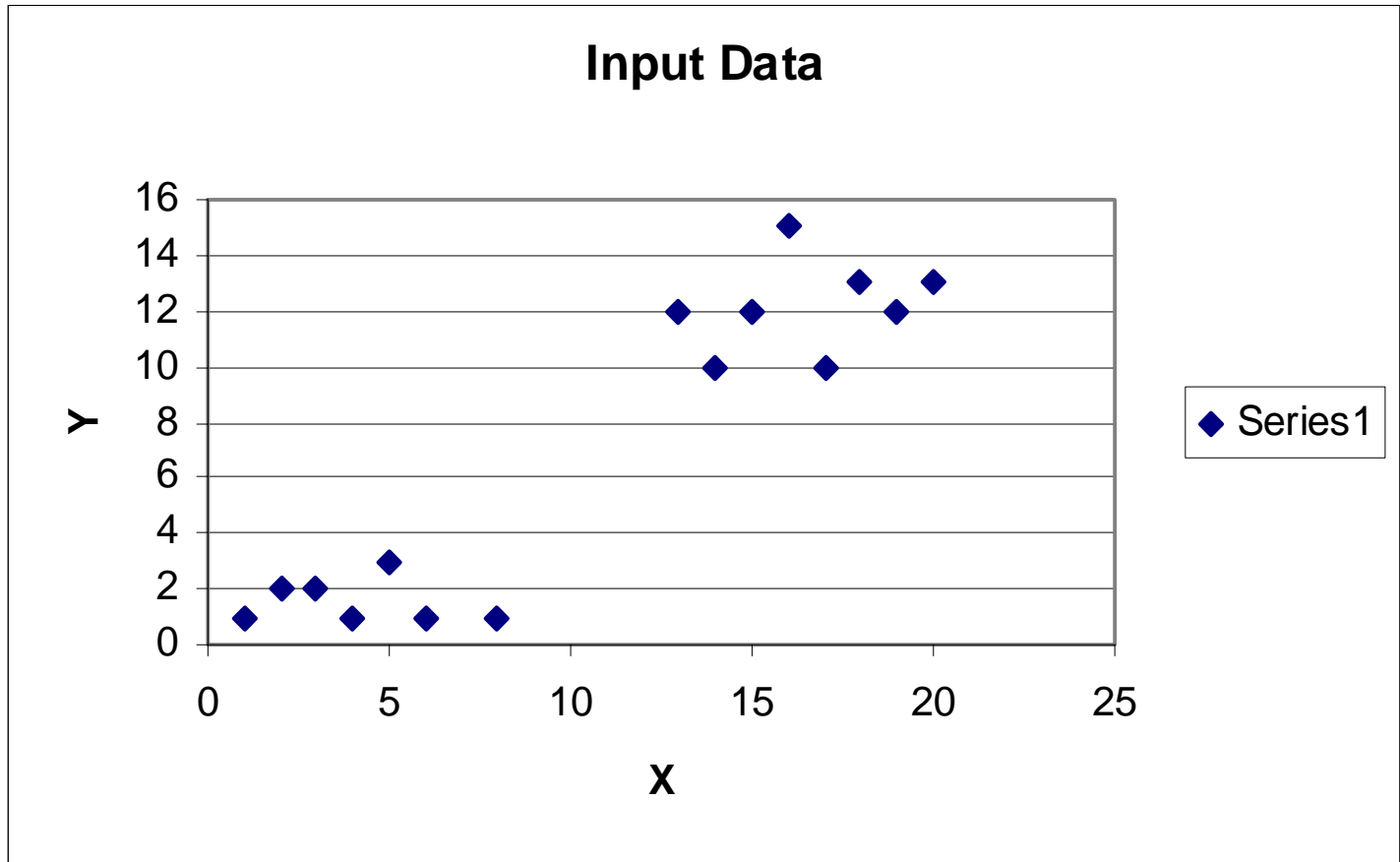
	symptoms	storage	sensation
Context1	1	0	0	...
Context2	0	1	0	...
Context3	1	0	0	...
....



Where is the problem?

- Cluster the context vectors.
- But into how many clusters? – “Cluster Stopping Rules”
- Unsupervised – No knowledge of input data
- More than necessary clusters - Too fine grained distinction.
Less than necessary clusters – Too coarse distinction.
- Directly affects the performance of the methods / system.

Input Data - Example

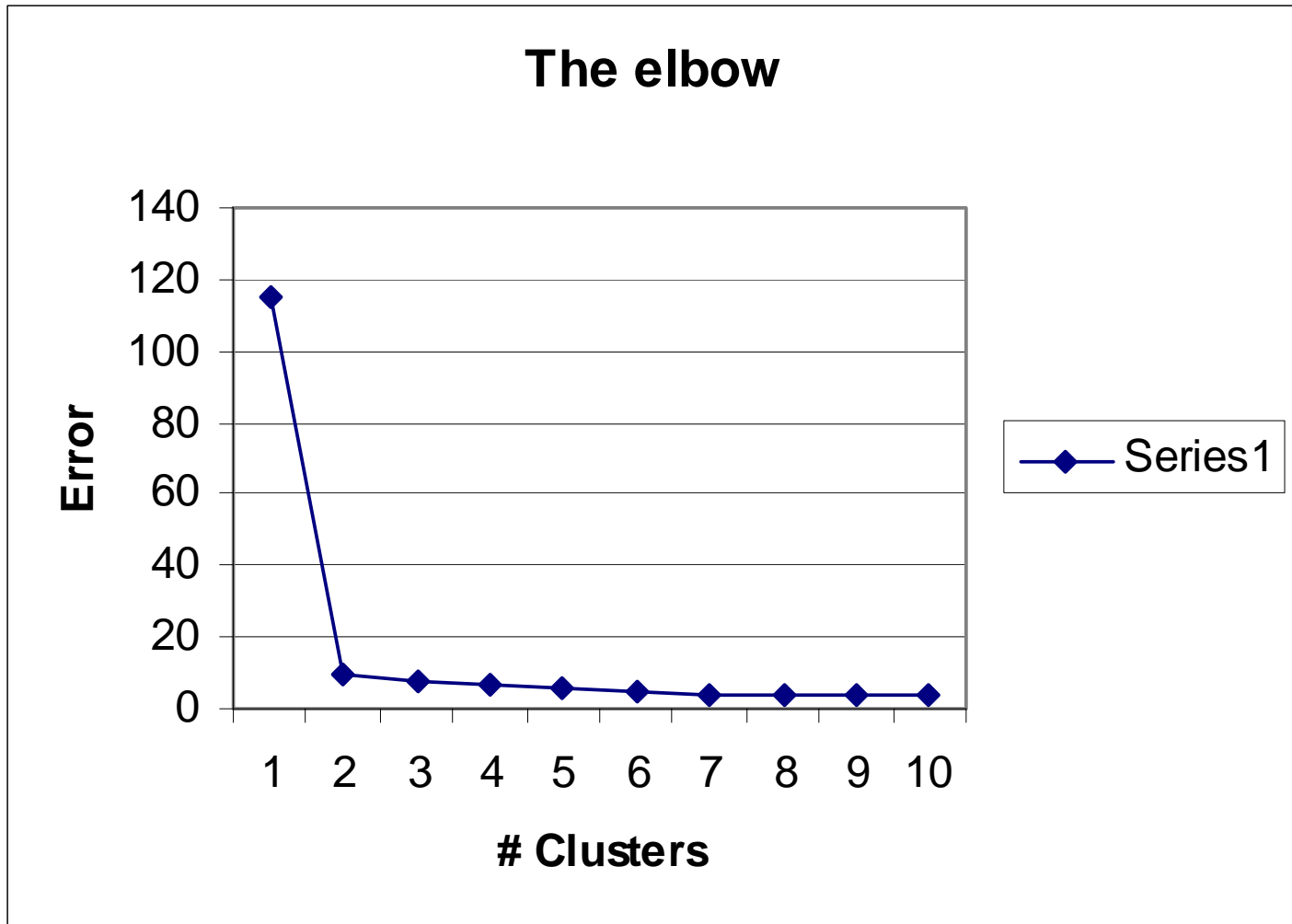




Cluster Stopping Rules – In general

- Estimate the number of clusters that a given dataset naturally separates out into.
- Various techniques have been proposed for this problem – we have studied the following 3 separate stopping rules:
 - **Gap Statistic** (Tibshirani et al., 2001)
 - **Calinski and Harabasz (C&H)** (Calinski and Harabasz, 1974)
 - **Hartigan** (Hartigan, 1975)

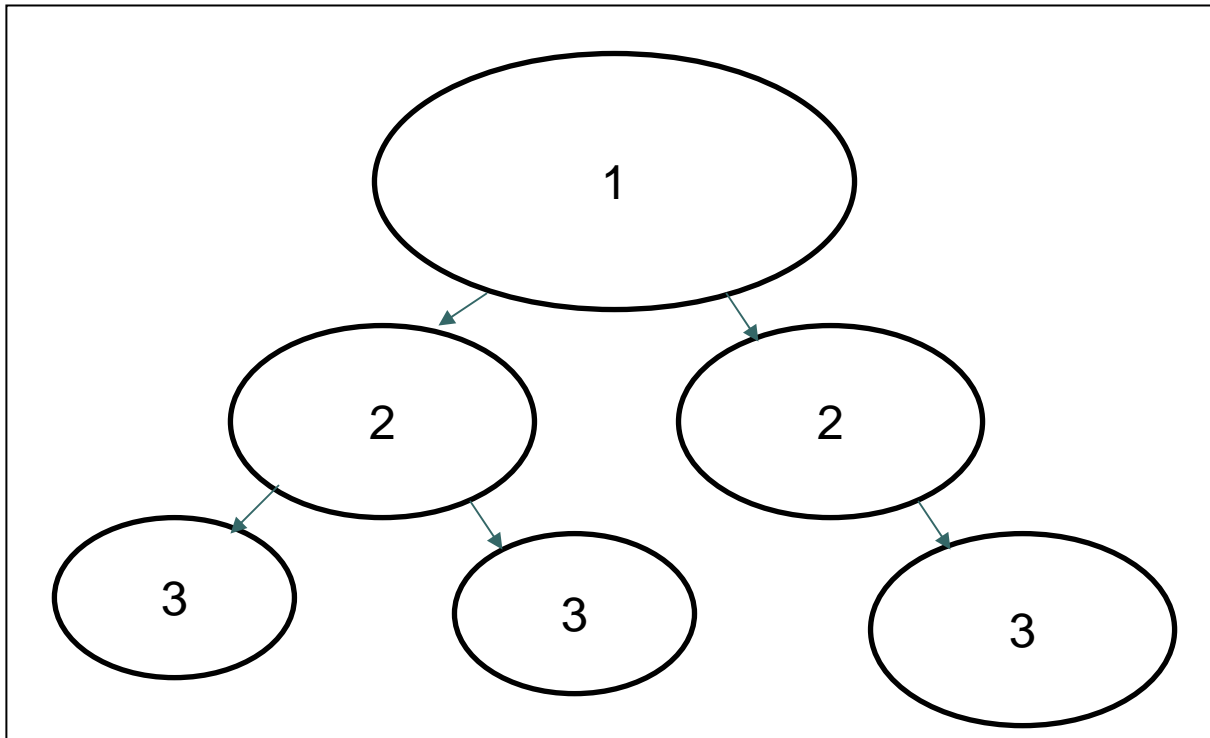
The Idea



Hartigan Method

(<http://search.cpan.org/dist/Statistics-Hartigan/>)

- $H(k) = (N - k - 1) * (e(k) / e(k+1) - 1)$
 - N: Total # contexts to be clustered
 - k: # clusters
 - e(k): Total Error in all clusters when clustered in k clusters.





Hartigan Method

(<http://search.cpan.org/dist/Statistics-Hartigan/>)

- Ratio indicates the error that will be introduced or removed by splitting the k clusters into $k+1$ clusters.
- Requires a threshold T , such that the k value for which the $H(k) \leq T$ is satisfied is the optimal number of clusters for the dataset.
- Can handle the case of single cluster.

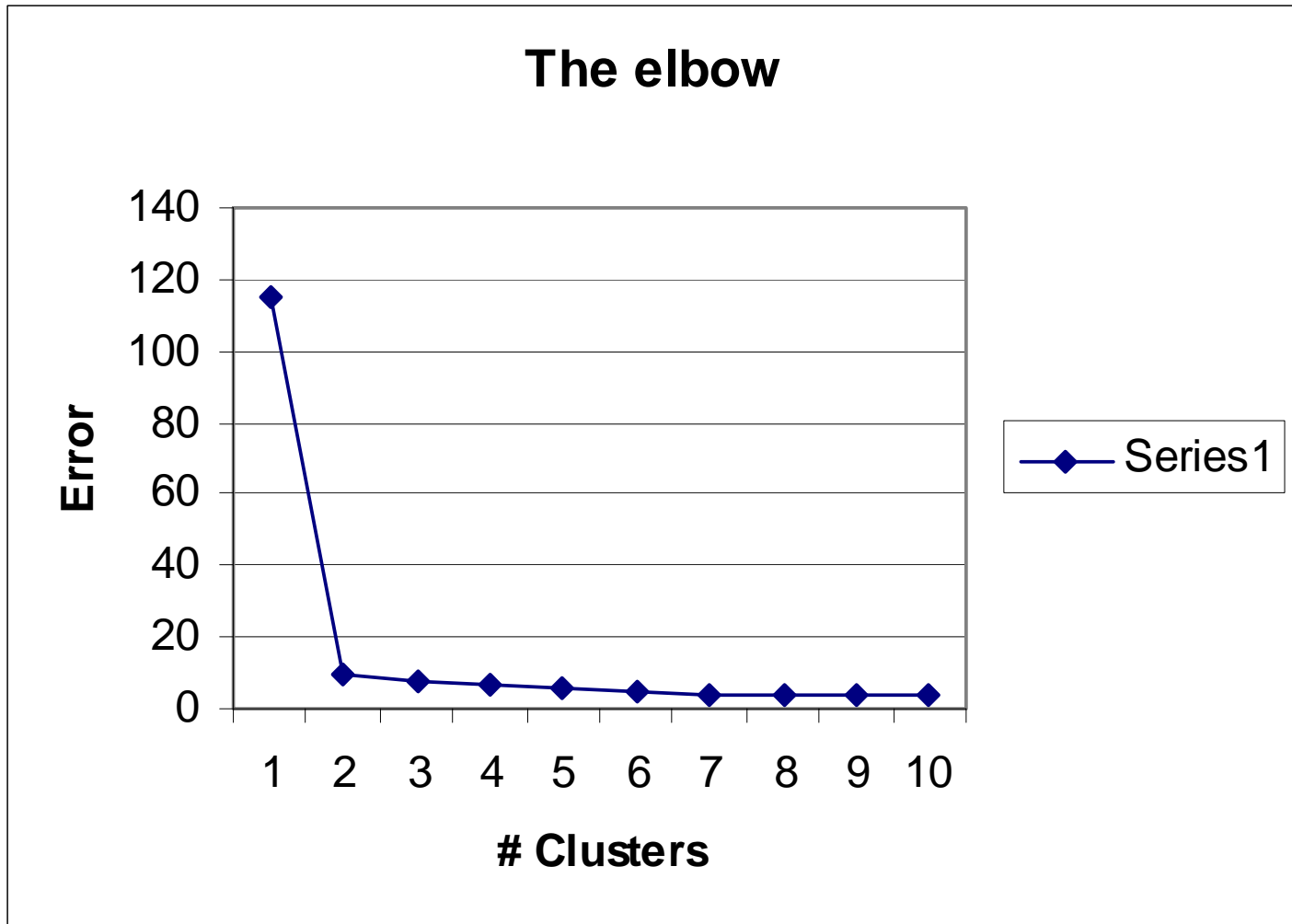


Calinski and Harabasz Method (C&H)

(<http://search.cpan.org/dist/Statistics-CalinskiHarabasz/>)

- Milligan and Cooper Study – Best method (Milligan and Cooper, 1985)
- $CH(k) = BGSS / WGSS * (n - k) / (k - 1)$
- Uses Within Group Sum of squares (WGSS) and also Between Group Sum of Squares (BGSS)
- Minimize WGSS and Maximize BGSS
- Picks the k value (# Clusters) that maximizes the ratio (CH(k)) of BGSS to WGSS
- Cannot capture the case where all the data naturally falls into one cluster.

The Idea





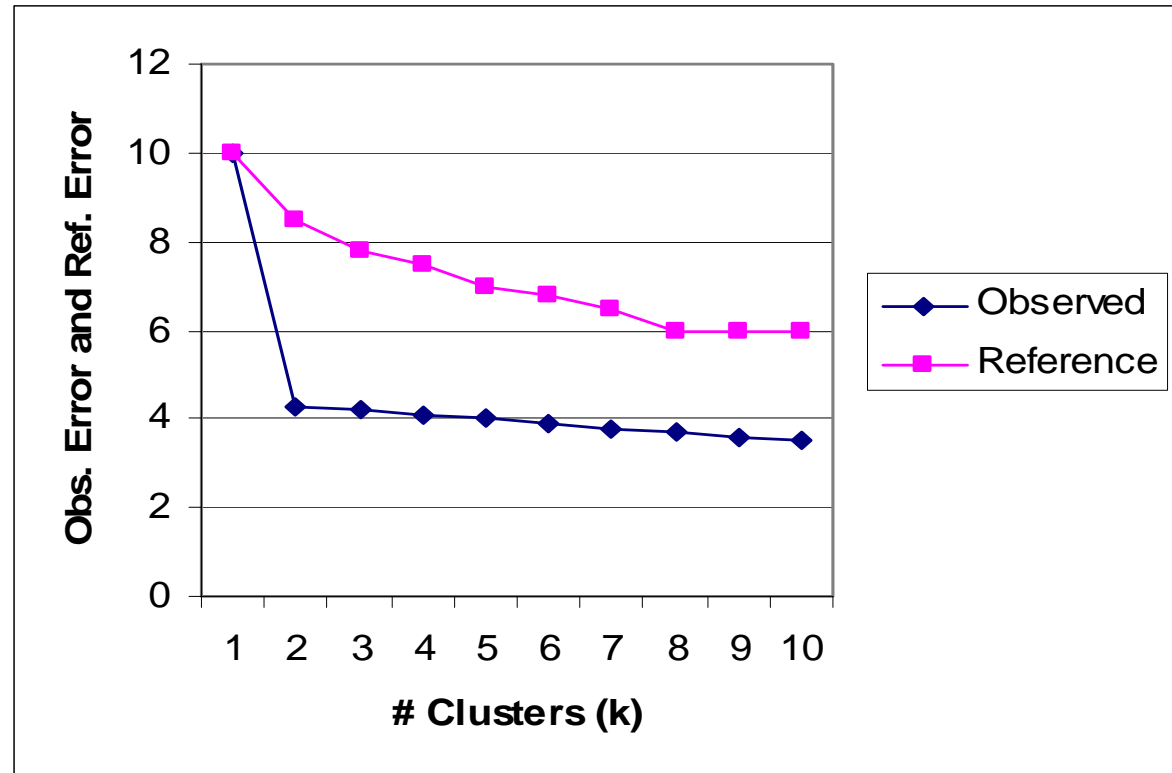
Gap Statistics

(<http://search.cpan.org/dist/Statistics-Gap/>)

- The main idea is to standardize the graph of Error ($\log(W_k)$) by comparing with the expected graph under appropriate null reference distribution.
- The adopted null model is the case of single cluster ($k=1$) which is rejected in favor of $k > 1$ value if sufficient evidence is present.
- The reference distribution to be used should be driven by the underlying data and problem domain.

Gap Statistics (cont.)

The k for which the error value falls farthest below the expected curve is the optimal k value. (In this case $k = 2$)





Reference Distribution Generation

- Parametric Bootstrapping
- Two methods:
 - Uniform
 - Proportional

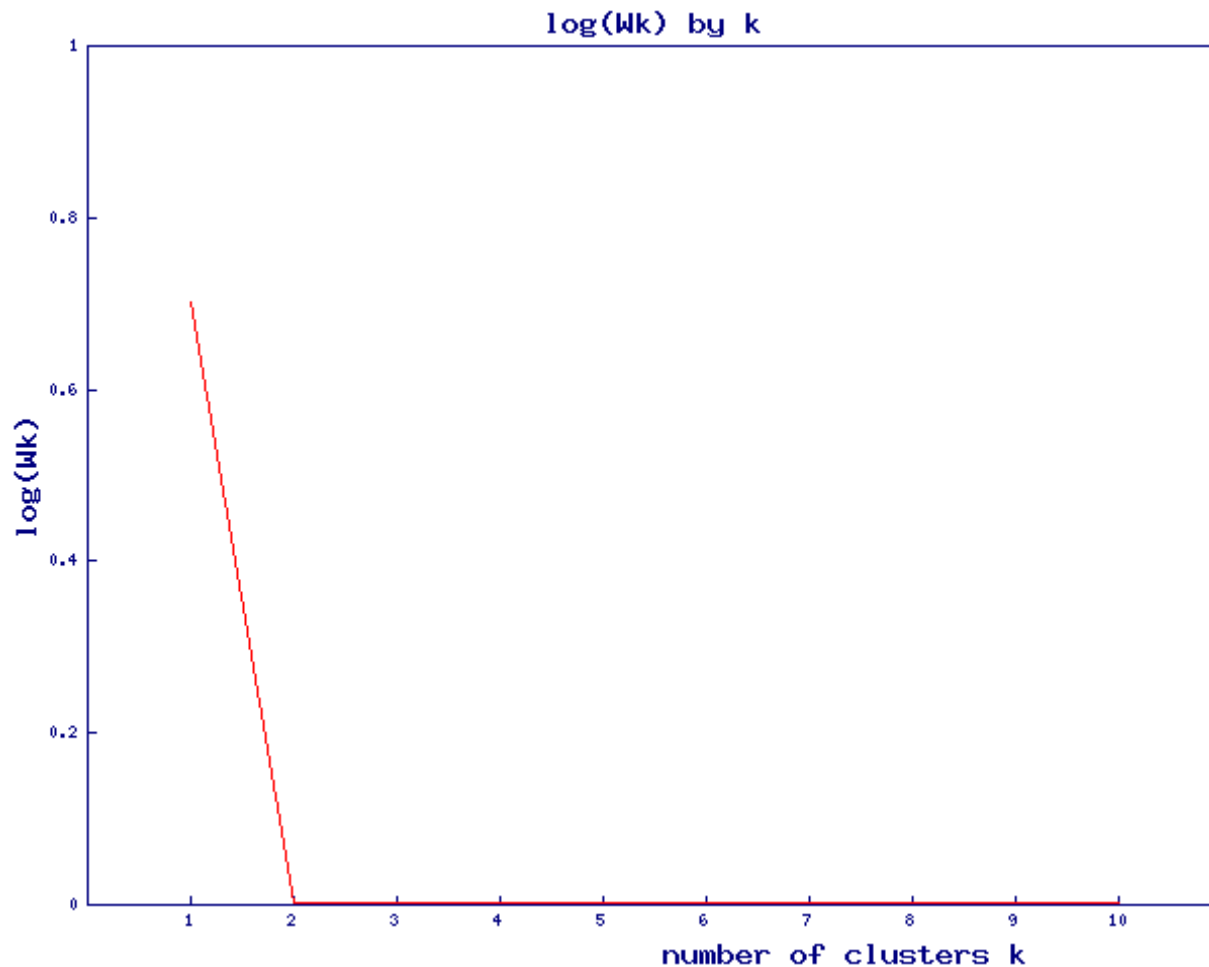
	Feature1	Feature2	Feature3	...	Row Marginal
Context1	1	0	1	...	24
Context2	1	0	0	...	11
Context3	0	1	1	...	56
Context4	1	0	1	...	18
.....					...
Column Marginal	12	28	39	...	83



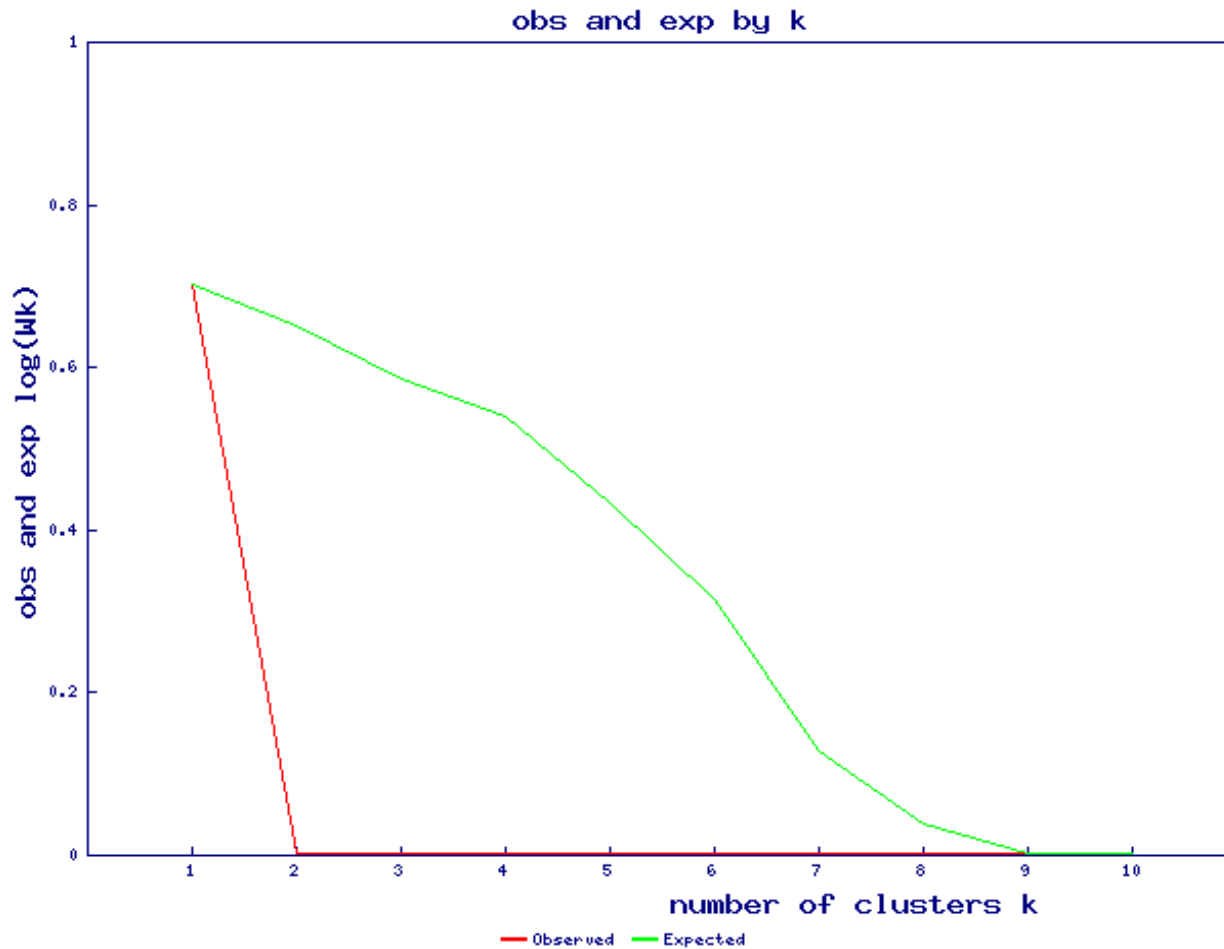
Gap Statistics - Algorithm

1. Calculate the Error (observed error) when the data is separated into k cluster.
2. Generate reference data B (100) times and find the average Error (reference error) when this data is separated into k clusters.
3. $\text{Gap}(k) = \text{reference_error}(k) - \text{observed_error}(k)$
4. Optimal k is the smallest k value for which the following is satisfied:
$$\text{Gap}(k) \geq \text{Gap}(k+1) - \text{standard_error}$$

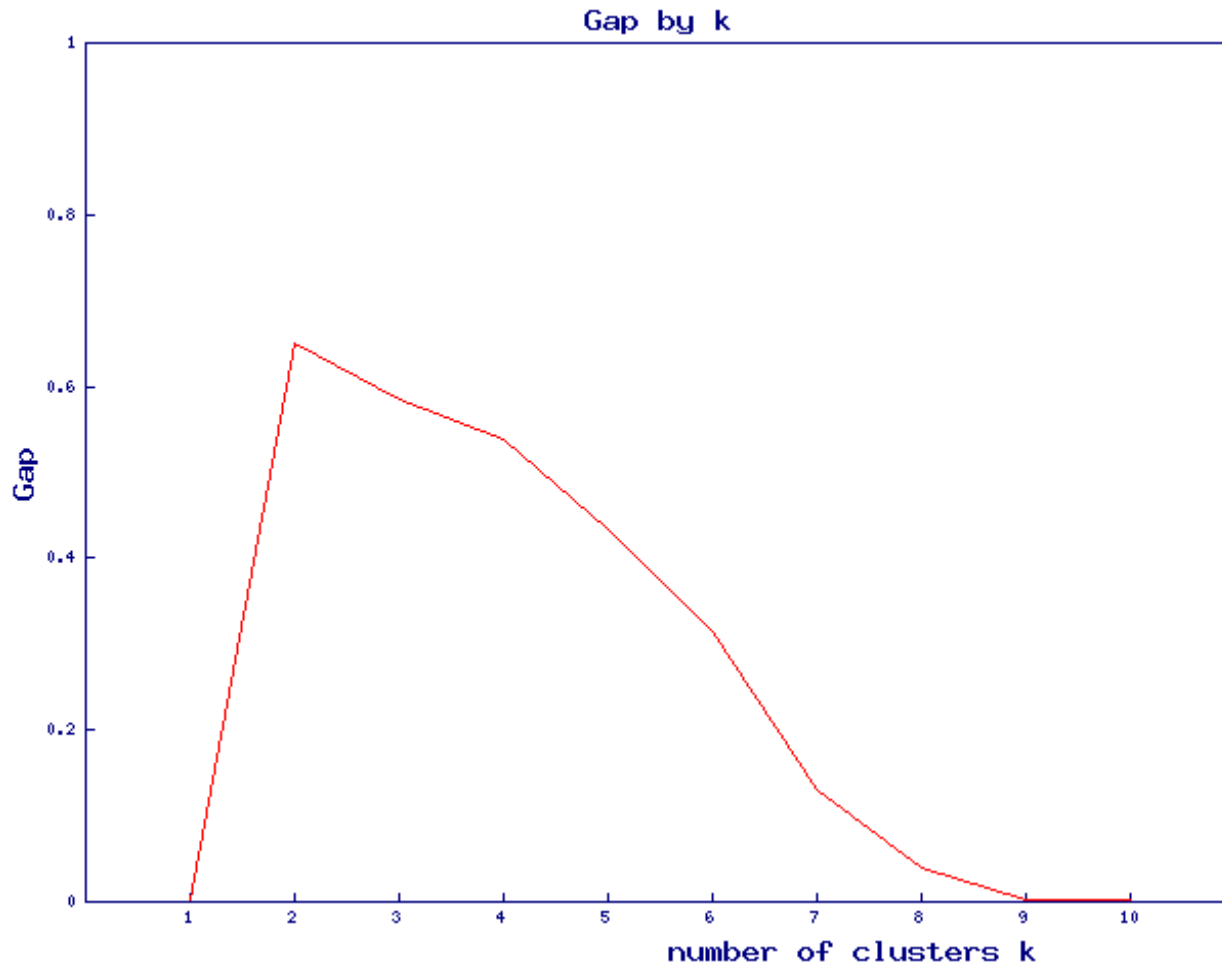
Gap Statistics – Graphs for the dataset with optimal # clusters = 2



Gap Statistics – Graphs for the dataset with optimal # clusters = 2



Gap Statistics – Graphs for the dataset with optimal # clusters = 2





Experiments





Experimental Setup

- We have experimented with 2 types of datasets from Medical Domain – Specifically the National Library of Medicine dataset:
 - Abstracts (41 ambiguous words)
 - Mixed Words (5 files - containing 2/3/4 ambiguated words per file) (The concept of creating mixed words developed by Purandare and Pedersen, 2004)
- Context representation as input to stopping rules:
 - PB3 - Feature type: bigrams
 - The setting created by Purandare and Pedersen, 2004;
 - Applied to NLM data by Savova et al., 2005
 - JPM - Feature type: unigrams
 - Joshi et al., 2005

Experimental Results in terms of:

words with correctly predicted sense

All words

		Abstracts (41 words) (baseline = 0.00)	Mixed Words (5 words) (baseline = 0.0)
PB3	C&H	0.49 (20)	0.20 (01)
	Hartigan	0.10 (04)	0.00 (00)
	Gap (unif)	0.02 (01)	0.00 (00)
	Gap (prop)	0.24 (10)	0.00 (00)
JPM	C&H	0.37 (15)	0.20 (01)
	Hartigan	0.02 (01)	0.00 (00)
	Gap (unif)	0.05 (02)	0.20 (01)
	Gap (prop)	0.12 (05)	0.20 (01)



Experimental Results (cont.) –

with ± 1 tolerance in the estimated # clusters

		Abstracts (41 words) (baseline = 0.82)	Mixed words (5 words) (baseline = 0.4)
PB3	C&H	0.73 (30)	0.6 (3)
	Hartigan	0.85 (35)	0.4 (2)
	Gap (unif)	0.83 (34)	0.4 (2)
	Gap (prop)	0.71 (29)	0.8 (4)
JPM	C&H	0.58 (24)	0.6 (3)
	Hartigan	0.85 (35)	0.4 (2)
	Gap (unif)	0.63 (26)	0.6 (3)
	Gap (prop)	0.59 (24)	0.6 (3)



Experimental Results (cont.) - Average number of senses

		Average number of senses predicted (Abstracts true # of senses = 2.19)	Average number of senses predicted (Mixed words true # of senses = 2.8)
PB3	C&H	2.90	2.40
	Hartigan	1.27	1.00
	Gap (unif)	1.49	1.00
	Gap (prop)	2.51	2.40
JPM	C&H	3.36	3.60
	Hartigan	1.10	1.00
	Gap (unif)	2.44	4.00
	Gap (prop)	2.59	4.00



Indirect Experimental Results: F-scores for the WSD task

		Abstracts (41 words) (Majority Sense = 82.63)	Mixed words (5 words) (Majority Sense = 38.47)
PB3	C&H	80.71	38.91
	Hartigan	82.15	38.47
	Gap (unif)	82.00	38.47
	Gap (prop)	81.31	38.70
JPM	C&H	80.27	39.01
	Hartigan	82.89	38.47
	Gap (unif)	81.63	39.02
	Gap (prop)	81.15	39.02



Discussion

- We have compared 3 stopping rules while using 2 context representations on two datasets
- This set of experiments provide a foundation for the problem of discovering the correct number of senses in an unsupervised manner
- Abstracts data set
 - C&H appears to be the best method without any tolerance results but with tolerance adjustments there is no single method that outperforms the others.
- Mixed Words data set
 - None of the three stopping rules performed well ... may be explained by very low kappa values.



Discussion (cont.)

- Context representation using bigrams (PB3) appear to be better for C&H and Gap Statistic.
- Hartigan appears to be insensitive to the context representation that we have experimented with.
- All three methods demonstrated reasonable approximation of the elbow region (with ± 1 tolerance)



Future Work

- Experimenting with different clustering algorithms.
- Experimenting with different context representations.
- Experimenting with different Mixed words dataset that includes words with high kappa values.
- Combining the stopping rules to get better estimates.
- Applying Singular Value Decomposition (SVD) to the reference distribution generation for the Gap Statistics.



References

- Schutze H.: Automatic Word Sense Discrimination Computational Linguistics, 24(1): (1998) 97 - 124.
- Pedersen T. and Bruce R.: Distinguishing word senses in untagged text. The Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, (1997) 197 - 207. Providence, RI
- Tibshirani R., Walther G. and Hastie T.: Estimating the number of clusters in a dataset via the Gap statistic. Journal of the Royal Statistics Society, 63(2): (2001) 411-424.
- Calinski T. and Harabasz J.: A dendrite method for cluster analysis. Communications in statistics, 3, (1974) 1 - 27.
- Hartigan J.: Clustering Algorithms. John Wiley and Sons, (1975) New York, NY



References (cont.)

- Milligan G. and Cooper M.: An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, 50(2): (1985) 159-179.
- Purandare A. and Pedersen T.: Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. The Proceedings of the Conference on Computational Natural Language Learning, May 6-7, 2004, Boston, MA
- Resolving Ambiguities in Biomedical Text with Unsupervised Clustering Approaches (Savova, Pedersen, Purandare and Kulkarni) - University of Minnesota Supercomputing Institute Research Report UMSI 2005/80 and CB Number 2005/21, May.
- Joshi M., Pedersen T. and Maclin R. - A Comparative Study of Support Vector Machines Applied to the Supervised Word Sense Disambiguation Problem in the Medical Domain. To appear in the proceedings of the 2nd Indian International Conference on Artificial Intelligence, 2005.