Unsupervised Context Discrimination and Cluster Stopping

Anagha Kulkarni
Department of Computer Science
University of Minnesota, Duluth

July 5, 2006

What is a “Context”?

- For the purpose of this thesis which deals with written text:
  - A Sentence
  - A Paragraph
  - Complete Text from a document

More generally any unit of text per se!

What is "Context Discrimination"?

Grouping contexts based on their mutual similarity or dissimilarity.

Example:

1. We had a very hot summer last year.

2. Germany is hosting FIFA 2006.

3. The weather in Duluth is highly dynamic and thus hard to predict.

4. England is out of World Cup 2006!

Word Sense Discrimination (WSD)

- **About**: Ambiguous words (target or head word).

- **Task**: To group the given contexts based on the meaning of the ambiguous word.

Example:

1. Let us roll this sheet and bind it with a ***tape***.
2. I prefer this brand of ***tape*** over any other because it binds the best.
3. As she sang the melodious song he recorded her on the ***tape***.
4. As he moved forward to adjust the volume of the ***tape*** playing this loud song…

Name Discrimination

- **About**: People, places, organizations sharing same name (target or head word).

- **Task**: To group the given contexts based on the underlying entity of the ambiguous name.

Example:

1. **George Miller** is an Emeritus Professor of Psychology at the Princeton University and is often referred to as the father of the WordNet.
2. The Mad-Max movie made the Australian director, **George Miller**, a celebrity overnight.
3. **George Miller** is an acclaimed movie director.

Email Clustering

- **About**: Email grouping

- **Task**: To group the given emails based on the similarity of their contents. ***Headless*** Clustering!

  Example:

  1. "Hi, I'm looking for a program which is able to **display** 24 bit **images**. We are using a **Sun Sparc** equipped with **Parallax graphics** board running **X11**. Thanks in advance."
  2. "I currently have some **grayscale image** files that are not in any standard **format**. They simply contain the 8-bit **pixel** values. I would like to **display** these **images** on a **PC**. The conversion to a **GIF format** would be helpful. "
  3. "I really feel the need for a knowledgeable **hockey** observer to explain this year's **playoffs** to me. I mean, the obviously superior Toronto **team** with the best center and the best **goalie** in the **league** keeps losing."

What is “Unsupervised Context Discrimination”?

Discriminating Contexts:

- Without using any labeled/tagged data.
- Without using external knowledge resources
- Using only what is present in the contexts!

- Why?
  - To avoid the knowledge acquisition bottleneck
  - To keep the method applicable across domains
  - To keep the method applicable across languages
  - To keep the method applicable across time

Approach to WSD by Purandare & Pedersen [2004]

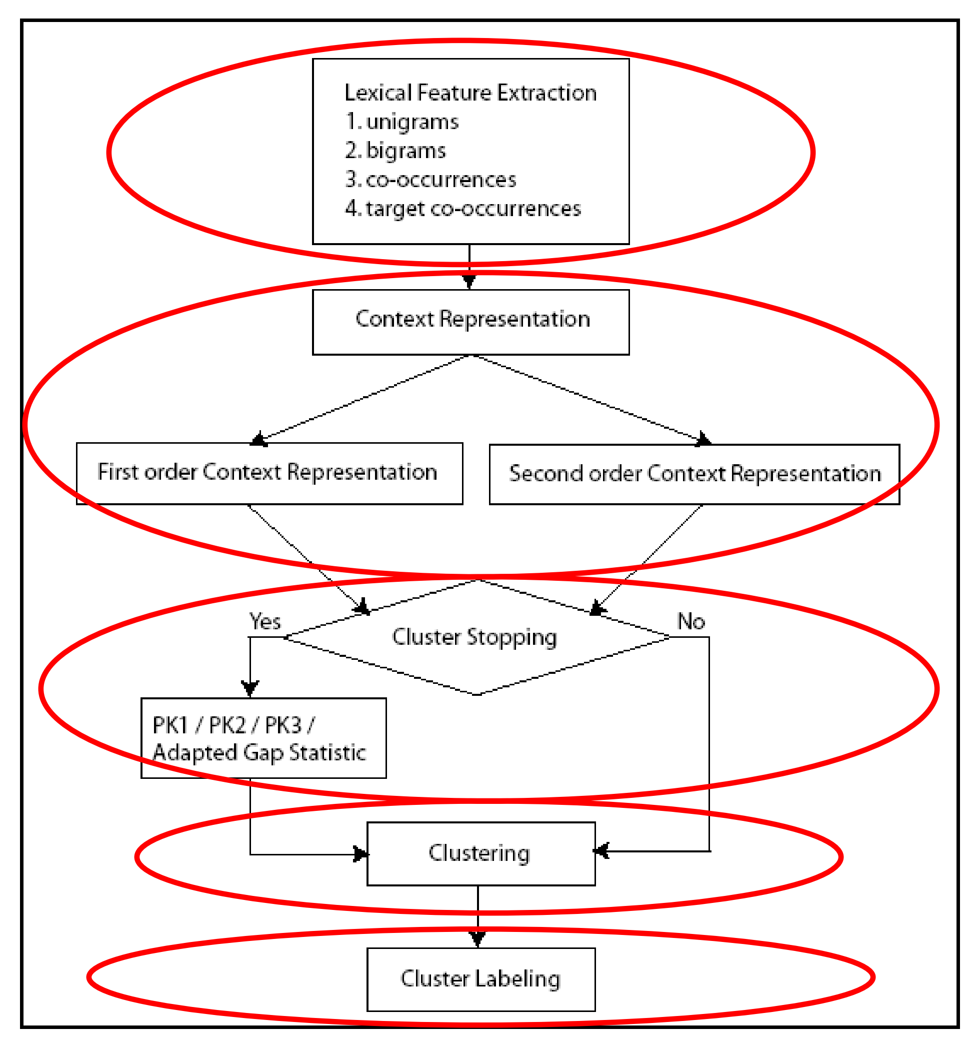
Based on the hypothesis of Contextual Similarity by Miller and Charles (1991):

"any two words are semantically similar to the extent that their contexts are similar"

Major contributions of this thesis

- Generalized Purandare and Pedersen [2004] approach for WSD to the broader problem of Context Discrimination.
- Introduced three measures for the cluster stopping problem.
- Introduced preliminary method of cluster labeling.

Methodology: 5 Steps


Lexical Feature Extraction
1. unigrams
2. bigrams
3. co-occurrences
4. target co-occurrences
Context Representation
First order Context Representation
Second order Context Representation
Yes
Cluster Stopping
No
PK1 / PK2 / PK3 / Adapted Gap Statistic
Clustering
Cluster Labeling

Step1

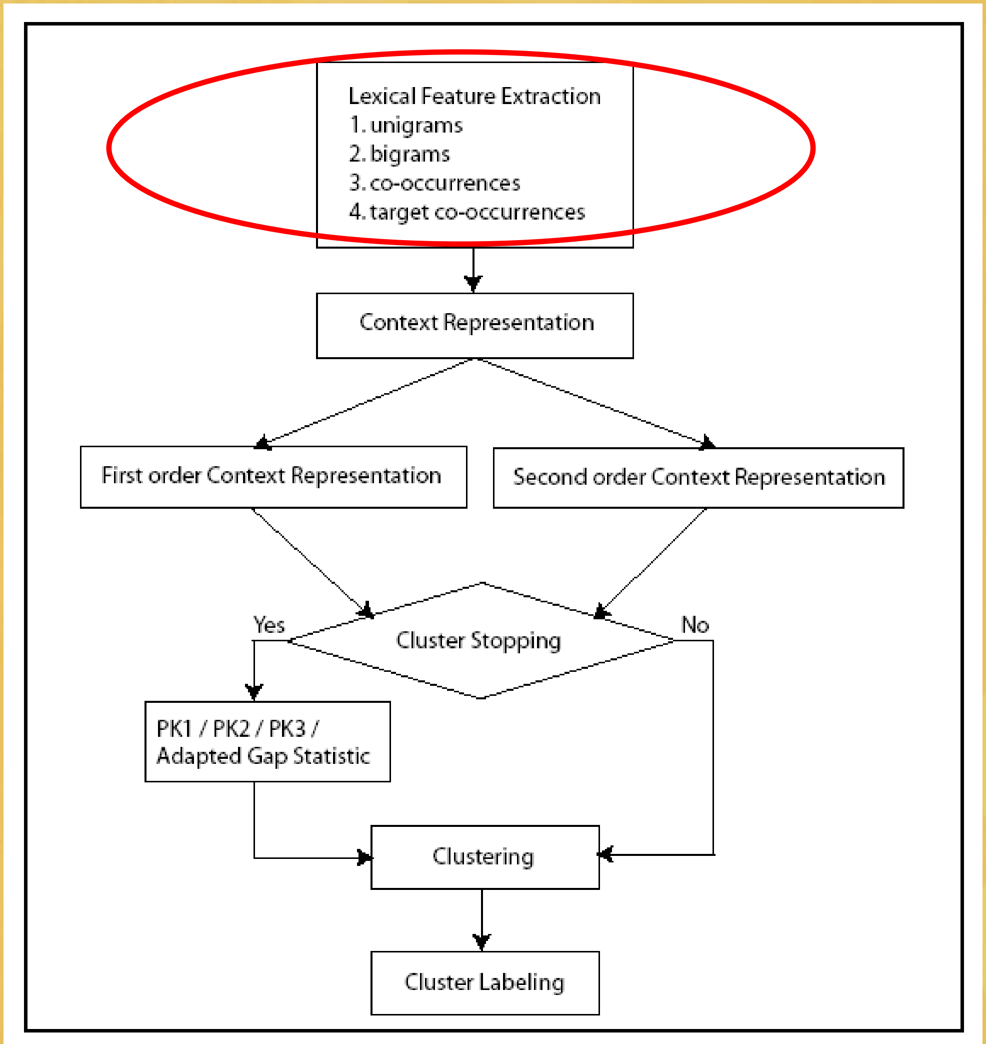
Step2

Step3

Step4

Step5

Methodology: Lexical Feature Extraction



Step1

Lexical Features

- Lexical Features: Are the words or word-pairs of a language that can be used to represent the given contexts.

- Can be selected from: the test data or a separate feature selection data.

- No external knowledge in any shape or form used.

- No syntactic information about the features used either.

Example:
Movie
Professor
Director
Psychology
Mad-Max
Princeton
Australia
WordNet

George Miller is a Emeritus **Professor** of **Psychology** at the **Princeton** University and is often referred to as the father of the **WordNet**.

Types of Lexical Features

- Unigrams: Single words.

  Example: Movie, Professor, Director, Psychology…

- Bigrams: Ordered word-pairs.

  Example: Movie Director, Princeton University…

- Co-occurrences: Unordered word-pairs.

  Example: Director Movie, Princeton University…

- Target Co-occurrences: Unordered word-pairs of which one of the words is the target word.

  Example: *tape* playing, binding *tape*…

Feature Filtering Techniques

- **Frequency cutoff:** Remove features occurring less than X times. To remove rare features.

- **Stoplisting:** To remove function words such as “the”, ”of”, ”in”, ”a”, ”an” etc.

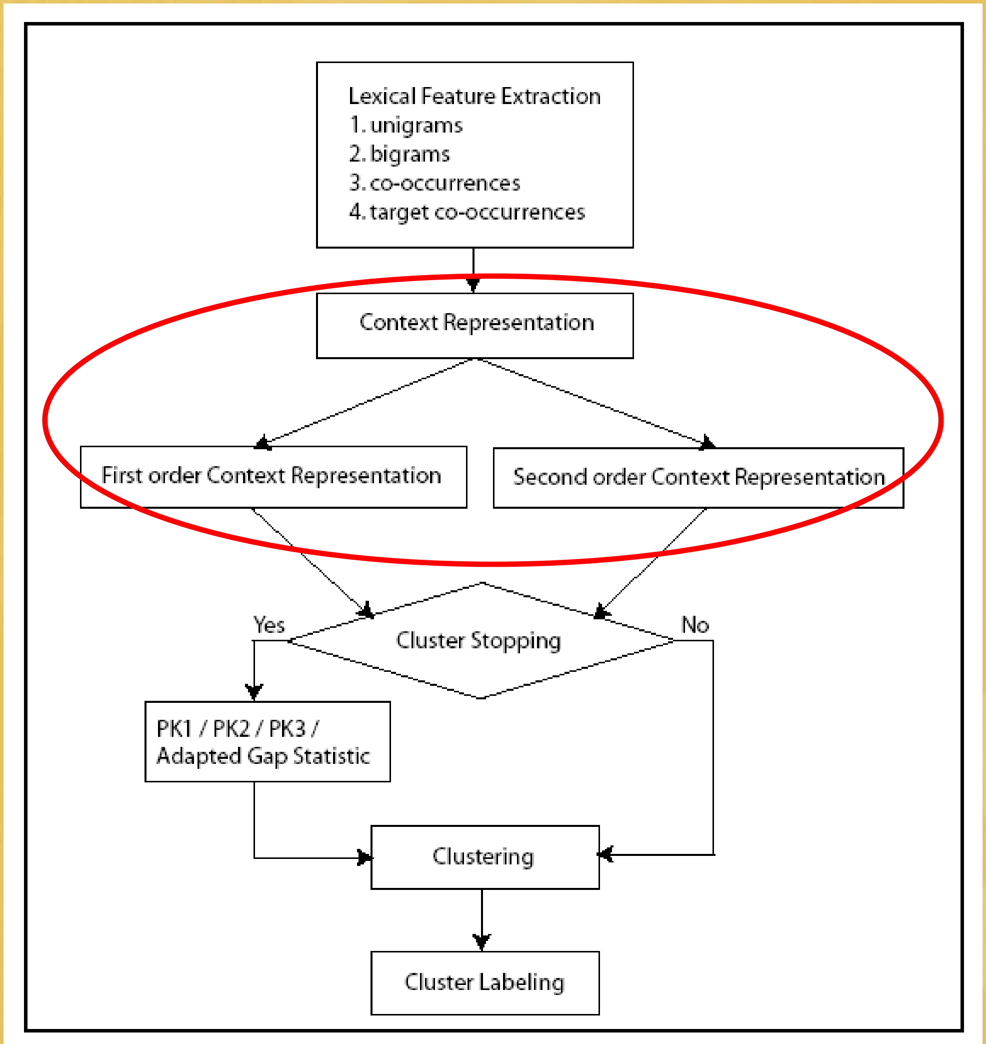
  For bigrams and co-occurrences:

  – OR Mode: Remove if either of the words is a stopword.

  – AND Mode: Remove only if both the words are stopwords.

- **Statistical tests of association** (bigrams, co-occurrences): To check if the two words in a word-pair occur together just by chance or they are truly related.

Methodology: Context Representation



Step2

Context Representation

The task of translating each textual context into a format that a computer can understand.

Context vector: C1

Example:

- Context1: George Miller is an Emeritus Professor of Psychology at the Princeton University and is often referred to as the father of the WordNet.

- Context2: The Mad-Max movie made the Australian director, George Miller, a celebrity overnight. → **Context vector: C2**

First Order Context Representation (Order1)

	Movie	Professor	Director	Psychology	Mad-Max	Princeton	Australian
Context1	0	1	0	1	0	1	0
Context2	1	0	1	0	1	0	1

Second Order Context Representation (Order2)

Tries to go beyond the “exact match” strategy of Order1 by capturing indirect relationships.

Example

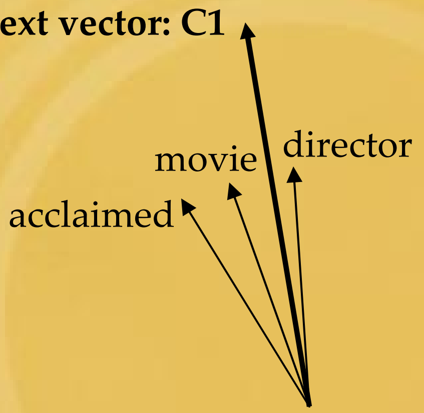
1. George Miller is an acclaimed movie director.
2. George Miller has since continued his work in the film industry.
3. Film director George Miller in the news for “Mad-Max”.

Order2: Step1: Creating the word-by-word matrix

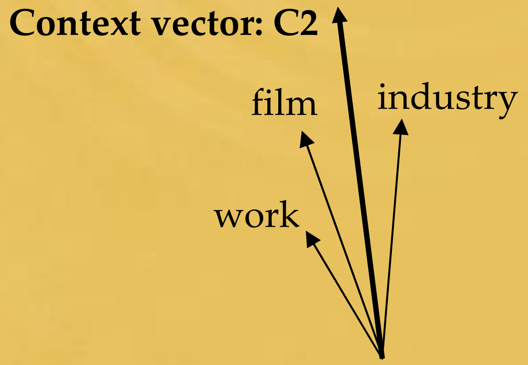
	Director	University	Mad-Max	Psychology	Industry	…
Movie	1	0	0	0	0	0
Professor	0	1	0	1	0	0
Princeton	0	1	0	0	0	1
Film	1	0	0	0	1	0
Australian	1	0	1	0	0	0
Celebrity	1	0	0	0	1	0
Father	0	0	0	0	0	1
…	1	0	1	0	1	0

Order2: Step2: Creating the context vectors

- George Miller is an acclaimed movie director.



- George Miller has since continued his work in the film industry.



Singular Value Decomposition (SVD)

Order1 matrix: M1

	Movie	Professor	Director	Psychology	Mad-Max	Princeton	Australian	University
Context1	0	1	0	0	0	1	0	1
Context2	0	0	0	1	0	1	0	1
Context3	0	1	0	1	0	0	0	0
Context4	1	0	0	0	1	0	1	0
Context5	0	0	1	0	0	0	1	1
Context6	1	0	1	0	1	0	0	0

SVD reduced matrix: M1$_{\textbf{reduced}}$

	d1	d2	d3	d4
Context1	0.7859	-0.5961	0.0579	-0.3261
Context2	0.7859	-0.5961	0.0579	-0.3261
Context3	0.3546	-0.3662	0.7115	0.7662
Context4	0.5385	0.8373	0.3087	-0.1271
Context5	0.7716	0.2139	-0.8758	0.4897
Context6	0.5385	0.8373	0.3087	-0.1271

SVD (cont.)

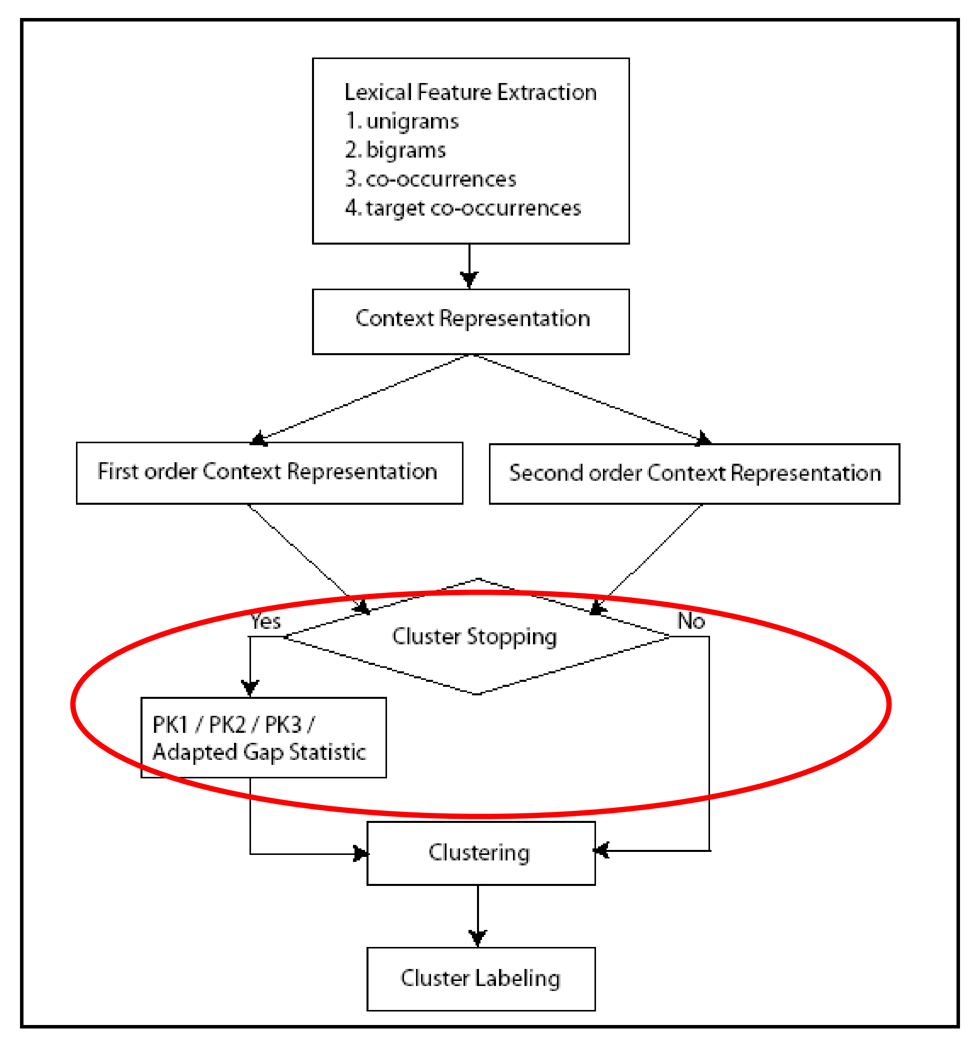
Order2: Step1: Word-by-word matrix: M2

	Director	University	Max	Psychology	Overnight	WordNet
Movie	1	0	0	0	0	0
Professor	0	1	0	1	0	0
Princeton	0	1	0	0	0	1
Mad	1	0	1	0	0	0
Australian	1	0	0	0	0	0
Celebrity	1	0	0	0	1	0
Father	0	0	0	0	0	1

	d1	d2	d3
Movie	-0.6360	0	0
Professor	0	-0.7933	-0.8230
Princeton	0	-0.9893	0.3663
Mad	-0.8145	0	0
Australian	-0.6360	0	0
Celebrity	-0.8145	0	0
Father	0	-0.4403	0.6600

SVD reduced matrix: $M2_{reduced}$

Methodology: Predicting *k* via Cluster Stopping



Step3

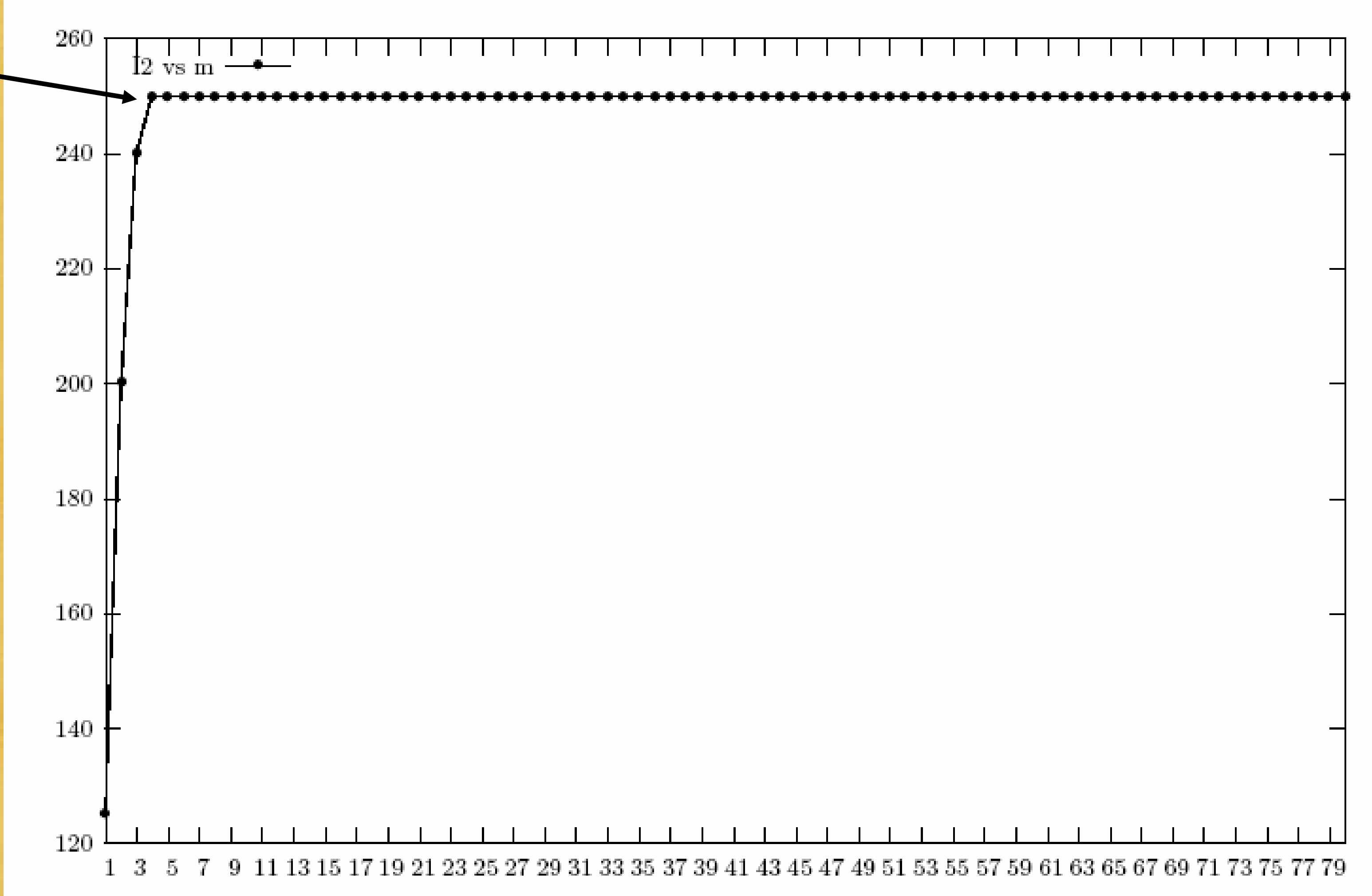
Building blocks of Cluster Stopping

- Criterion functions (crfun): Metric that the clustering algorithms use to assess and optimize the quality of the generated clusters.

- Types:
  - Internal: Maximize within cluster similarity (I1, I2)
  - External: Minimize between cluster similarity (E1)
  - Hybrid: Internal + External (H1, H2)

- Cluster a dataset iteratively into *m* clusters and record *crfun(m)* values…

Contrived dataset: #contexts = 80, expected $k = 4$

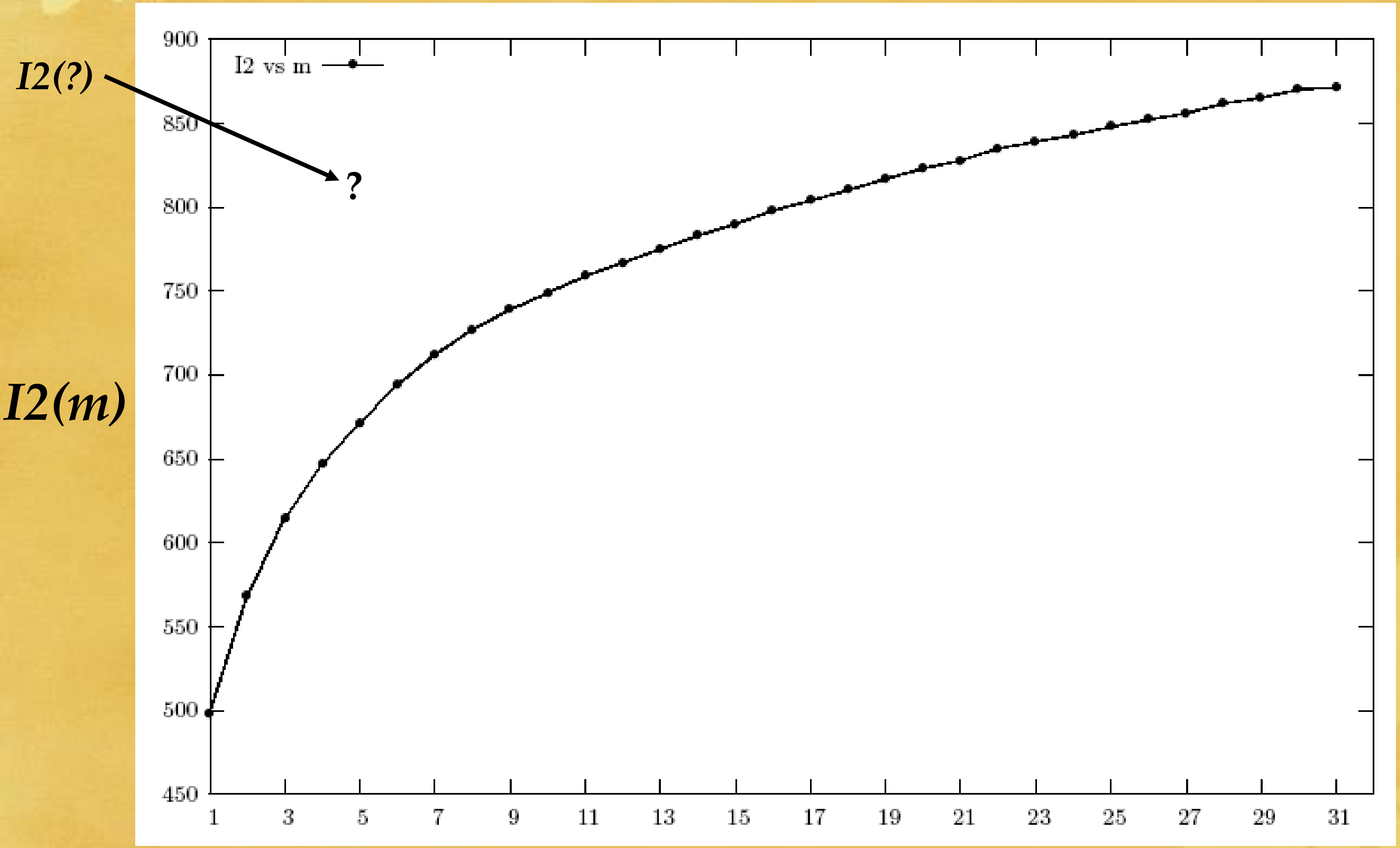


I2(4)

I2(m)

m

Real dataset: #contexts = 900, expected k = 4 (DS)



Cluster Stopping Measures

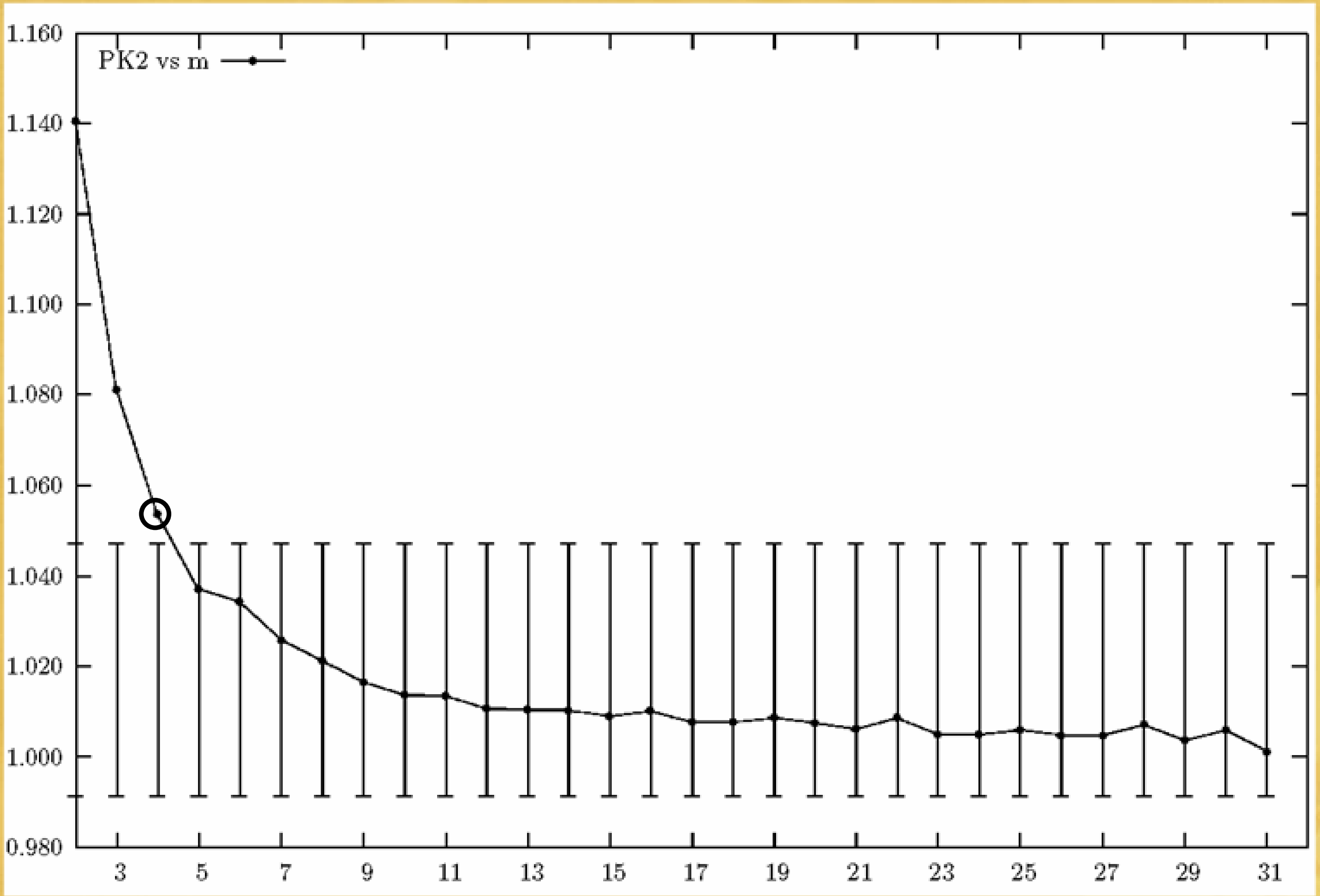
- Based on the criterion functions.

- Do not require any form of user input such as setting a threshold value.

- 3 measures:
  - PK2
  - PK3
  - Adapted Gap Statistic

$$PK2(m) = \frac{crfun(m)}{crfun(m-1)}$$

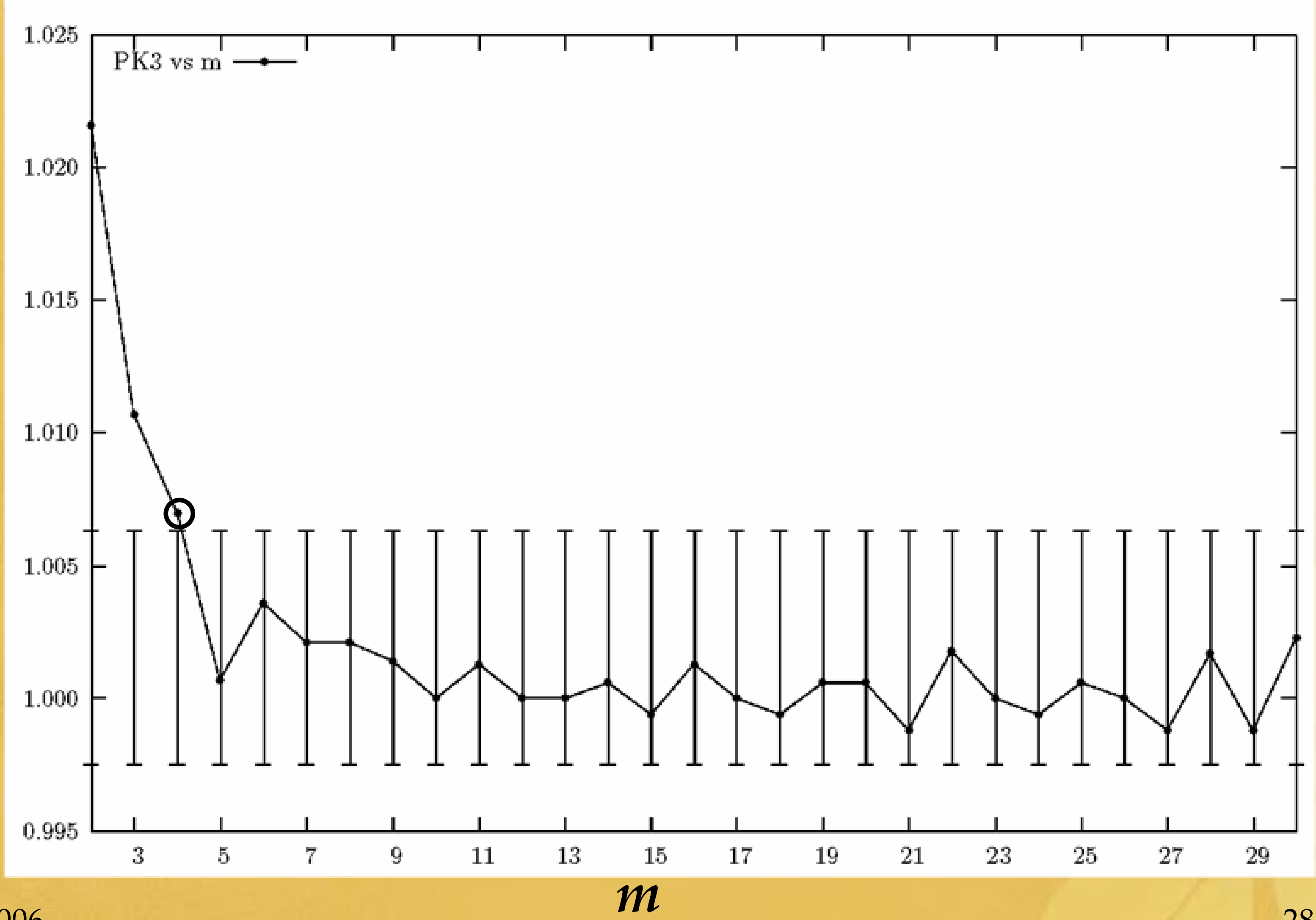
***PK2(m)* for DS**



m

$$PK3(m) = \frac{2 * crfun(m)}{crfun(m-1) + crfun(m+1)}$$

***PK3(m)* for DS**



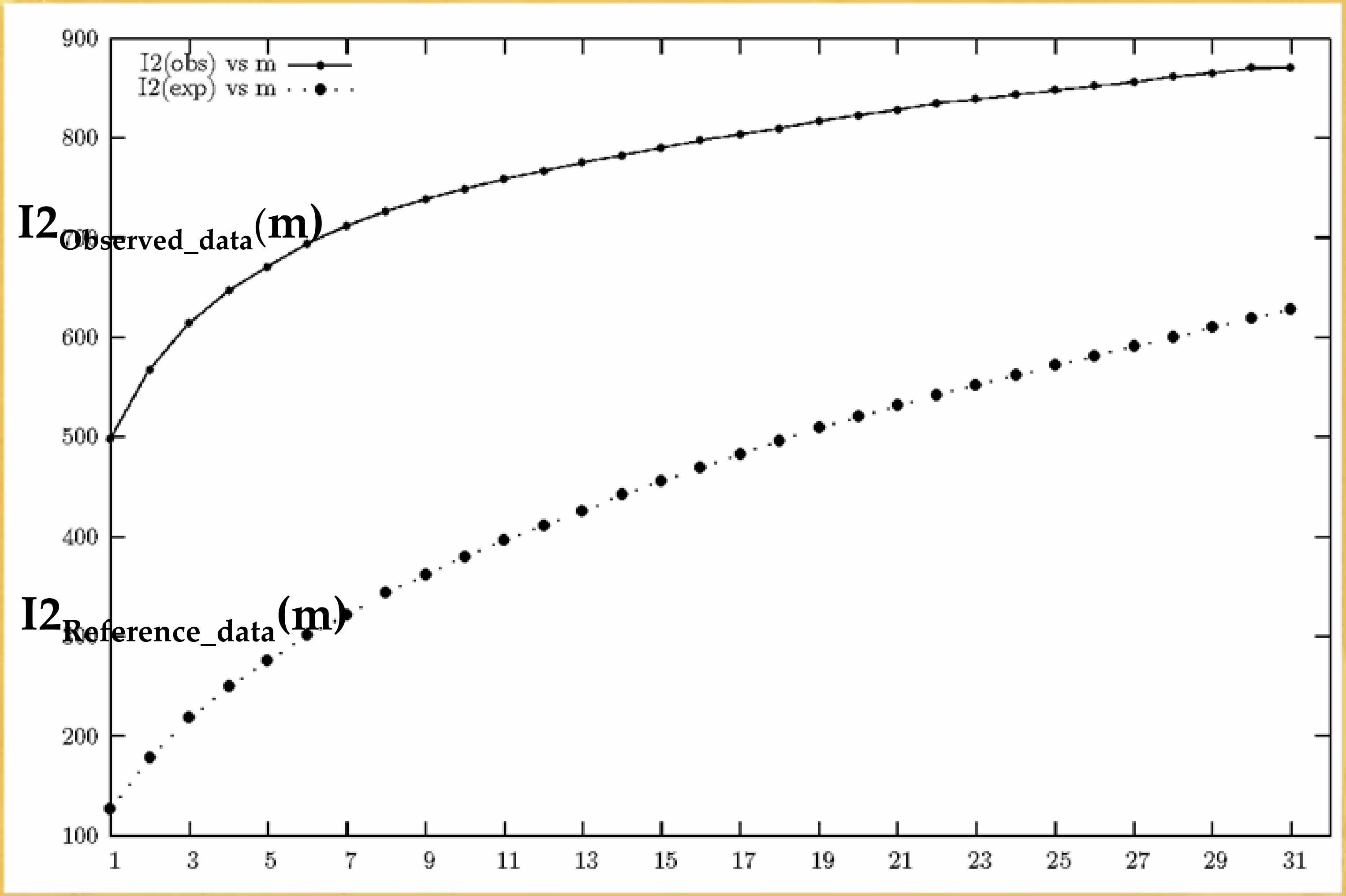
Adapted Gap Statistic

- Based on Gap Statistic by Tibshirani et al. (2001)

- The main idea:
  - Null hypothesis: H0: For the given dataset optimal *k* = *1.*
  - Alternative hypothesis: H1: For the given dataset optimal k > 1

- Algorithm:
  - Generate a data for the null reference model with expected *k* = *1.*
  - Generate a plot ($P_{Observed}$) of crfun(m) values for the given or observed data.
  - Generate a plot ($P_{Reference}$) of crfun(m) values for the generated reference data.
  - Compare $P_{Observed}$ with the $P_{reference}$ and find the largest “gap” between them.
  - The first point of maximum gap is the optimal *k* value!

Adapted Gap Statistic



$I2_{Observed\_data}(m)$

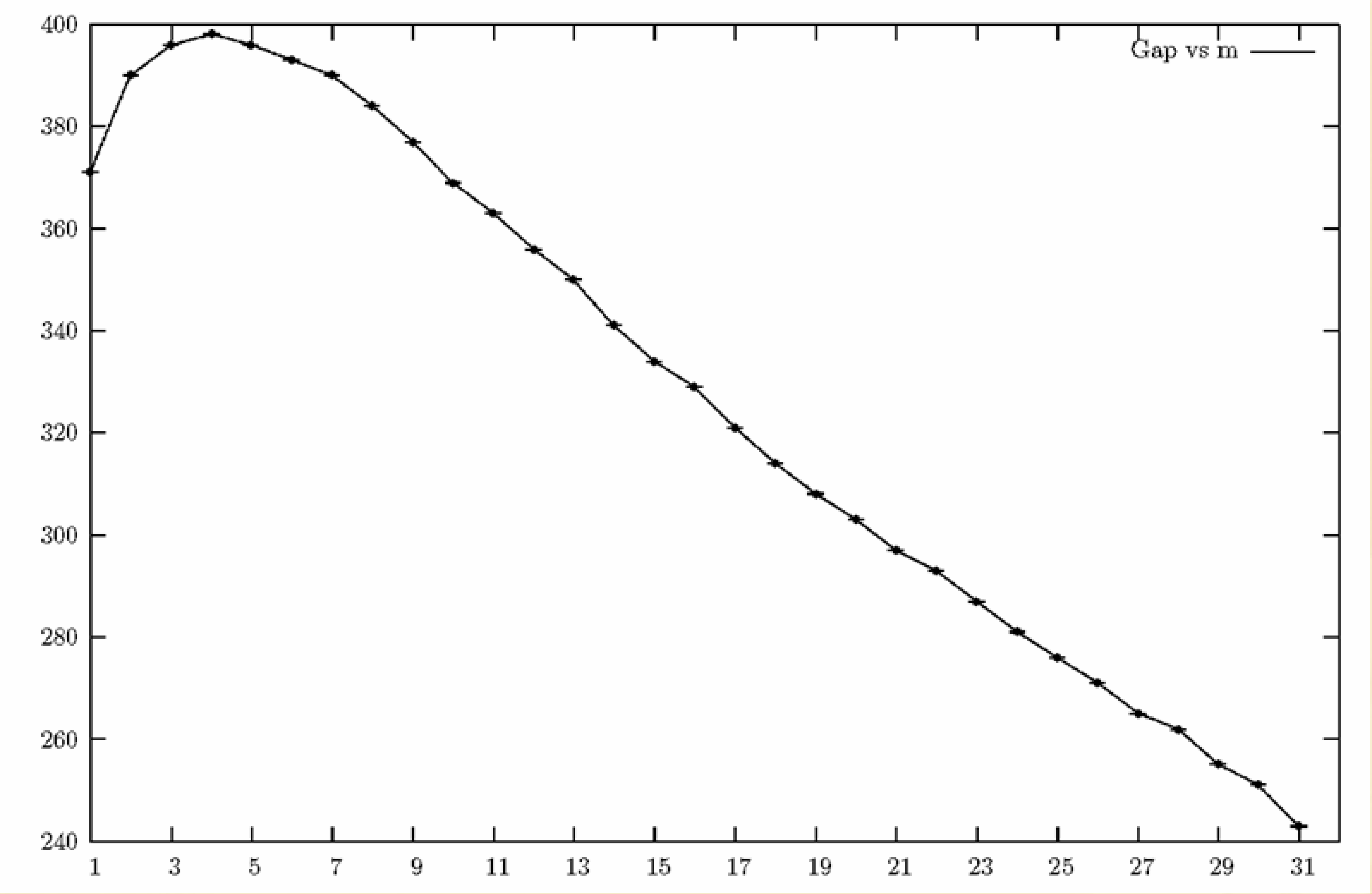
for DS

$I2_{Reference\_data}(m)$

m

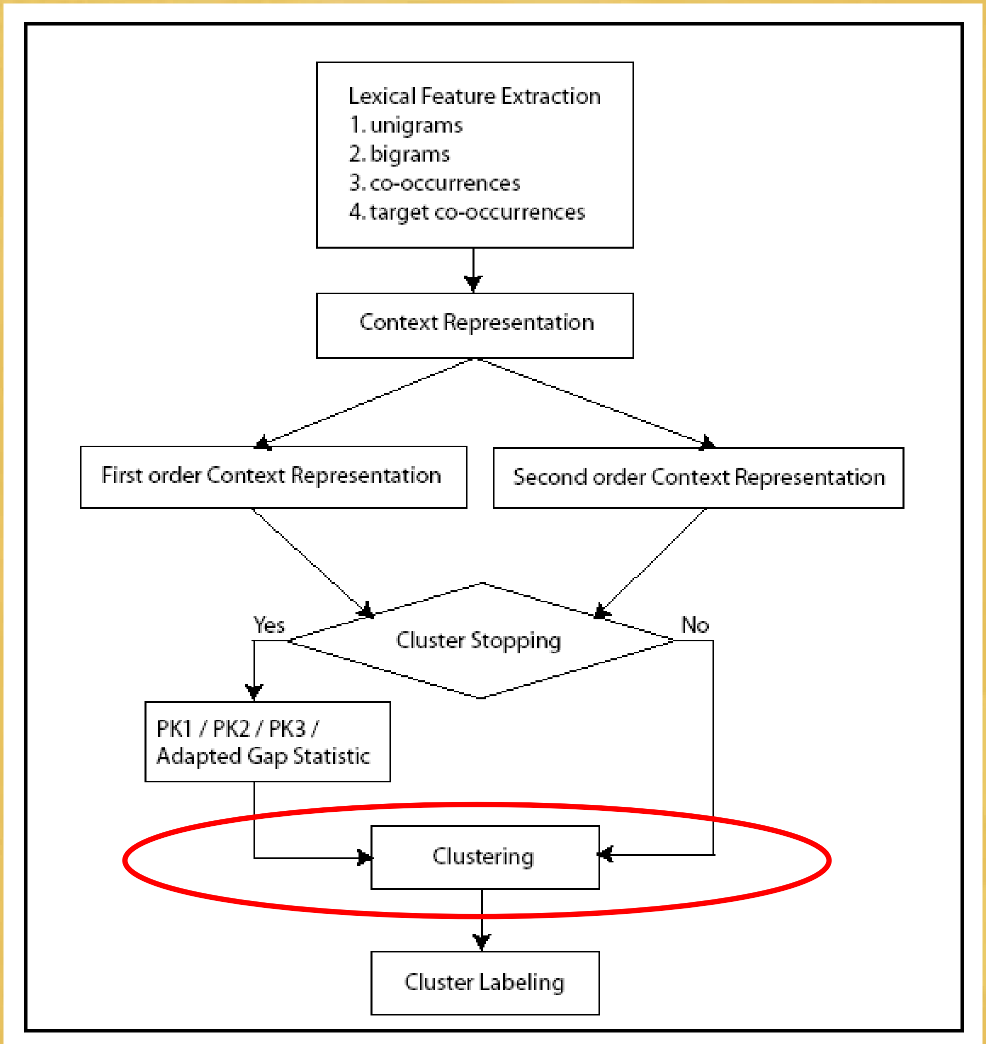
Adapted Gap Statistic (cont.)

Gap(m)



m

Methodology: Clustering



Step4

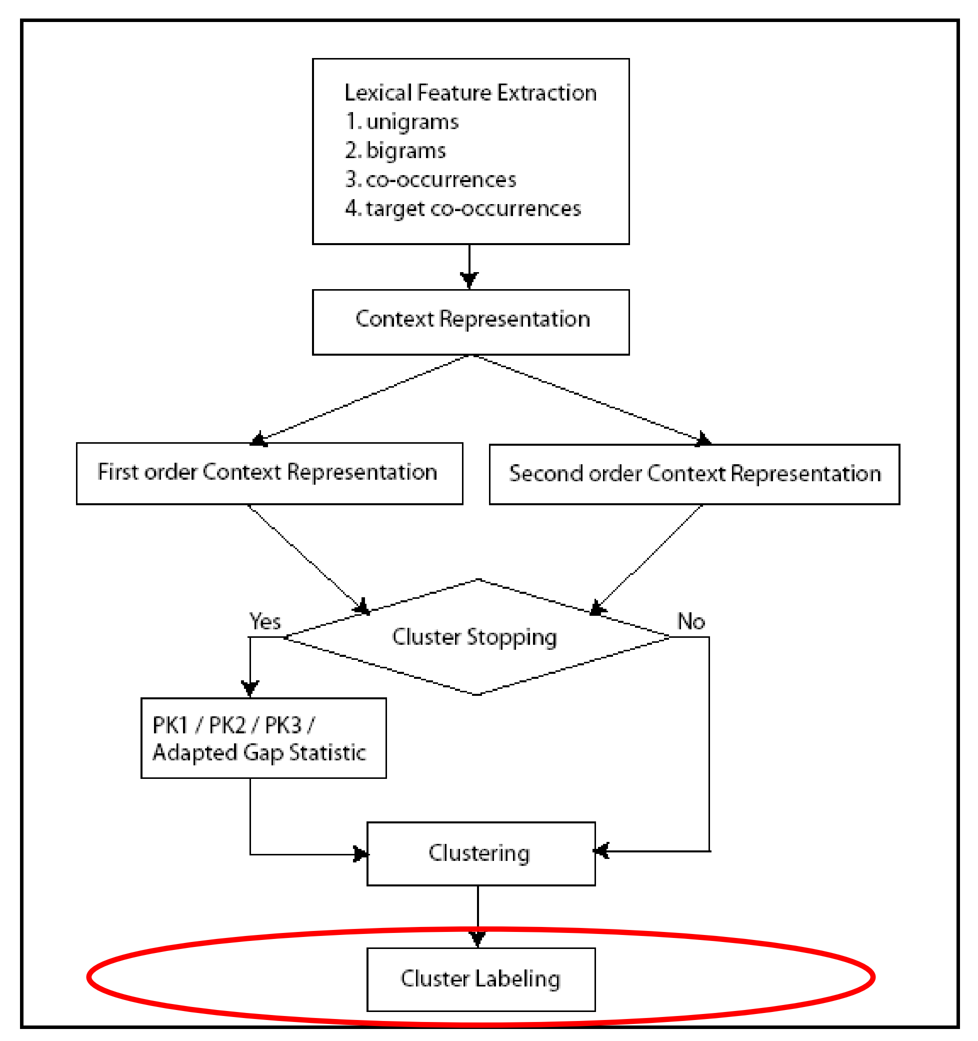
Clustering

- One of the primary methods of unsupervised learning.

- We support 3 types of clustering algorithms:
  - Hierarchical (e.g.: Agglomerative)
  - Partitional (e.g.: K-means)
  - Hybrid (e.g.: Repeated Bisections)

- Aim: To appropriately group the given set of context vectors into *k* clusters.

Methodology: Cluster Labeling


Lexical Feature Extraction
1. unigrams
2. bigrams
3. co-occurrences
4. target co-occurrences
Context Representation
First order Context Representation
Second order Context Representation
Cluster Stopping
Yes
No
PK1 / PK2 / PK3 / Adapted Gap Statistic
Clustering
Cluster Labeling

Step5

Cluster Labeling

- **Aim**: To identify the underlying entity for each cluster.
- **Descriptive labels**: Top N bigrams of that cluster.
- **Discriminating labels**: Top N bigrams unique to that cluster.
- Can use frequency or statistical tests of association (like in feature selection) to select the top N bigrams.

Cluster labels for an ambiguous name *Richard Alston:*

Clusters	**Assigned Cluster Labels**
C0: Australian Senator	Communications Information, Media Release, Minister Communications, Information Technology
C1: Choreographer	Artistic Director, Dance Company

Experimental Data – 4 genre

NameConflate genre

- Name discrimination data.
- **Source**: *The New York Times* archives (Jan `02 to Dec `04)
- **Method**: Creating pseudo ambiguity by conflation.
- **Multi-dimensional ambiguity**: 2, 3, 4, 5 or 6 names.
- **Distinct** (e.g. “Bill Gates” & “Jason Kidd”)
  - 7 datasets
- **Subtle** (e.g. “Bill Gates” & “Steve Jobs”)
  - 6 datasets

Web genre

- Name discrimination data.

- **Source**: The World Wide Web using Google search engine
  - Contents from top 50 (html) pages.
  - Traversed one level deep.

- **Method:** Manually cleaned and annotated.

- **Name variations:** “Mr. Miller”, “Dr. Miller”, “G. Miller”…

- **5** datasets
  - **Richard Alston**, 2 entities, 247 contexts.
  - **Sarah Connor**, 2 entities, 150 contexts
  - **George Miller**, 3 entities, 286 contexts
  - **Michael Collins**, 4 entities, 333 contexts
  - **Ted Pedersen**, 4 entities, 359 contexts

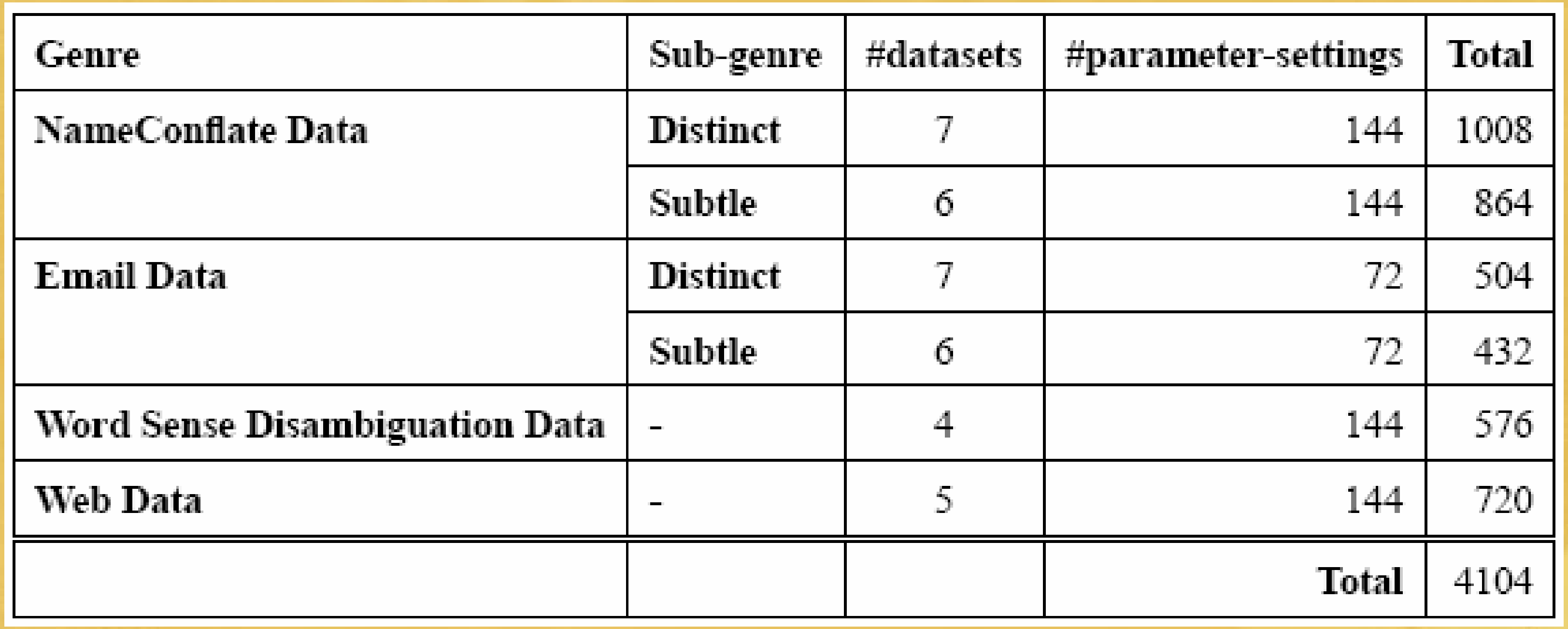
Email genre

- Email Clustering data.
- **Source**: 20 Newsgroups dataset
  - 20, 000 USENET posting manually categorized into 20 groups.
  - e.g.: comp.graphics and rec.sport.hockey
- **Method**: Creating artificial mixing of contexts by combining posting from two or more groups.
- **Multi-dimensional ambiguity**: Conflated 2, 3 or 4 groups.
- **Distinct** (e.g. “sci.electronics” & “soc.religion.christian”)
  - 7 datasets
- **Subtle** (e.g. “sci.crypt” & “sci.electronics”)
  - 6 datasets

WSD genre

- Word Sense Discrimination data.
- Datasets for **4** ambiguous words: “hard”, “serve”, “line” and “interest”.
- **Source**: The cleaned and SENSEVAL2 formatted versions of these datasets distributed by Dr. Ted Pedersen.

Experiments

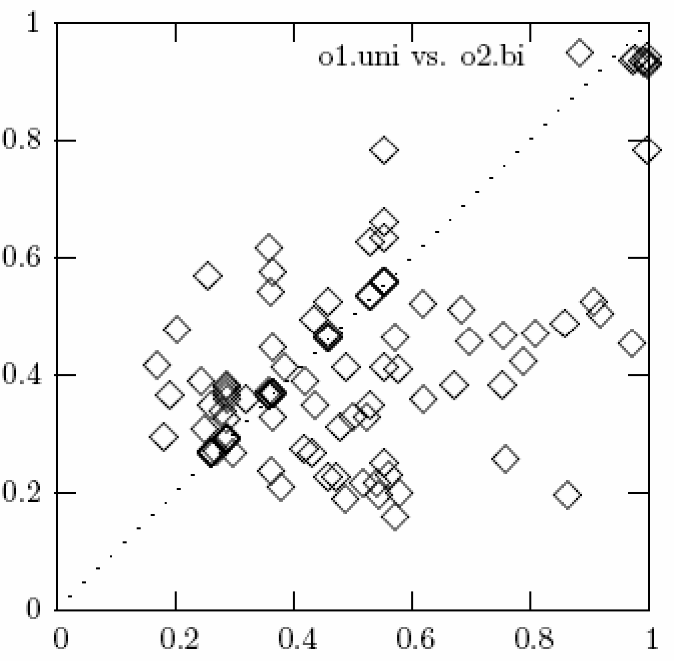


Genre	Sub-genre	#datasets	#parameter-settings	Total
NameConflate Data	Distinct	7	144	1008
	Subtle	6	144	864
Email Data	Distinct	7	72	504
	Subtle	6	72	432
Word Sense Disambiguation Data	-	4	144	576
Web Data	-	5	144	720
			Total	4104

Experimental Results

Order1 and unigrams vs. Order2 and bigrams

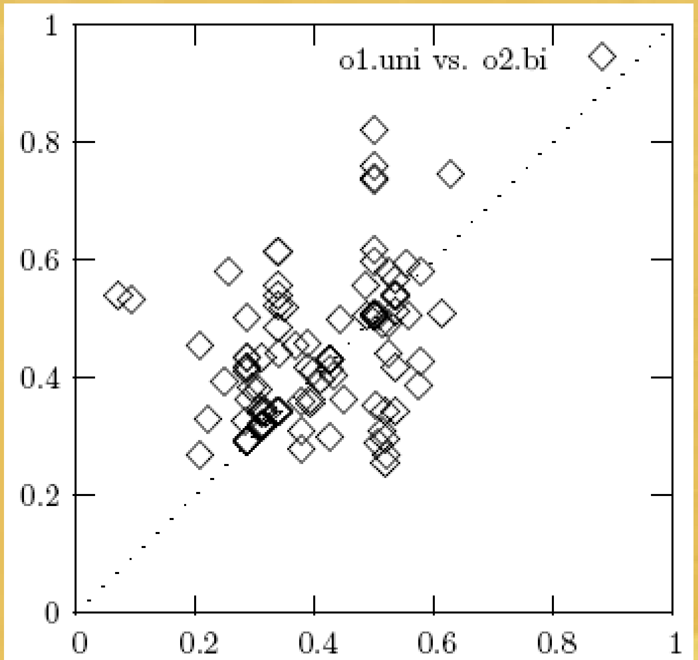
F-measure using Order2 & bigrams



F-measure using Order1 & unigram

NameConflate-Distinct

F-measure using Order2 & bigrams

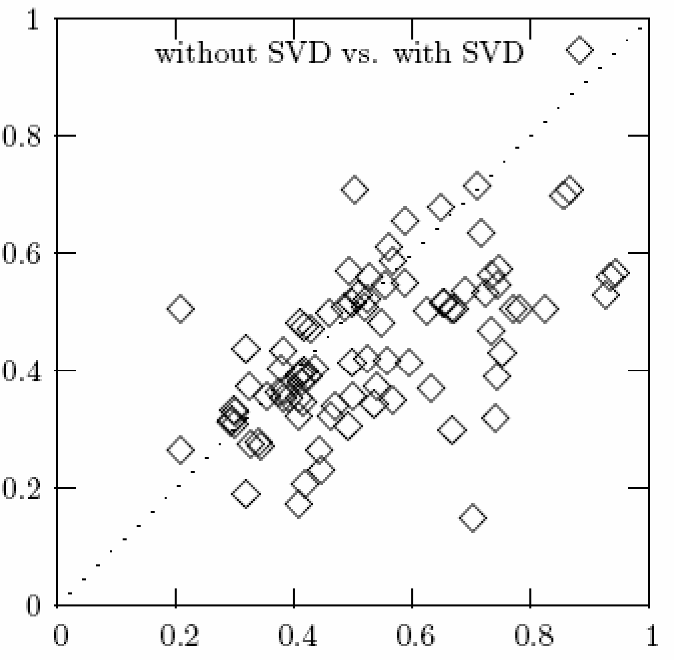


F-measure using Order1 & unigram

NameConflate-Subtle

Without SVD vs. With SVD

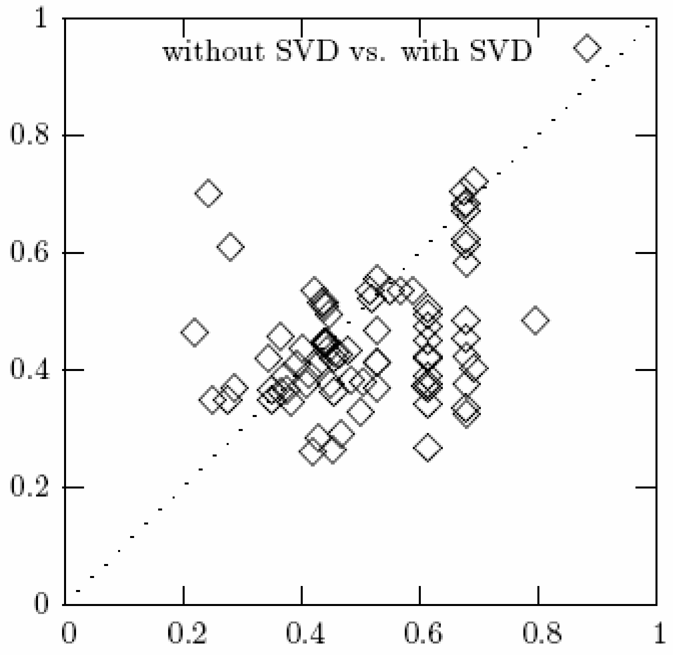
F-measure
With SVD



F-measure
Without SVD

Email-Distinct

F-measure
With SVD

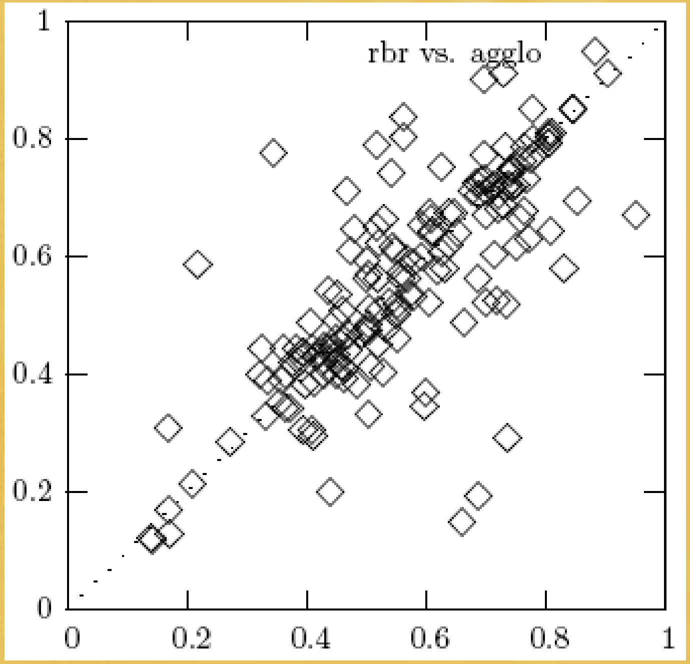


F-measure
Without SVD

WSD

Repeated Bisection vs. Agglomerative Clustering

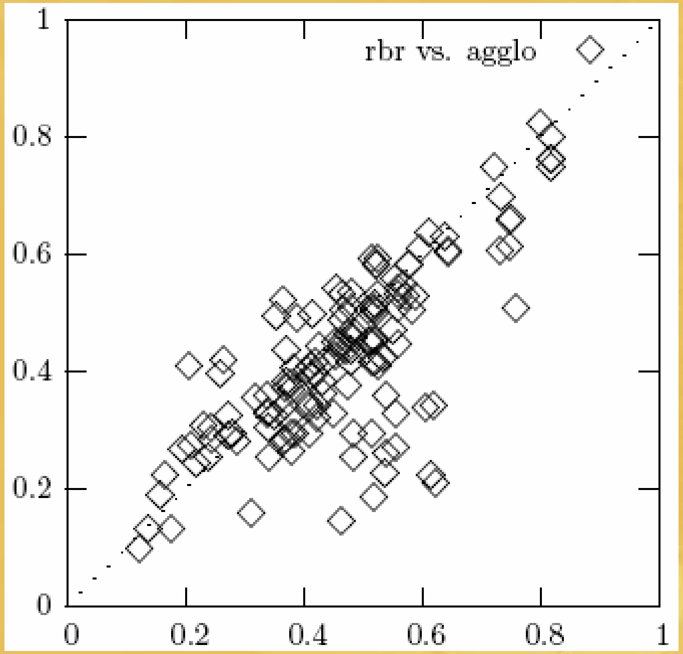
F-measure using Agglomerative



F-measure using Repeated Bisections

Web

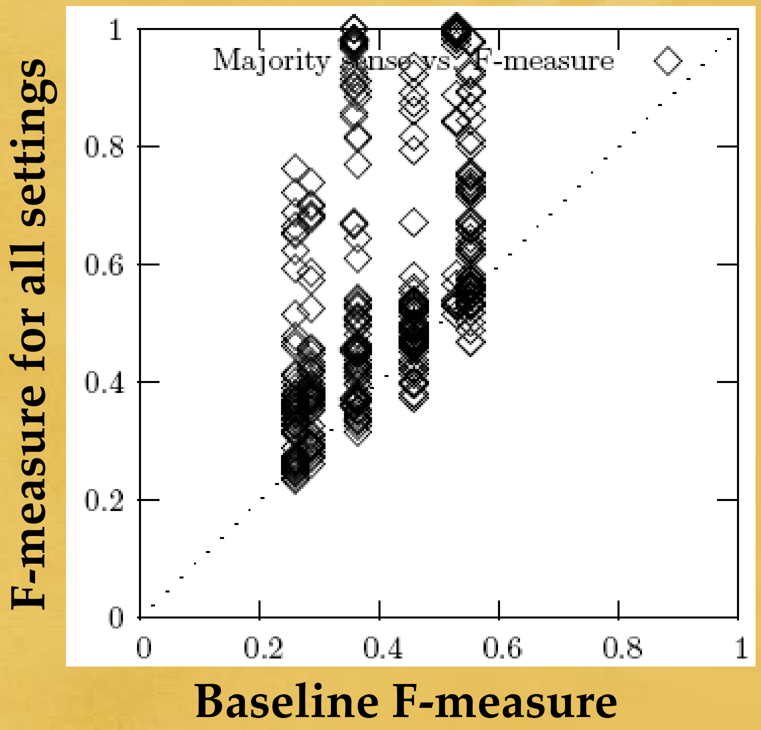
F-measure using Agglomerative



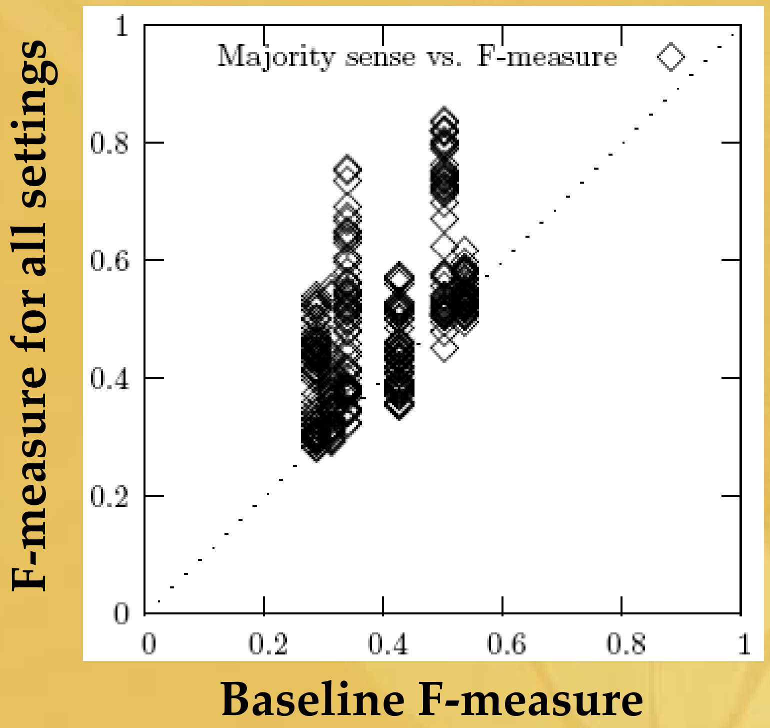
F-measure using Repeated Bisections

NameConflate-Subtle

NameConflate: Distinct vs. Subtle

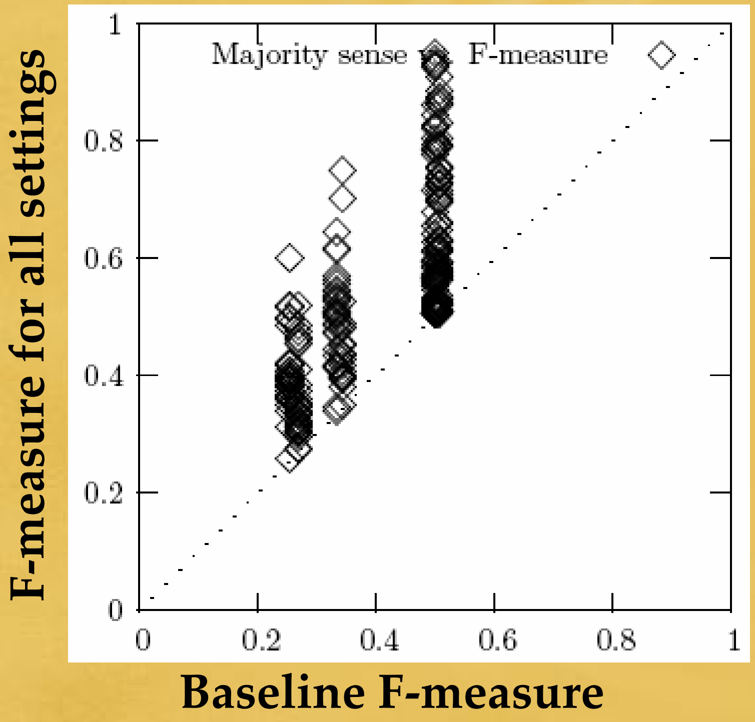


NameConflate-Distinct

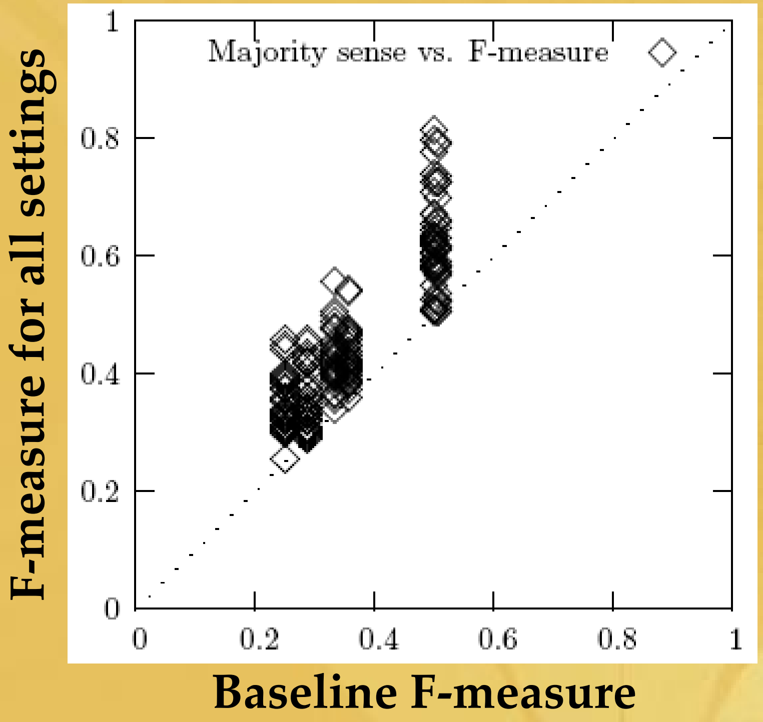


NameConflate-Subtle

Email: Distinct vs. Subtle



Email-Distinct

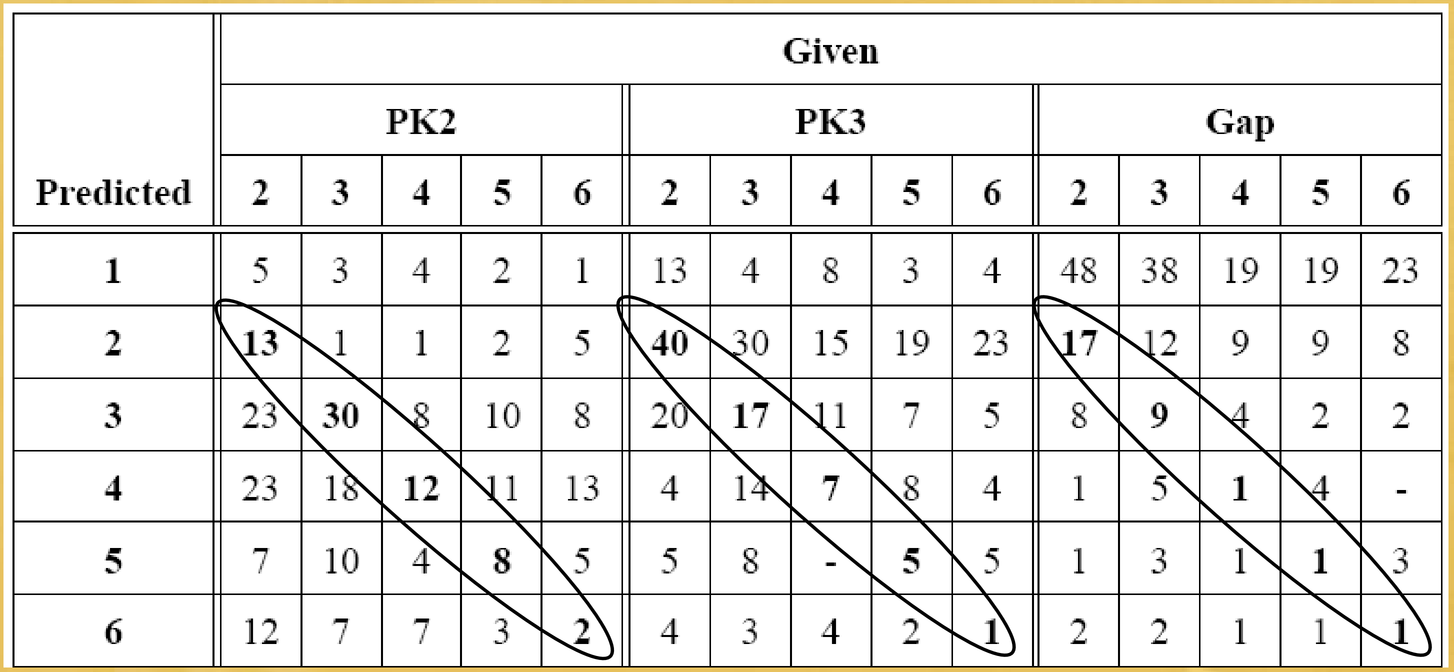


Email-Subtle

Cluster Stopping Results

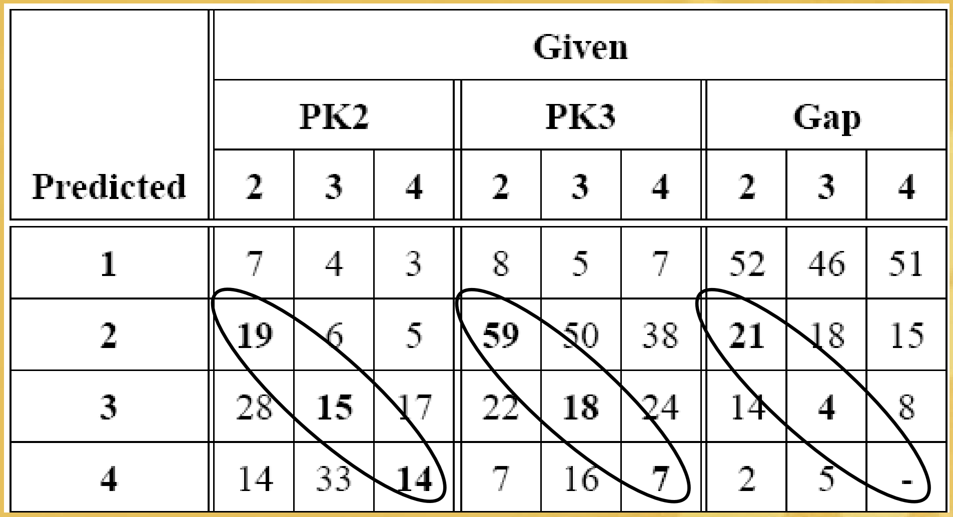
NameConflate: *k* predictions

NameConflate-Distinct



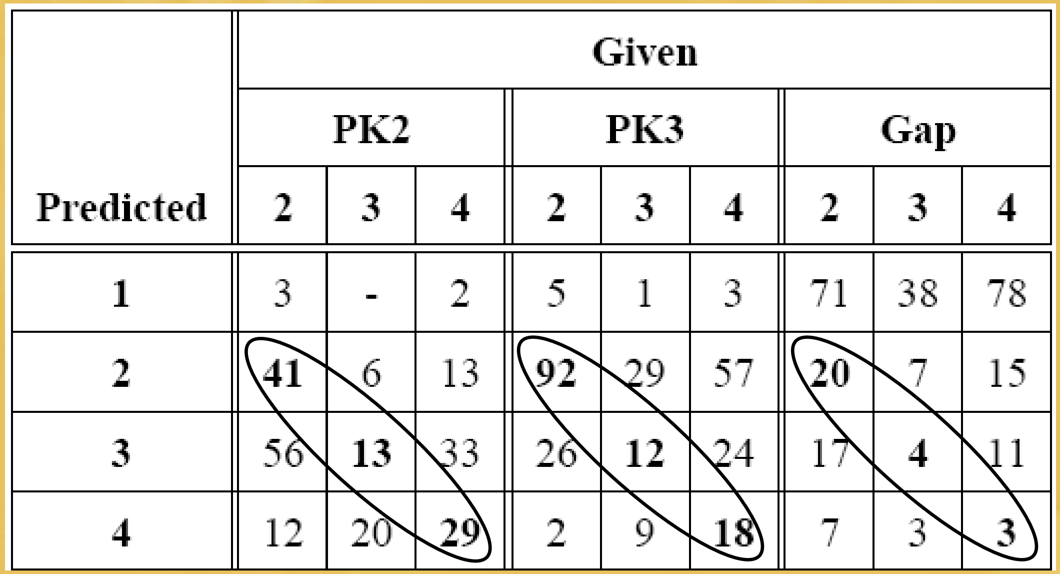
Predicted	Given														
	PK2					PK3					Gap				
	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
1	5	3	4	2	1	13	4	8	3	4	48	38	19	19	23
2	**13**	1	1	2	5	**40**	30	15	19	23	**17**	12	9	9	8
3	23	**30**	8	10	8	20	**17**	11	7	5	8	**9**	4	2	2
4	23	18	**12**	11	13	4	14	**7**	8	4	1	5	**1**	4	-
5	7	10	4	**8**	5	5	8	-	**5**	5	1	3	1	**1**	3
6	12	7	7	3	**2**	4	3	4	2	**1**	2	2	1	1	**1**

NameConflate-Subtle



Predicted	Given								
	PK2			PK3			Gap		
	2	3	4	2	3	4	2	3	4
1	7	4	3	8	5	7	52	46	51
2	**19**	6	5	**59**	50	38	**21**	18	15
3	28	**15**	17	22	**18**	24	14	**4**	8
4	14	33	**14**	7	16	**7**	2	5	**-**

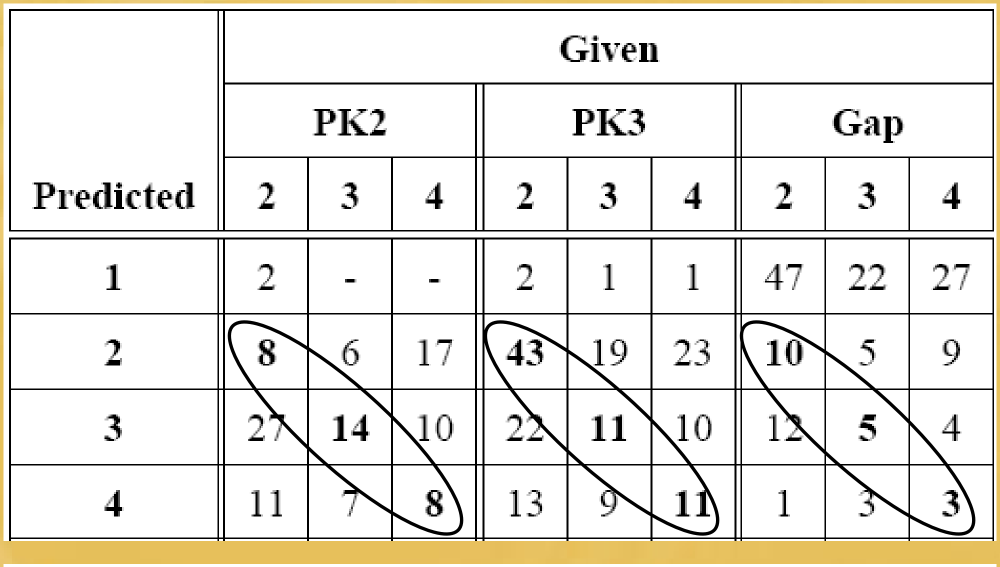
Web: *k* predictions



	Given								
	PK2			PK3			Gap		
Predicted	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**
1	3	-	2	5	1	3	71	38	78
2	**41**	6	13	**92**	29	57	**20**	7	15
3	56	**13**	33	26	**12**	24	17	**4**	11
4	12	20	**29**	2	9	**18**	7	3	**3**

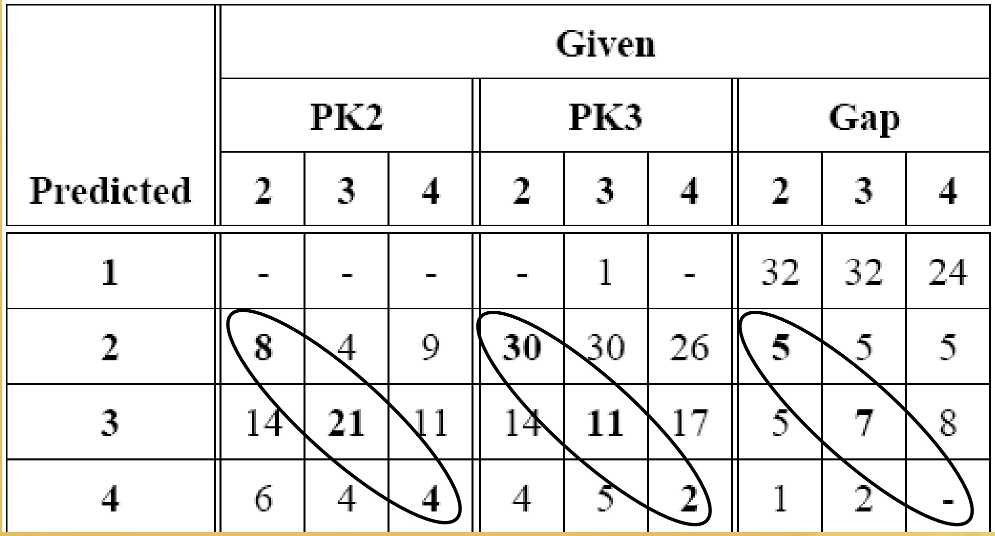
Email: *k* predictions

Email-distinct



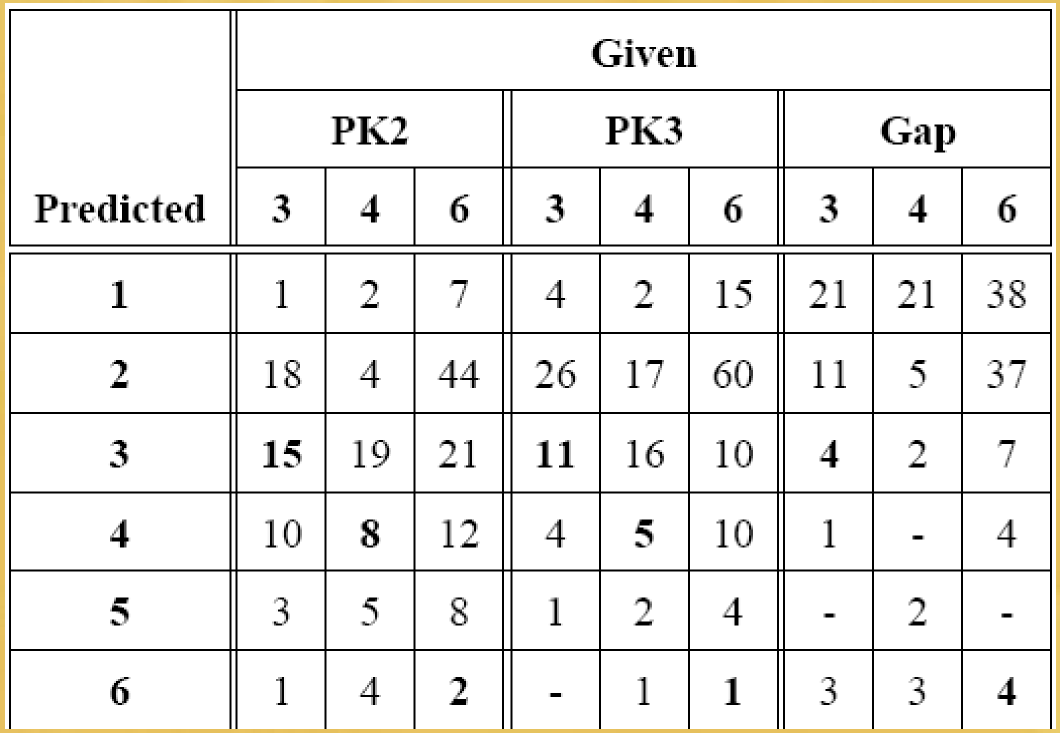
	Given								
	PK2			PK3			Gap		
Predicted	2	3	4	2	3	4	2	3	4
1	2	-	-	2	1	1	47	22	27
2	**8**	6	17	**43**	19	23	**10**	5	9
3	27	**14**	10	22	**11**	10	12	**5**	4
4	11	7	**8**	13	9	**11**	1	3	**3**

Email-subtle



	Given								
	PK2			PK3			Gap		
Predicted	2	3	4	2	3	4	2	3	4
1	-	-	-	-	1	-	32	32	24
2	**8**	4	9	**30**	30	26	**5**	5	5
3	14	**21**	11	14	**11**	17	5	**7**	8
4	6	4	**4**	4	5	**2**	1	2	-

WSD: *k* predictions



Predicted	Given								
	PK2			PK3			Gap		
	3	4	6	3	4	6	3	4	6
1	1	2	7	4	2	15	21	21	38
2	18	4	44	26	17	60	11	5	37
3	**15**	19	21	**11**	16	10	**4**	2	7
4	10	**8**	12	4	**5**	10	1	-	4
5	3	5	8	1	2	4	-	2	-
6	1	4	**2**	-	1	**1**	3	3	**4**

Conclusions

- Generalized the approach of by Purandare and Pedersen [2004] for WSD
  - Name Discrimination (headed clustering)
  - Email Clustering (headless clustering)
  - Thus in general for "Context Discrimination"
- Proposed and experimented with 3 cluster stopping measures.
- PK3 exhibits maximum agreement with the given number of clusters.

Conclusions (cont.)

- Order1 and Order2 provide a complimenting pair of context representations.
- Applying SVD generally does not help our methods.
- Performance of the clustering algorithm of repeated bisections is generally comparable with agglomerative except for the subtle type of datasets.
- We also find that our methods are better equipped to deal with "distinct" type of datasets than with "subtle" type of datasets.

Related Work

- Mann and Yarowsky, CoNLL 2003.

  Perform name disambiguation based on biographical data from WWW.

- Salvador and Chan, IEEE-ICTAI 2004.

  Introduce L-method for cluster-stopping which is based on fitting lines through evaluation graphs.

- Hamerly and Elkan, NIPS 2003.

  Introduce G-means method for cluster-stopping which is based on fitting a Gaussian distribution to each cluster.

Future Work

- Comparison with Latent Semantic Analysis (LSA)
- Improving the quality of automatically generated cluster labels
- Develop ensembles of cluster stopping methods
- Explore the effect of automatically generated stoplists

Links

- **SenseClusters**

  Project: http://senseclusters.sourceforge.net/

  Web-interface: http://marimba.d.umn.edu/cgi-bin/SC-cgi/index.cgi

- NameConflate and other Data generation utilities
  - http://www.d.umn.edu/~tpederse/tools.html

- Data and Publications
  - http://www.d.umn.edu/~tpederse/data.html
  - http://www.d.umn.edu/~tpederse/senseclusters-pubs.html