

Automatic Text Classification

Yutaka Sasaki

NaCTeM

School of Computer Science

Classification of Clinical Records

- Medical NLP Challenge (Computational Medicine Centre)
 - Classify anonymized real clinical records into International Clinical Codes (ICD-9-CM)
 - 44 research institutes participated
- Sample
 - Record:

Clinical History
This is a patient with meningocele and neurogenic bladder.

Impression
Normal renal ultrasound in a patient with neurogenic bladder.
 - Correct codes (possibly multiple codes):
 - 596.54 (Neurogenic bladder NOS)
 - 741.90 (Without mention of hydrocephalus)

Text Classification Demo (Sasaki, UoM) - Mozilla Firefox

ファイル(F) 編集(E) 表示(V) 移動(G) ブックマーク(B) ツール(T) ヘルプ(H)

http://text0.mib.man.ac.uk/~sasaki/demo/

Customize Links Free Hotmail Windows Marketplace Windows Media Windows Webcalldirect

Clinical Document Classification by PHENETICA

MANCHESTER 1824 (c)2007 Dr. Yutaka Sasaki NaCTeM

Clinical Record

Select a document here → 99614096
(previously unseen by classifier)

ICD-9: [596.54:741.90]

Clinical History
This is a patient with meningocele and neurogenic bladder.

Impression
Normal renal ultrasound in a patient with neurogenic bladder.

JA ● EN ●

Classify

(ICD9 Code Search: search)

(ICD9: International Classification of Diseases, Ninth Revision)

Slides

Please click [here](#) to see slides on document classification.

Results

– System's Choice

- Neurogenic bladder NOS (596.54)
- Without mention of hydrocephalus (741.90)

– Correct Answer

- Neurogenic bladder NOS (596.54)
- Without mention of hydrocephalus (741.90)

System's Top 5 Candidates

Rank	ICD9 Code(s)	Probability	Disease
[1]	596.54:741.90	74.9%	Neurogenic bladder NOS Without mention of hydrocephalus
[2]	596.54	14.3%	Neurogenic bladder NOS
[3]	788.30	2.9%	Urinary incontinence, unspecified
[4]	599.0	1.4%	Urinary tract infection, site not specified
[5]	599.7	1.2%	Hematuria

Classification Details

Feature	Weight
[I]neurogenic bladder	+6.956
[I]neurogenic	+6.956
[C]neurogenic	-4.369
[C]neurogenic bladder	-4.369

Document

Predicted codes (multi-topics)

Correct codes

Top 5 Candidates

Significance of each feature

Evaluation results

Table 2. Scores of Top 10 Systems in Terms of the Cost Sensitive Measure and 3 Annotators in the Medical NLP Challenge 2007

Team Short Name	Cost Sensitive	Micro-average F1	Macro-average F1
Szeged	0.9180	0.8908	0.7691
University of Turku	0.9126	0.8769	0.7034
University at Albany	0.9091	0.8855	0.7291
PENN	0.9088	0.8760	0.7210
Annotator A	0.9056	0.8264	0.6124
<i>MANCS</i>	<i>0.9049</i>	<i>0.8594</i>	<i>0.6676</i>
otters	0.9010	0.8509	0.6816
LMCO-IS & S	0.9009	0.8719	0.7760
SULTRG	0.8998	0.8676	0.7322
Annotator B	0.8997	0.8963	0.8973
GMJLJL	0.8975	0.8711	0.7334
ohsu_dmice	0.8938	0.8457	0.6542
Annotator C	0.8621	0.8454	0.8829

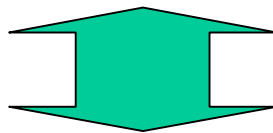
Introduction

Introduction

- Text Classification is the task:
 - to classify documents into predefined classes
- Text Classification is also called
 - Text Categorization
 - Document Classification
 - Document Categorization
- Two approaches
 - manual classification and automatic classification

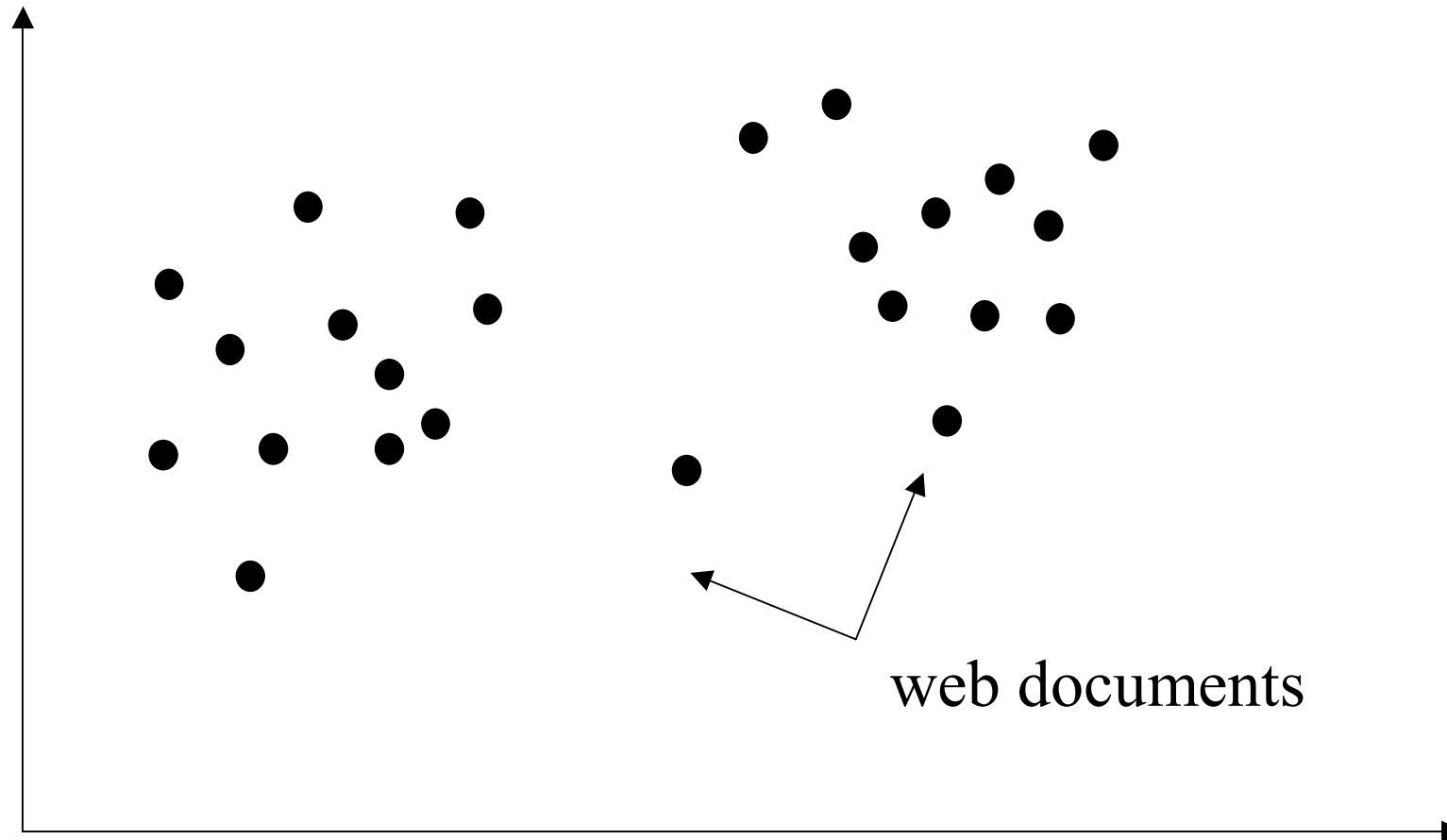
Relevant technologies

- Text Clustering
 - Create clusters of documents **without any external information**
- Information Retrieval (IR)
 - Retrieve a set of documents relevant to a **query**
- Information Filtering
 - Filter out irrelevant documents through **interactions**
- Information Extraction (IE)
 - Extract fragments of **information**, e.g., person names, dates, and places, in documents

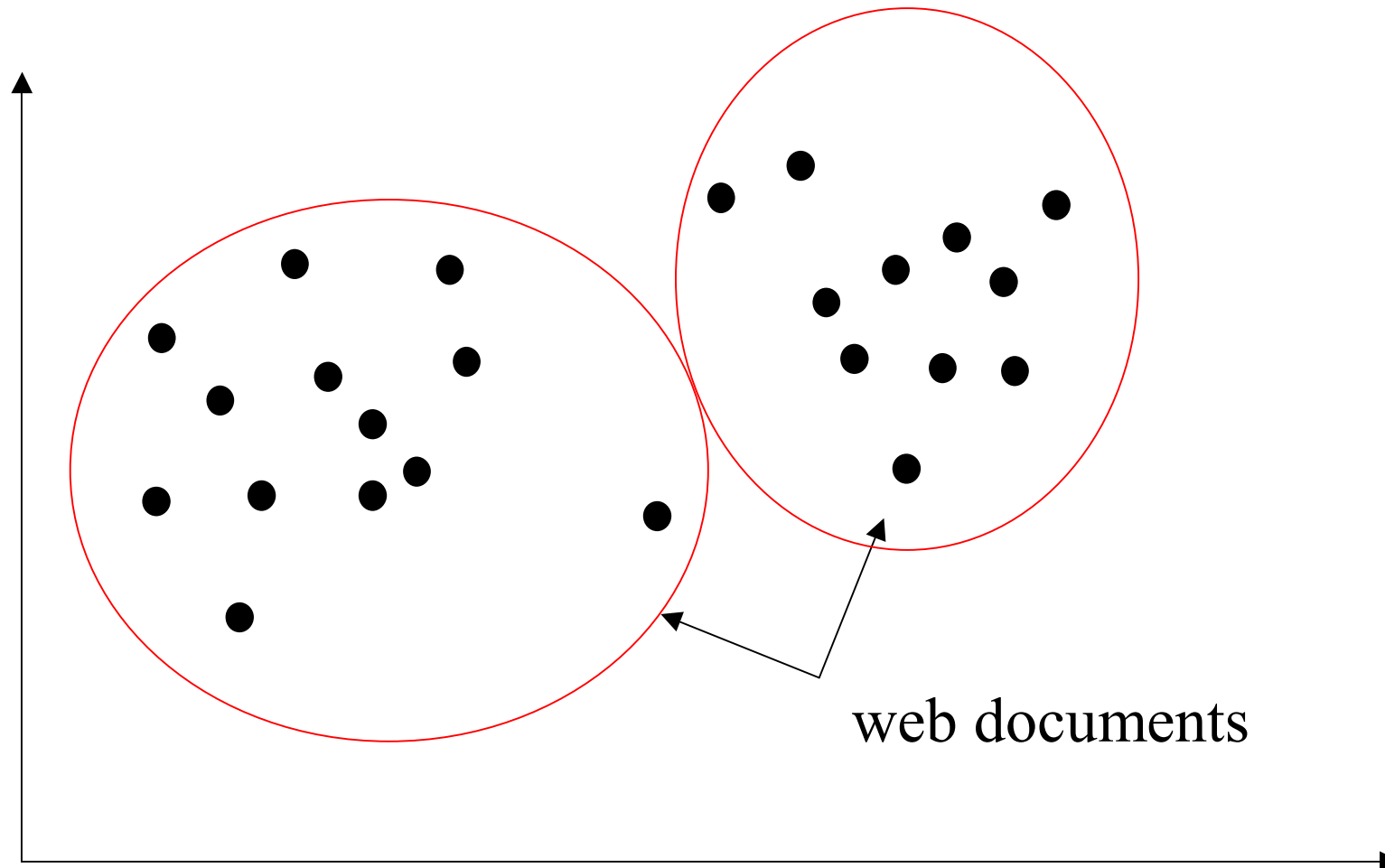


- Text Classification
 - No **query, interactions, external information**
 - Decide **topics** of documents

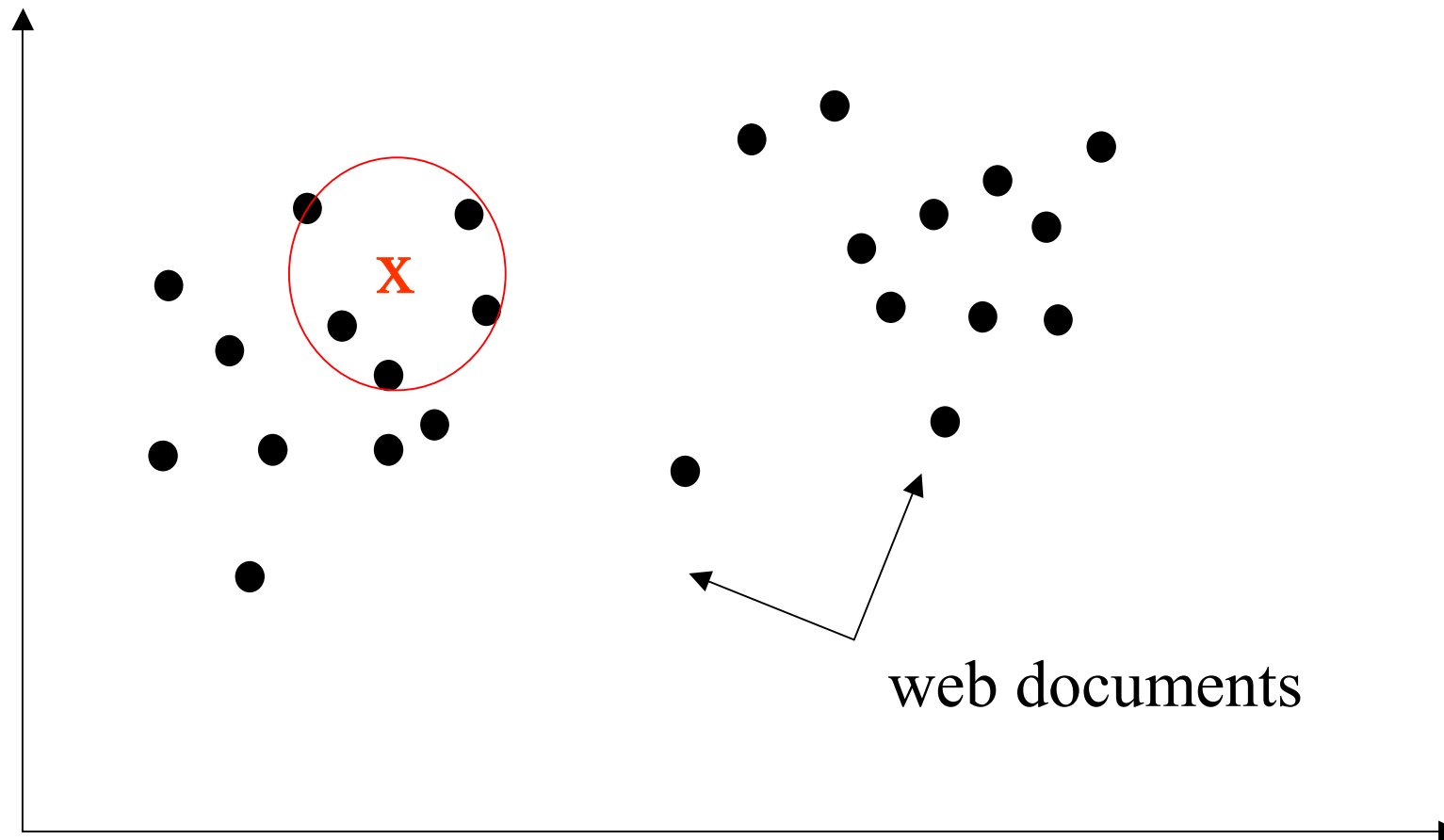
Examples of relevant technologies



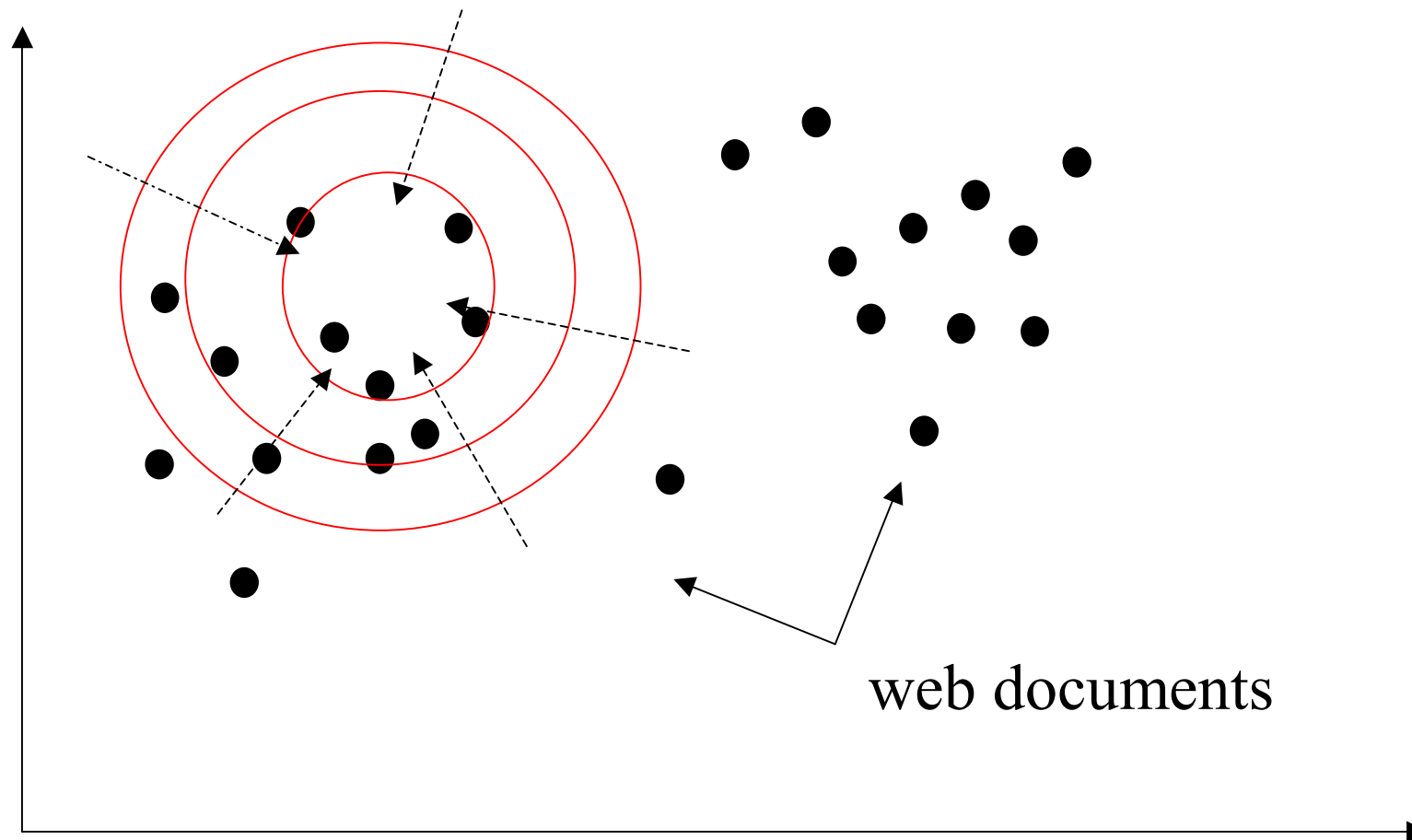
Example of clustering



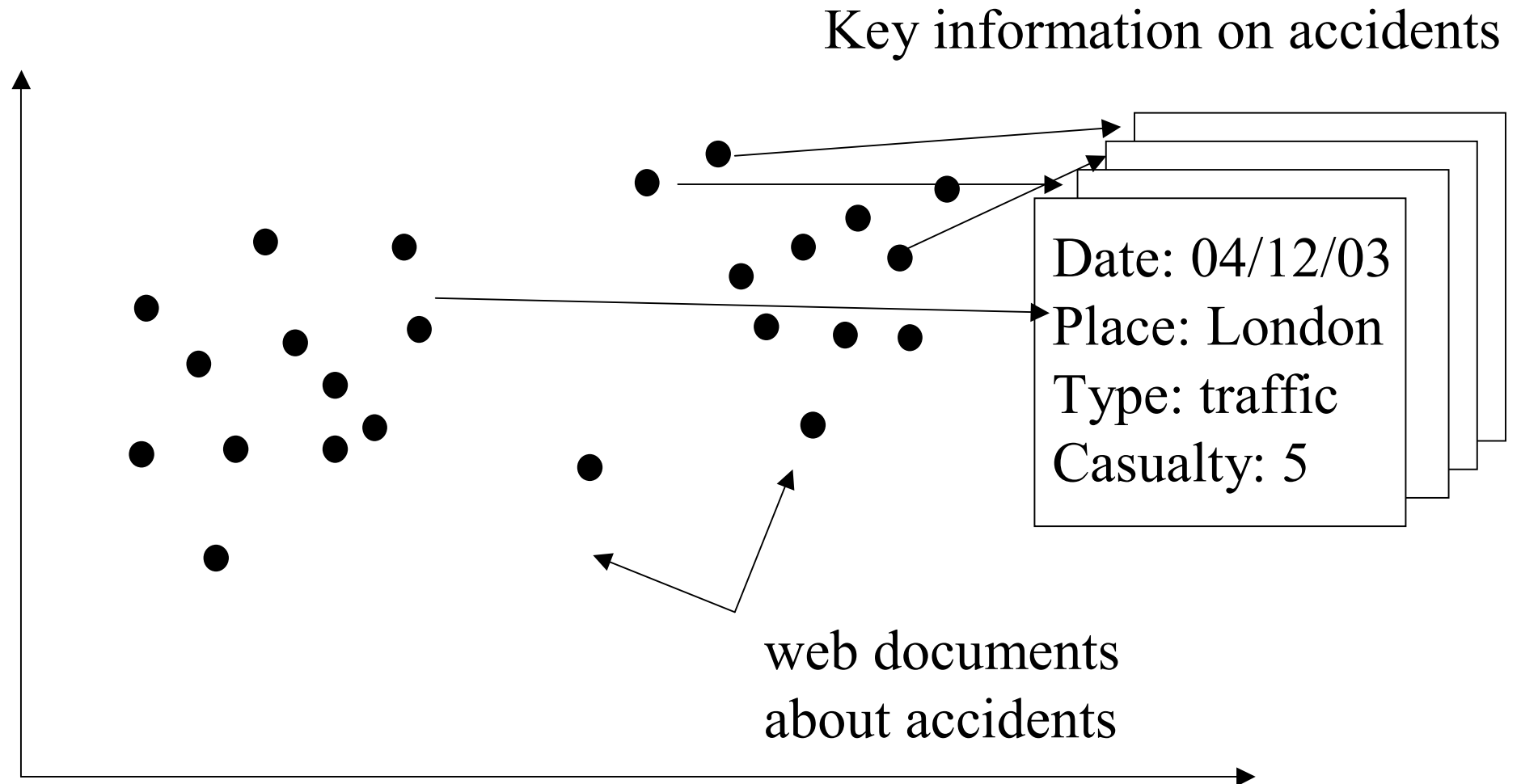
Examples of information retrieval



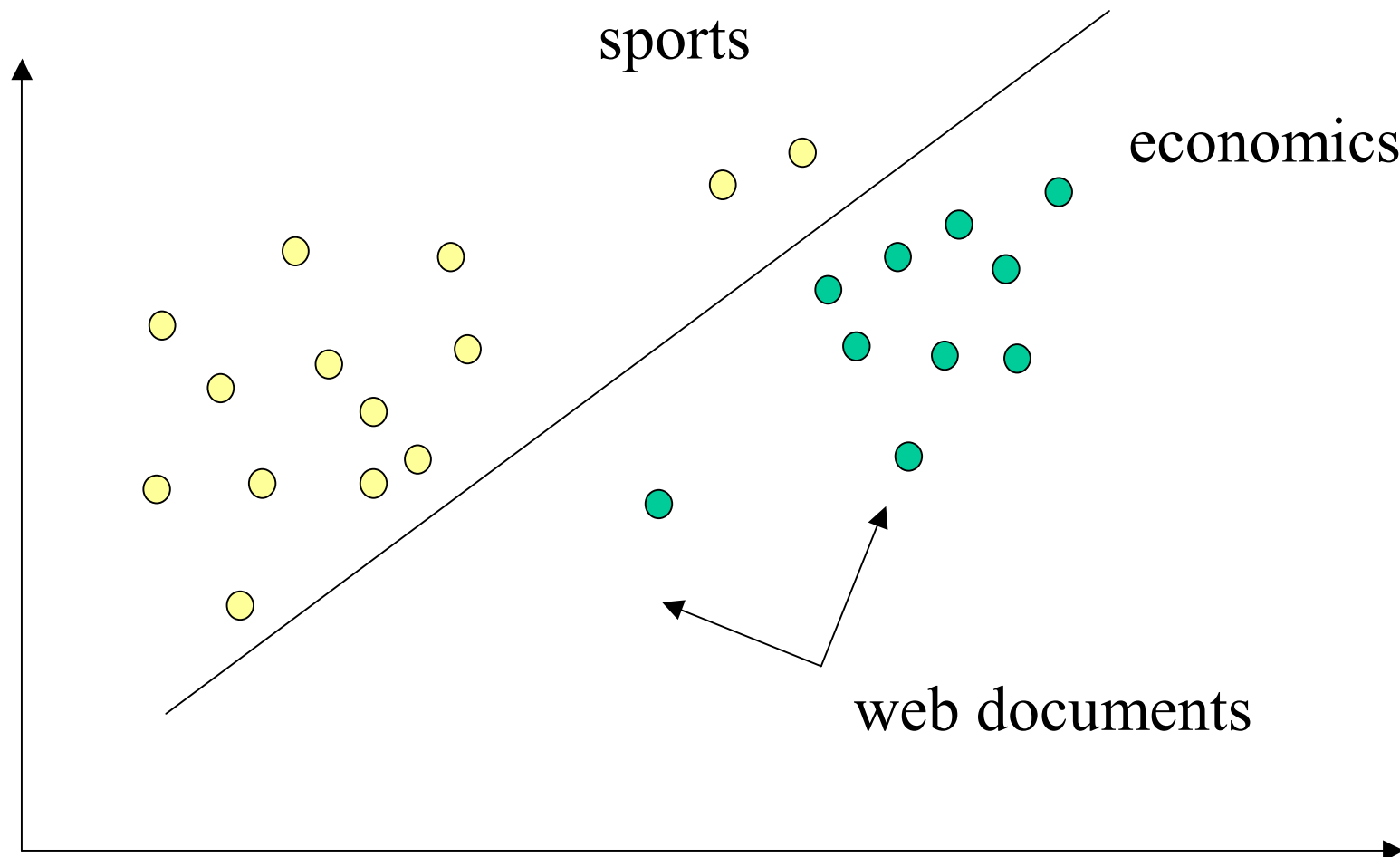
Examples of information filtering



Examples of information extraction



Examples of text classification



Text Classification Applications

- E-mail spam filtering
- Categorize newspaper articles and newswires into topics
- Organize Web pages into hierarchical categories
- Sort journals and abstracts by subject categories (e.g., MEDLINE, etc.)
- Assigning international clinical codes to patient clinical records

Simple text classification example

- You want to classify documents into 4 classes: economics, sports, science, life.
- There are two approaches that you can take:
 - rule-based approach
 - write a set of rules that classify documents
 - machine learning-based approach
 - using a set of sample documents that are classified into the classes (training data), automatically create classifiers based on the training data

Comparison of Two Approaches (1)

Rule-based classification

Pros:

- very accurate when rules are written by experts
- classification criteria can be easily controlled when the number of rules are small.

Cons:

- sometimes, rules conflicts each other
 - maintenance of rules becomes more difficult as the number of rules increases
- the rules have to be reconstructed when a target domain changes
- low coverage because of a wide variety of expressions

Comparison of Two Approaches (2)

Machine Learning-based approach

Pros:

- domain independent
- high predictive performance

Cons:

- not accountable for classification results
- training data required

Formal Definition

- Given:
 - A set of documents $D = \{d_1, d_2, \dots, d_m\}$
 - A fixed set of topics $T = \{t_1, t_2, \dots, t_n\}$
- Determine:
 - The topic of d : $t(d) \in T$, where $t(x)$ is a classification function whose domain is D and whose range is T .

Rule-based approach

Example: Classify documents into sports

“ball” must be a word that is frequently used in sports

\Rightarrow Rule 1: “ball” $\in d \rightarrow t(d) = \text{sports}$

But there are other meanings of “ball”

Def.2-1 : a large formal gathering for social dancing
(WEBSTER)

\Rightarrow Rule 2: “ball” $\in d$ & “dance” $\notin d \rightarrow t(d) = \text{sports}$

Def.2-2 : a very pleasant experience : a good time (WEBSTER)

\Rightarrow Rule 3: “ball” $\in d$ & “dance” $\notin d$ & “game” $\in d$ &
“play” $\in d \rightarrow t(d) = \text{sports}$

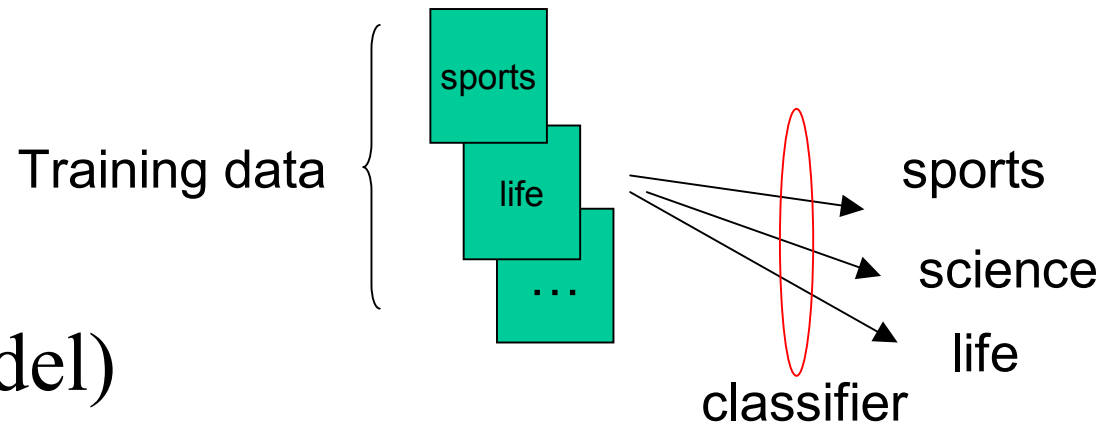
Natural language has a rich variety of expressions:

e.g., “Many people have a ball when they play a bingo game.”

Machine Learning Approach

1. Prepare a set of training data

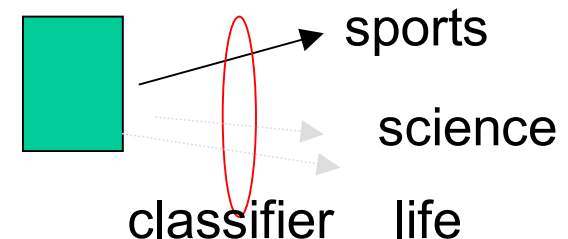
- Attach topic information to the documents in a target domain.

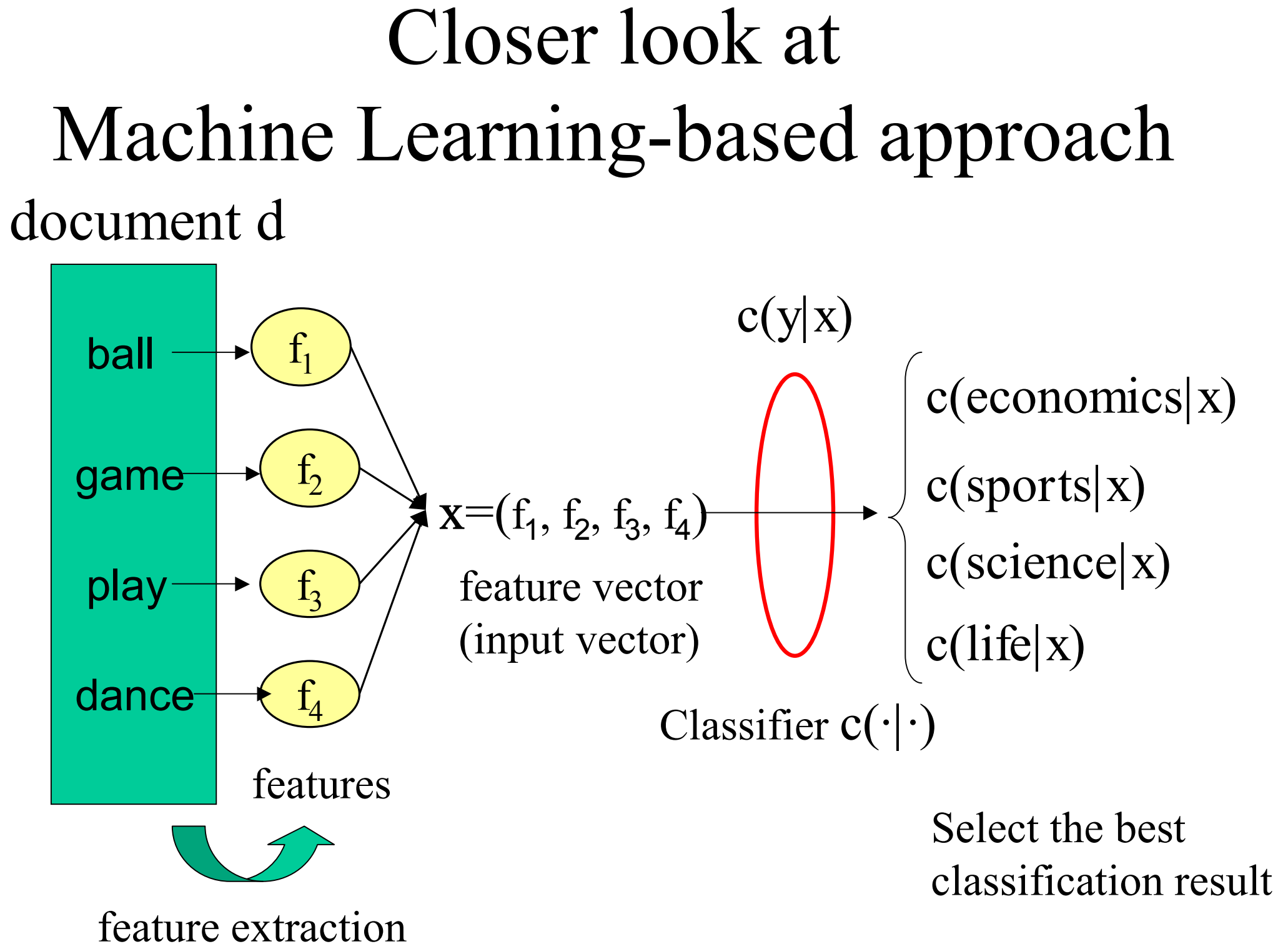


2. Create a classifier (model)

- Apply a Machine Learning tool to the data
 - Support Vector Machine (SVM), Maximum Entropy Models (MEM)

3. Classify new documents by the classifier





Rule-based vs. Machine Learning-based

[Creecy et al., 1992]

- Data: US Census Bureau Decennial Census 1990
 - 22 million natural language responses
 - 232 industry categories and 504 occupation categories
 - It costs about \$15 million if fully done by hand
- Define classification rules manually:
 - Expert System AIOCS
 - Development time: 192 person-months (2 people, 8 years)
 - Accuracy = 57%(industry), 37%(occupation)
- Learn classification function
 - Machine Learning-based System PACE
 - Development time: 4 person-months
 - Accuracy = 63%(industry), 57%(occupation)

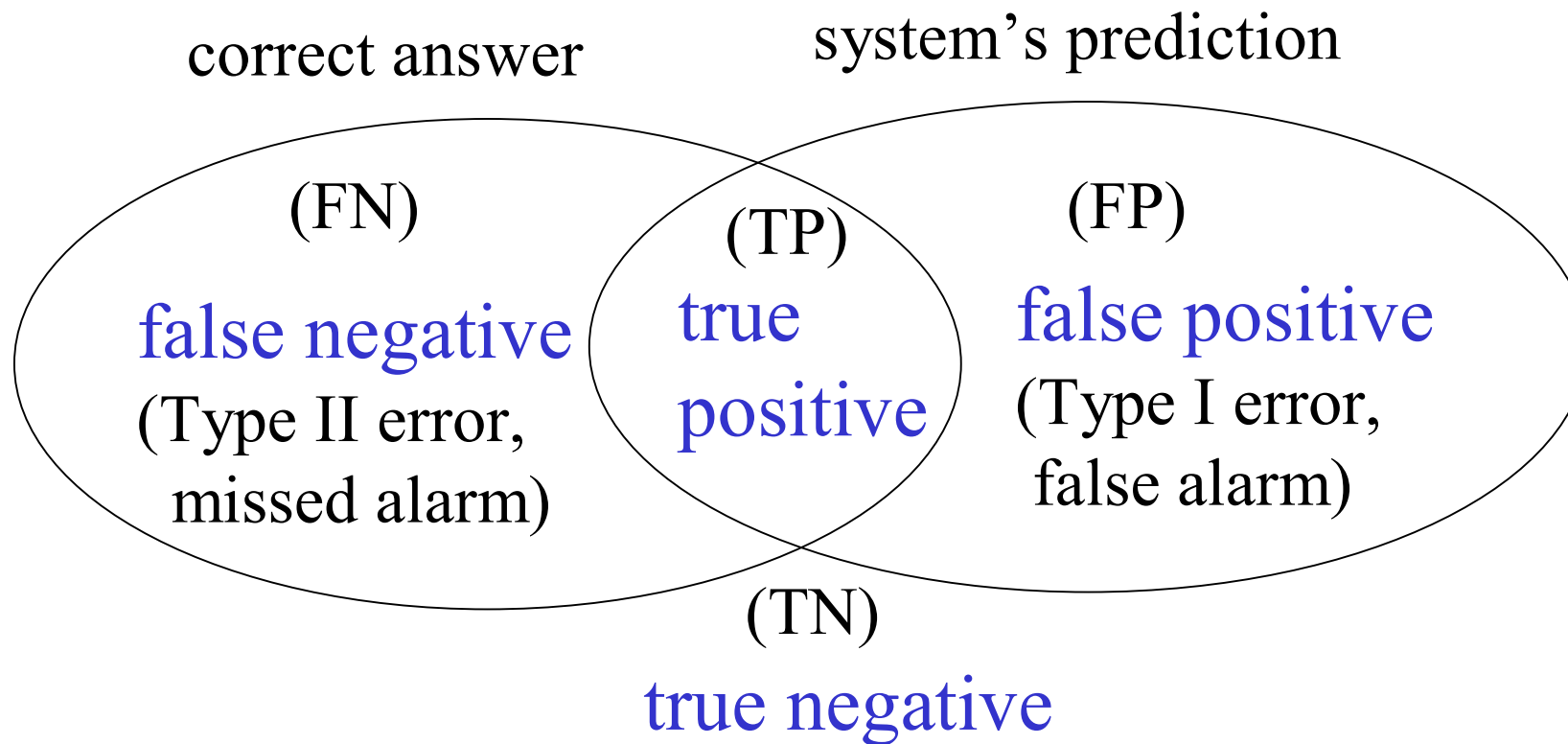
Evaluation

Common Evaluation Metrics

- Accuracy
- Precision
- Recall
- F-measure
 - harmonic mean of recall and precision
 - micro-average F1
 - global calculation of F1 regardless of topics
 - macro-average F1:
 - average on F1 scores of all the topics

Accuracy

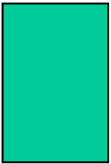
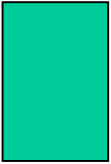
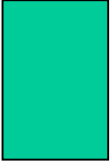
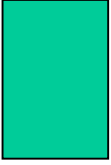
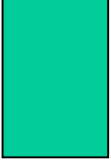
- The rate of correctly predicted topics



$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Accuracy

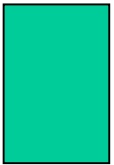
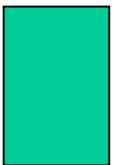
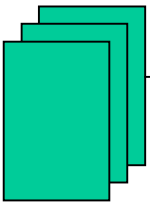
- Example: classify docs into spam or not spam

system's prediction			correct answer	TP	FP	FN	TN
d1		→ Y	N		1		
d2		→ Y	Y	1			
d3		→ N	Y			1	
d4		→ N	N				1
d5		→ Y	N		1		

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{1 + 1}{1 + 2 + 1 + 1} = 0.4$$

Issue in Accuracy

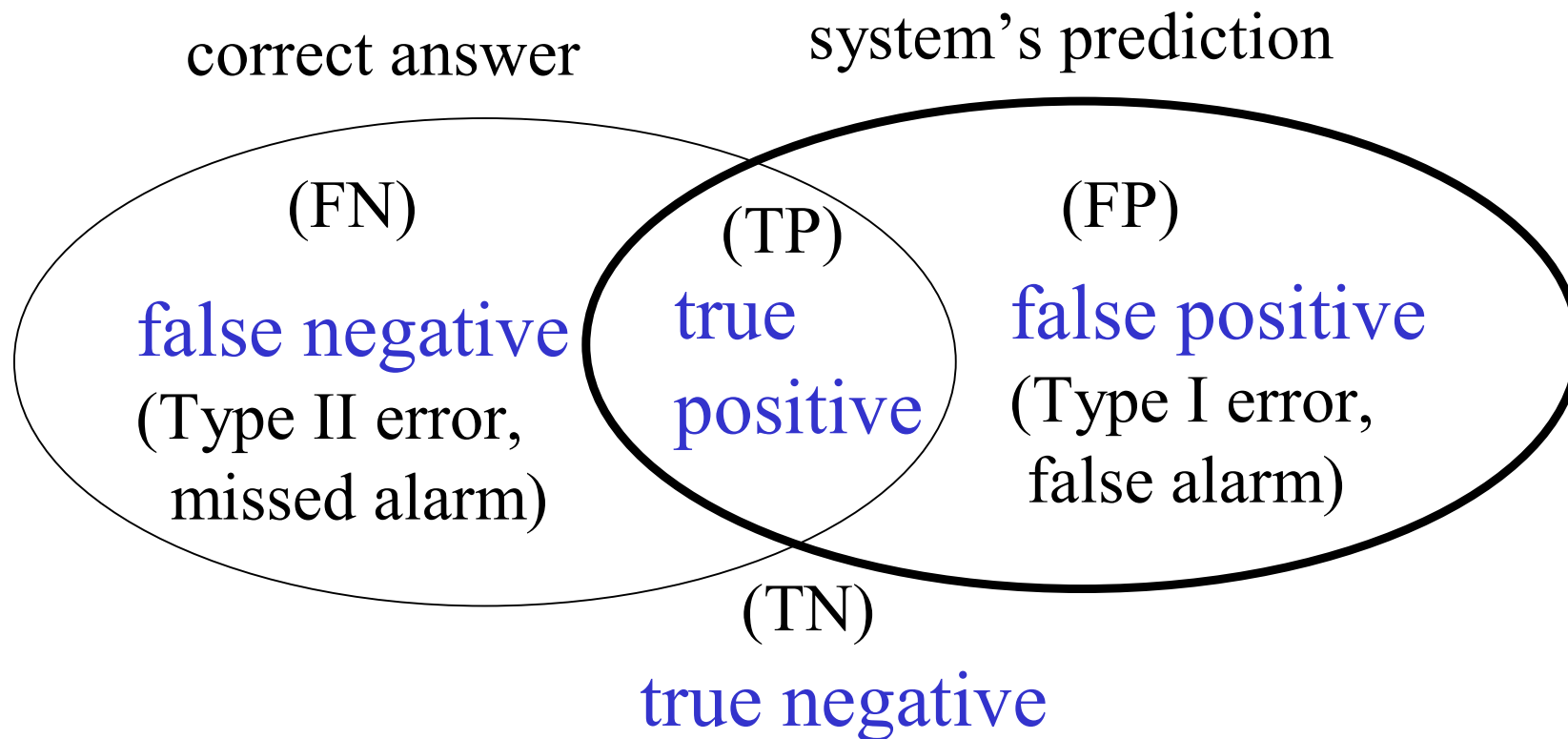
- When a certain topic (e.g., not-spam) is a majority, the accuracy easily reaches a high percentage.

system's prediction		correct answer	TP	FP	FN	TN
d1	 →	N			1	
...	
d10	 →	N			1	
d11-	 →	N				
d1000		...				
		N				
			}			990

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \frac{990}{1000} = 0.99$$

Precision (PPV)






- The rate of correctly predicted topics



$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision

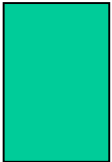
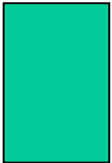
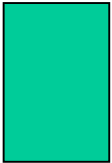
- Example: classify docs into spam or not spam

system's prediction			correct answer	TP	FP	FN	TN
d1		→	Y		1		
d2		→	Y	1			
d3		→	N			1	
d4		→	N				1
d5		→	Y		1		

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} = \frac{1}{1+2} = 0.333$$

Issue in Precision

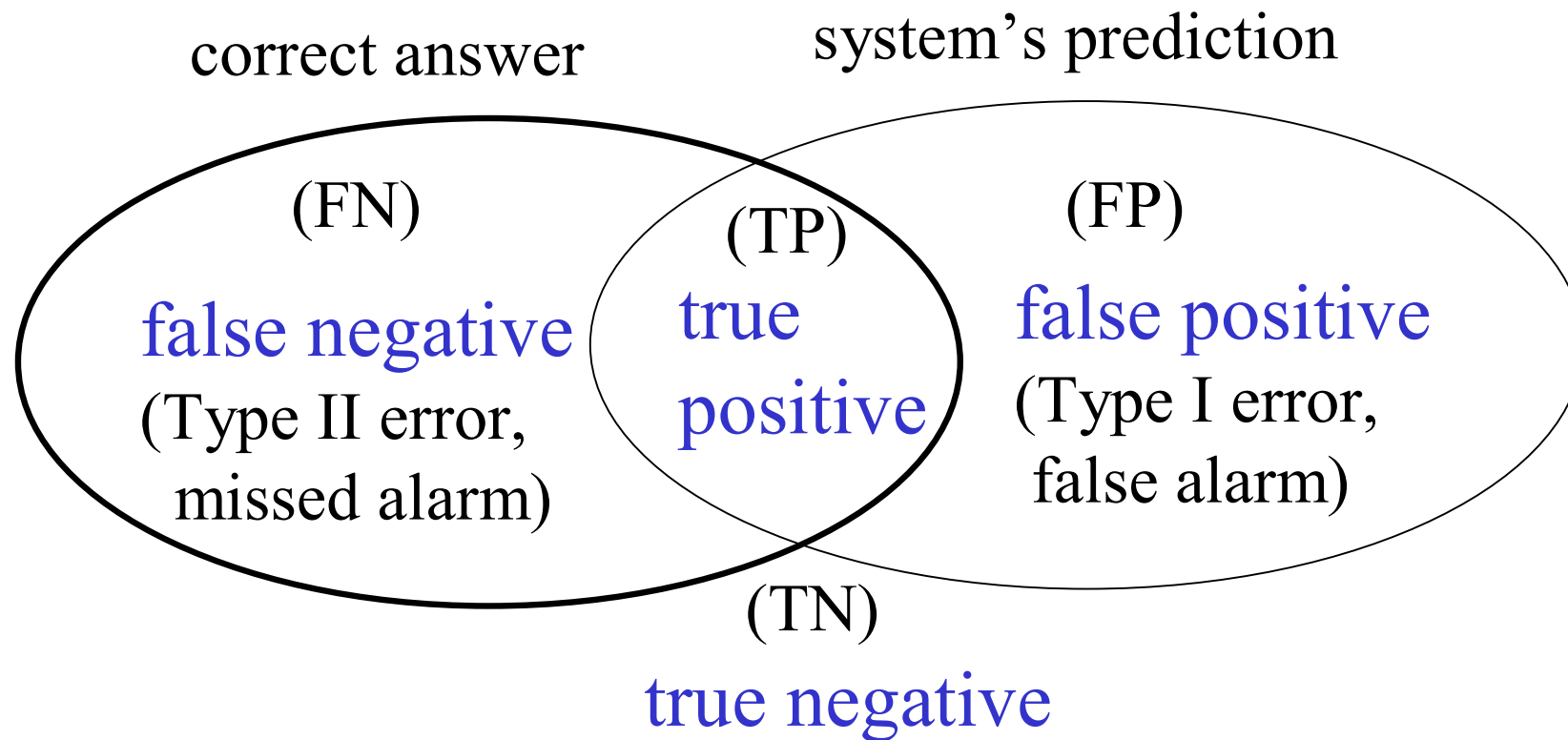
- When a system outputs only confident topics, the precision easily reaches a high percentage.

system's prediction			correct answer	TP	FP	FN	TN
d1		→	N			1	...
...(Y or N)...			...	
d999		→	N				1
d1000		→	Y	1			

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{1}{1} = 1$$

Recall (sensitivity)

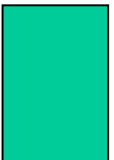
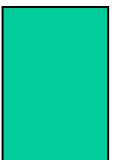
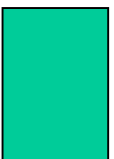
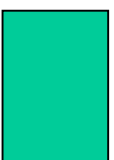
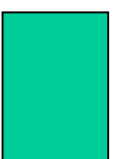
- The rate of correctly predicted topics



$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Recall

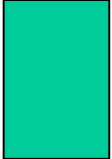
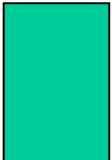
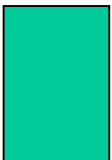
- Example: classify docs into spam or not spam

system's prediction			correct answer	TP FP FN TN
d1		→ Y	N	1
d2		→ Y	Y	1
d3		→ N	Y	1
d4		→ N	N	1
d5		→ Y	N	1

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{1}{1 + 1} = 0.5$$

Issue in Recall

- When a system outputs loosely, the recall easily reaches a high percentage.

system's prediction			correct answer	TP	FP	FN	TN
d1		→	Y	1			
...(Y or N)...		
d999		→	N		1		
d1000		→	Y	1			

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{n}{n} = 1$$

F-measure

- Harmonic mean of recall and precision

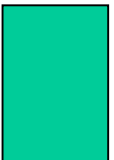
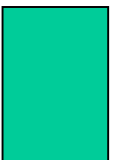
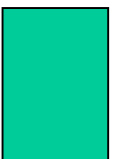
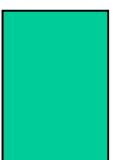
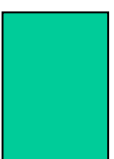
$$2 \cdot \text{Precision} \cdot \text{Recall}$$

$$\text{Precision} + \text{Recall}$$

- Since there is a trade-off between recall and precision, F-measure is widely used to evaluate text classification system.
- Micro-average F1: Global calculation of F1 regardless of topics
- Macro-average F1: Average on F1 scores of all topics

F-measure

- Example: classify docs into spam or not spam

system's prediction			correct answer	TP	FP	FN	TN
d1		→	Y		1		
d2		→	Y	1			
d3		→	N			1	
d4		→	N				1
d5		→	Y		1		

$$F = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} = \frac{2 \cdot 1/3 \cdot 1/2}{1/3 + 1/2} = 0.4$$

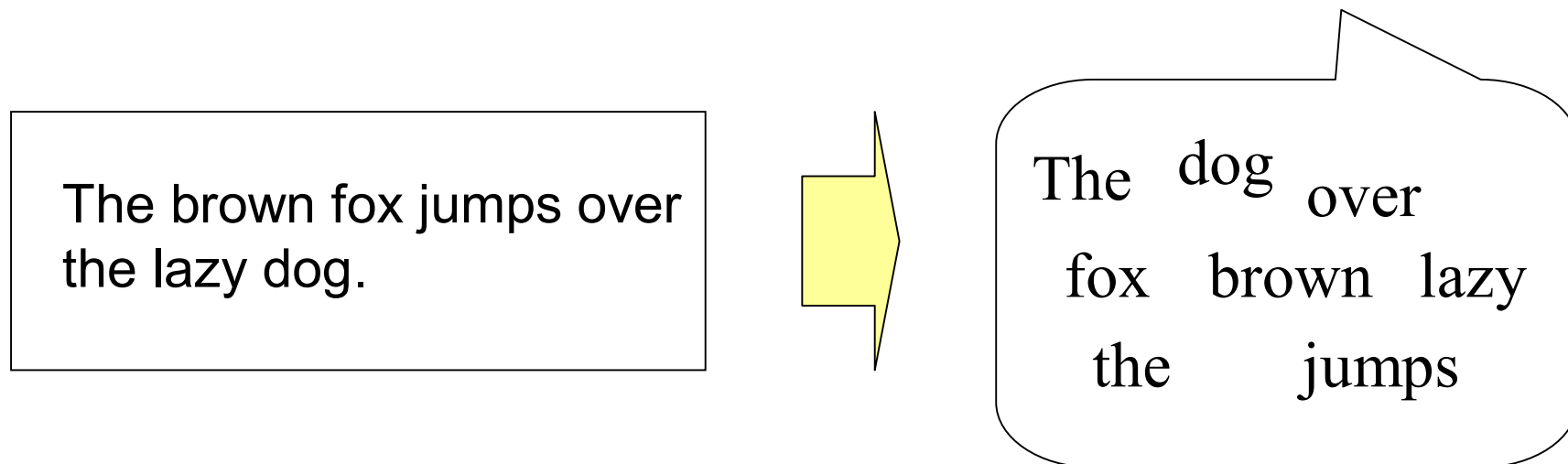
Summary: Evaluation Metrics

- Accuracy
- Precision
$$\frac{TP \text{ (\# system's correct predictions)}}{TP+FP \text{ (\# system's outputs)}}$$
- Recall
$$\frac{TP \text{ (\# system's correct predictions)}}{TP+FN \text{ (\# correct answers)}}$$
- F-measure
$$\frac{2 * Recall * Precision}{Recall + Precision}$$
- Micro F1: Global average of F1 regardless of topics
- Macro F1: Average on F1 scores of all topics
- Cost-Sensitive Accuracy Measure (*)
- Multi-Topic Accuracy (*)

Feature Extraction: from Text to Data

Basic Approach (1)

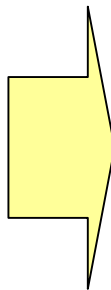
- Bag-of-Word approach
 - a document is regarded as a set of words regardless of the word order and grammar.



Basic Approach (2)

- Bi-grams, tri-grams, n-grams
 - Extract all of two, three, or n words in a row in the text

The brown fox jumps over
the lazy dog.



Bi-grams:

the brown,
brown fox,
fox jumps,
jumps over,
the lazy,
lazy dog

Tri-grams:

the brown fox,
brown fox jumps,
fox jumps over,
jumps over the,
the lazy dog

Basic Approach (3)

- Normalization

Convert words into a normalized forms

- down-case, e.g, The \rightarrow the, NF-kappa B \rightarrow nf-kappa b
- lemmatization: to basic forms, e.g., jumps \rightarrow jump
- stemming: mechanically remove/change suffixes
 - e.g., y \rightarrow i, s \rightarrow , “the brown fox jump over the lazi dog.”
 - the Porter’s Stemmer is widely used.

- Stop-word removal

- ignore predefined common words, e.g., the, a, to, with, that ...
- the SMART Stop List is widely used

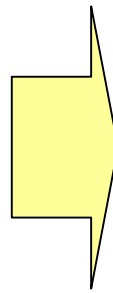
From Symbols to Numeric

- Term occurrence: occur (1) or not-occur (0)
- Term Frequency
 - tf_i = the number of times where word/n-gram w_i appears in a document.
- Inverse document frequency
 - the inverted rate of documents that contain word/n-gram w_i against a whole set of documents
$$idf_i = |D| / |\{d \mid w_i \in d \in D\}|.$$
- tf-idf
 - $tf-idf_i = tf_i \cdot idf_i$
 - frequent words that appear only in a small number of documents achieve high value.

Create Feature Vectors

1. enumerate all word/n-grams in a whole set of documents
2. remove duplications and sort the words/n-grams
3. convert each word into its value, e.g., tf, idf, or tf-idf.
4. create a vector whose i-th value is the value of i-th term

The brown fox jumps over
the lazy dog.



feature vector with tf weights:

a an ...brown,... dog ... fox jump lazi over the
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓
(0, 0,...,0, 1, ,0,...,0, 1,0,...,0, 1,0,...,0,1, 0,...,0,1,0,...,0,1, 2, 0, ..)

Generally, feature vectors are very sparse, i.e., most of the values are 0.

Multi-Topic Text Classification

Multi-topic Text Classification

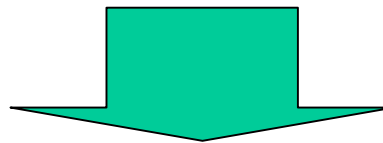
- One single document belongs to multiple topics
- An interesting and important research theme that is not nicely solved yet.

<TOPICS>ship</TOPICS>
The Panama Canal Commission, a U.S. government agency, said in its daily operations report that there was a backlog of 39 ships waiting to enter the canal early today.

<TOPICS>crude:ship</TOPICS>
The port of Philadelphia was closed when a Cypriot oil tanker, Seapride II, ran aground after hitting a 200-foot tower supporting power lines across the river, a Coast Guard spokesman said.

<TOPICS>crude</TOPICS>
Diamond Shamrock Corp said that effective today it had cut its contract prices for crude oil by 1.50 dlrs a barrel.

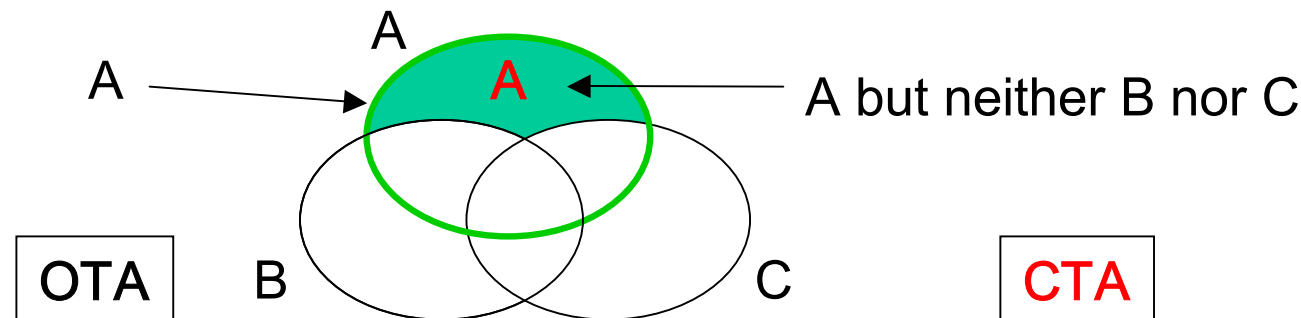
(Excerpt from Reuters-21578)



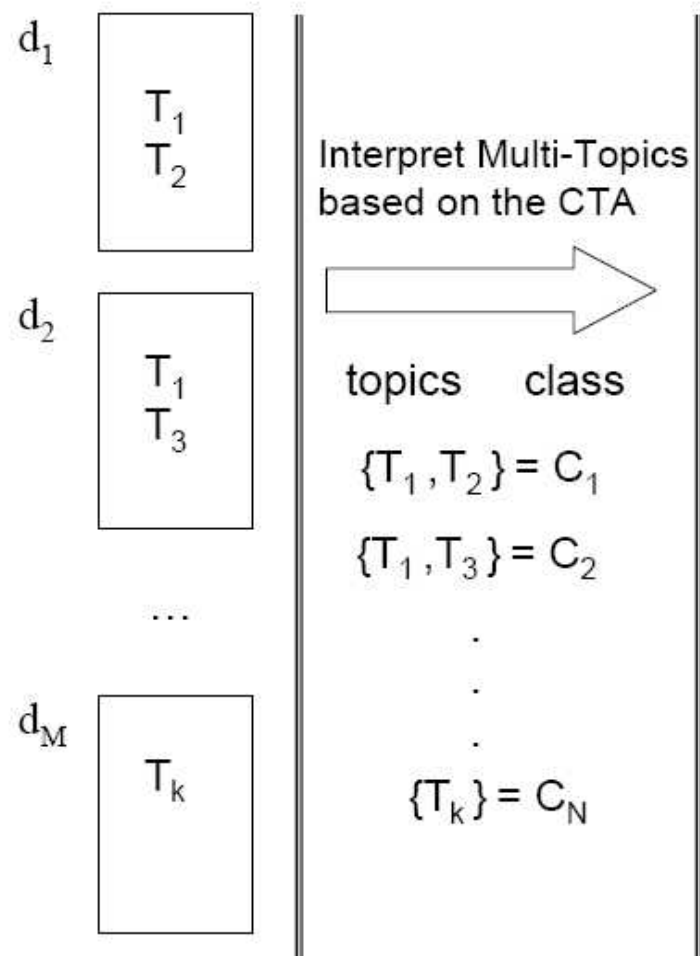
Topic A&B is not always a mixture of A and B

A View on Multi-topic Text Classification

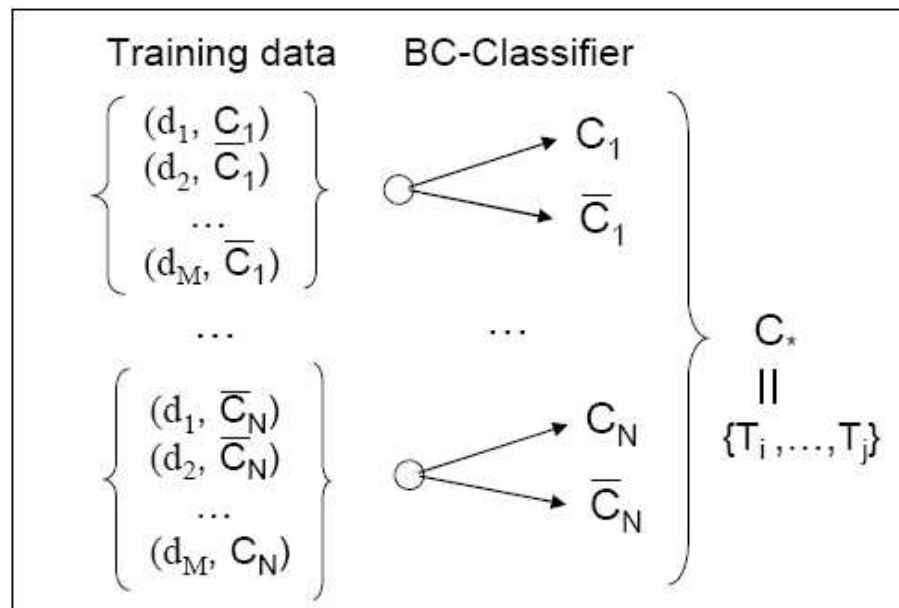
- **Open Topic Assumption (OTA)** (conventional view)
 - A document has multiple topics
 - The topics other than the given topics are neutral.
- **Closed Topic Assumption (CTA)**
 - A document has multiple topics
 - The other topics are considered to be explicitly excluded.
 - E.g., if there exist three topics A,B,C and a text d is given the topic A, then this assignment is regarded that d belongs to A but does **not** belong to B and C.



Multi-topic Documents



Binary-Class Classifiers



Multi-Class Classifier

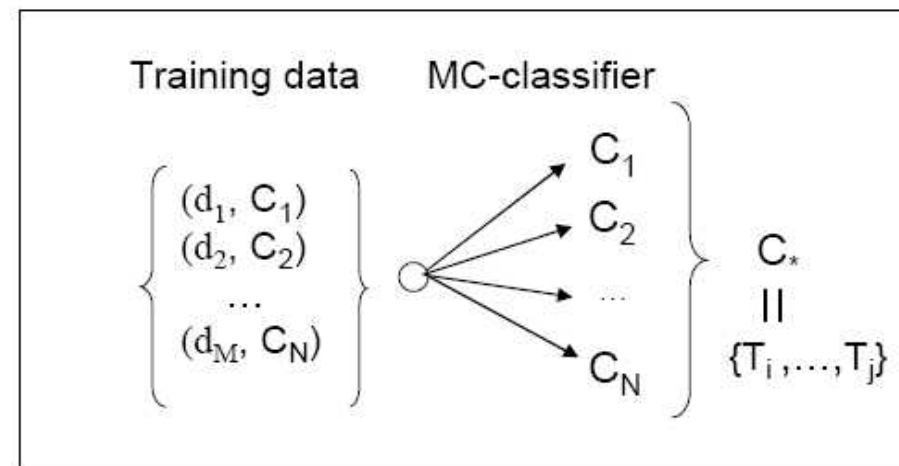


Figure 2. Implementing Multi-Topic Text Classification Problem

Case Studies

Experiments

- **Objective**
 - compare the performance of approaches based on Closed Topic Assumption and Open Topic Assumption.
- **Data 1 (Clinical records)**
 - Training: about 986 documents
 - Test: 984 documents
- **Data 2 (Reuters newswires)**
 - Training: 9,603 documents
 - Test: 3,299 documents
- **Machine Learning methods**
 - SVM: Support Vector Machines
 - MEM: Maximum Entropy Models
- **Approaches**
 - BC: Binary Class Classification
 - MC: Multi Class Classification

	SVM	MEM
BC	BCSVM (CTA/OTA)	BCMEM (CTA/OTA)
MC	MCSVM (CTA)	MCMEM (CTA)

Evaluation Metrics

- AC: multi-labelling accuracy
- Cost-Sensitive Accuracy Measure (for clinical data)
- Precision
$$\frac{\# \text{ system correct labeling}}{\# \text{ system output}}$$
- Recall
$$\frac{\# \text{ system correct labeling}}{\# \text{ correct labeling}}$$
- F1
$$\frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$
- Micro F1: Global calculation of F1 regardless of topics
- Macro F1: Average on F1 scores of all topics

Classification Experiments on Clinical Records

Table 2. Scores of Top 10 Systems in Terms of the Cost Sensitive Measure and 3 Annotators in the Medical NLP Challenge 2007

Team Short Name	Cost Sensitive	Micro-average F1	Macro-average F1
Szeged	0.9180	0.8908	0.7691
University of Turku	0.9126	0.8769	0.7034
University at Albany	0.9091	0.8855	0.7291
PENN	0.9088	0.8760	0.7210
Annotator A	0.9056	0.8264	0.6124
<i>MANCS</i>	<i>0.9049</i>	<i>0.8594</i>	<i>0.6676</i>
otters	0.9010	0.8509	0.6816
LMCO-IS & S	0.9009	0.8719	0.7760
SULTRG	0.8998	0.8676	0.7322
Annotator B	0.8997	0.8963	0.8973
GMJLJL	0.8975	0.8711	0.7334
ohsu_dmice	0.8938	0.8457	0.6542
Annotator C	0.8621	0.8454	0.8829

Experimental Results on Clinical Records (cont.)

Table 1. Results on the Medical NLP Challenge Data

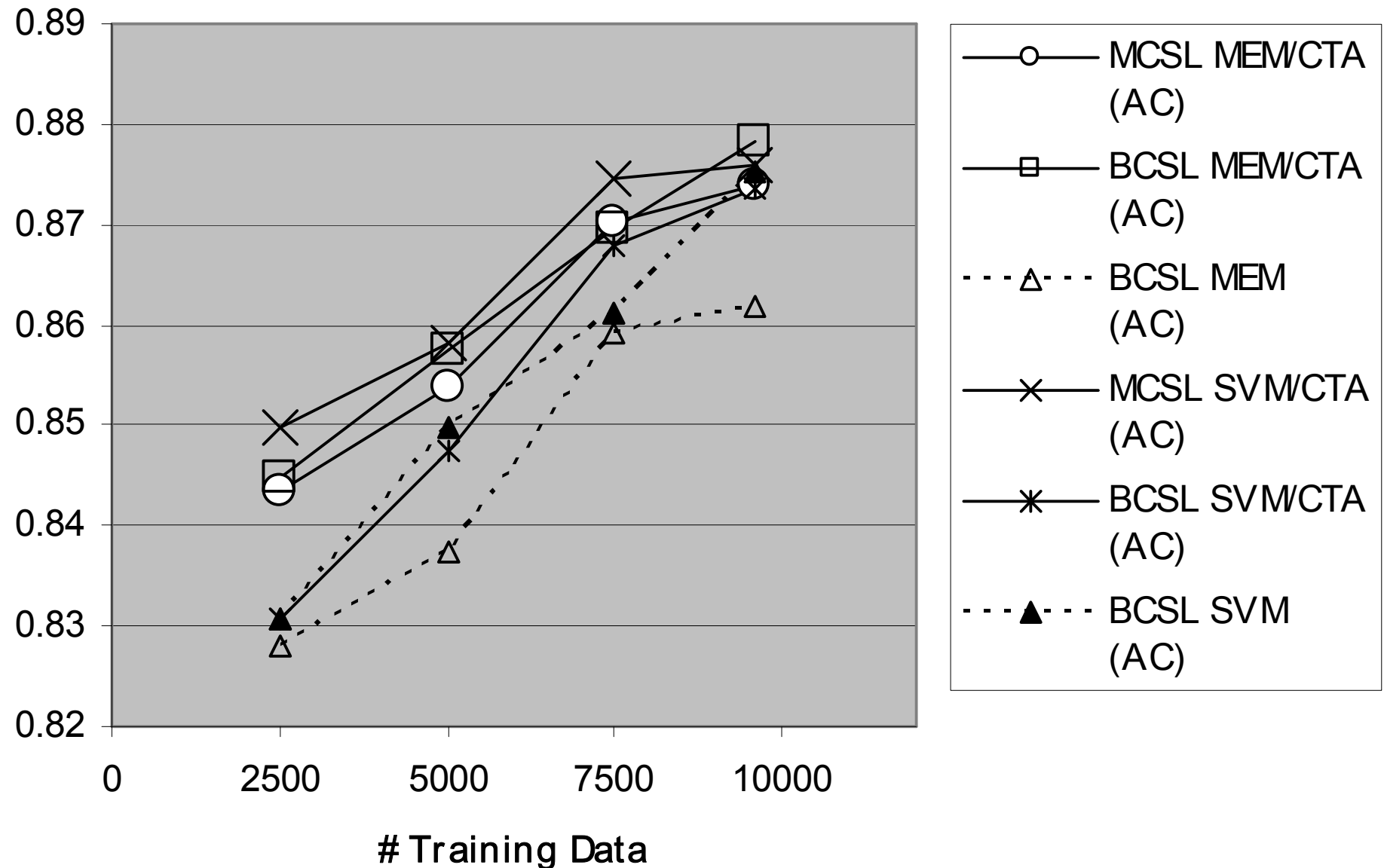
Feature	Features used								
uni-gram	x	x	x	x					
bi-gram		x	x						
tri-gram			x						
c-value				x				x	
tf-idf					x	x	x	x	x
section						x			x
negation							x	x	x
Interpretation	Scores								
MCSL-MEM/CTA Micro-average F1	0.8268	0.8263	0.8275	0.8292	0.8336	0.8386	0.8390	0.8396	0.8433
Multi-Topic AC	0.7367	0.7377	0.7367	0.7367	0.7500	0.7572	0.7551	0.7561	0.7643
BCSL-MEM/CTA Micro-average F1	0.7651	0.8264	0.8246	0.7863	0.7376	0.7444	0.7454	0.7651	0.7561
Multi-Topic AC	0.6629	0.7408	0.7398	0.6844	0.6373	0.6475	0.6486	0.6670	0.6577
BCSL-MEM Micro-average F1	0.7200	0.8105	0.8164	0.7440	0.6603	0.6550	0.6698	0.7049	0.6632
Multi-Topic AC	0.5912	0.6301	0.6424	0.6290	0.5307	0.5277	0.5420	0.5809	0.5410
MCSL-SVM/CTA Micro-average F1	0.7727	0.7947	0.7953	0.7851	0.8039	0.8147	0.8035	0.8037	0.8147
Multi-Topic AC	0.6762	0.7079	0.7080	0.6854	0.7182	0.7295	0.7172	0.7172	0.7295
BCSL-SVM/CTA Micro-average F1	0.8158	0.8198	0.8208	0.8196	0.8344	0.8414	0.8322	0.8306	0.8417
Multi-Topic AC	0.7254	0.7326	0.7336	0.7285	0.7520	0.7623	0.7541	0.7520	0.7643
BCSL-SVM Micro-average F1	0.8380	0.8454	0.8437	0.8452	0.8624	0.8584	0.8634	0.8672	0.8594
Multi-Topic AC	0.7396	0.7613	0.7602	0.7581	0.7859	0.7848	0.7859	0.7900	0.7848

Experimental Results on Reuters

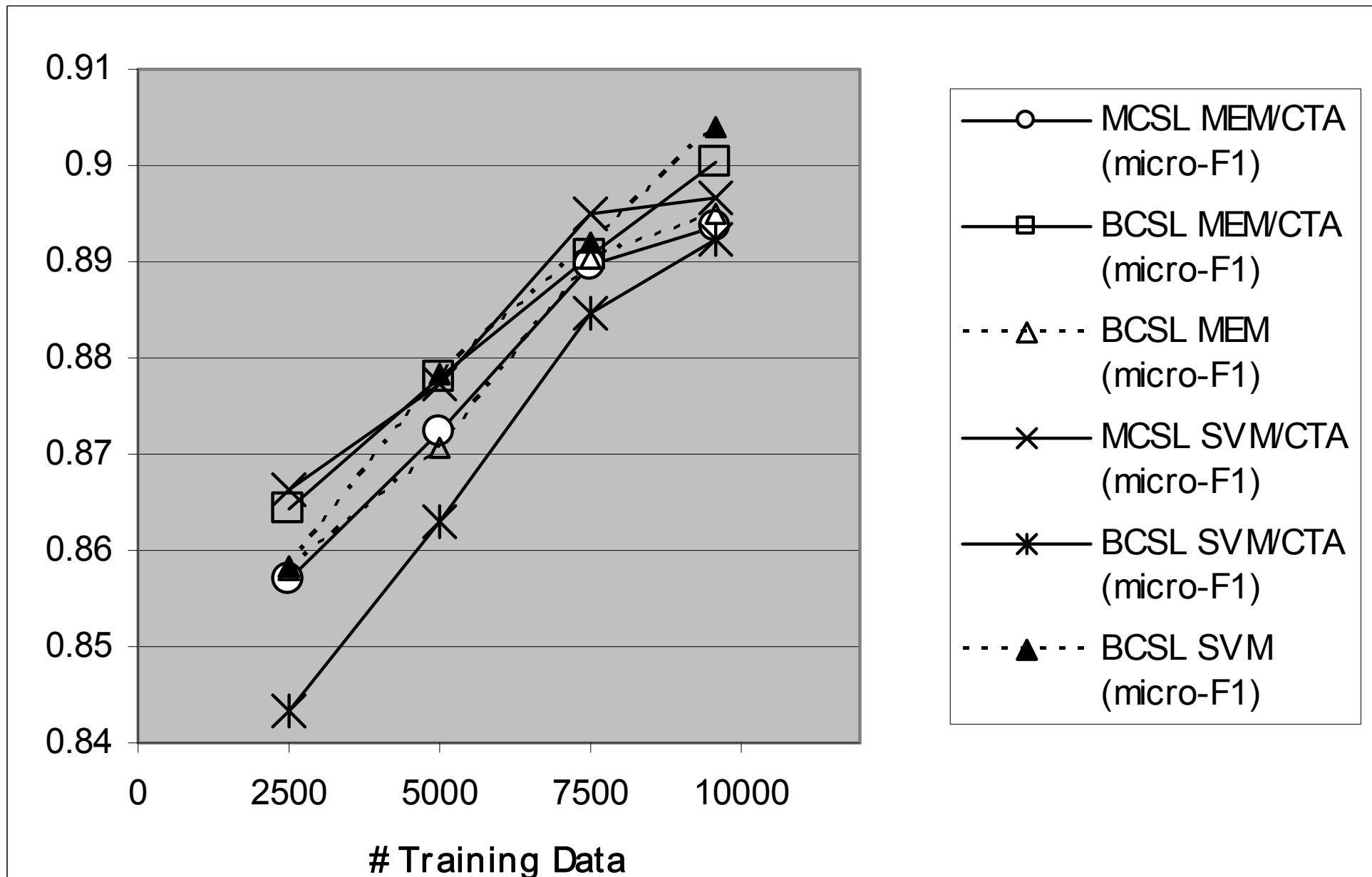
Table 1: Results on Reuters-21578 with 11 Topics

		F_1		
		micro	macro	AC
CTA	MCSL-MEM/CTA	0.8938	0.8279	0.8739
	BCSL-MEM/CTA	0.9003	0.8407	0.8781
OTA	BCSL-MEM	0.8949	0.8495	0.8618
CTA	MCSL-SVM/CTA	0.8930	0.8216	0.8745
	BCSL-SVM/CTA	0.8922	0.8150	0.8736
OTA	BCSL-SVM	0.9040	0.8435	0.8754

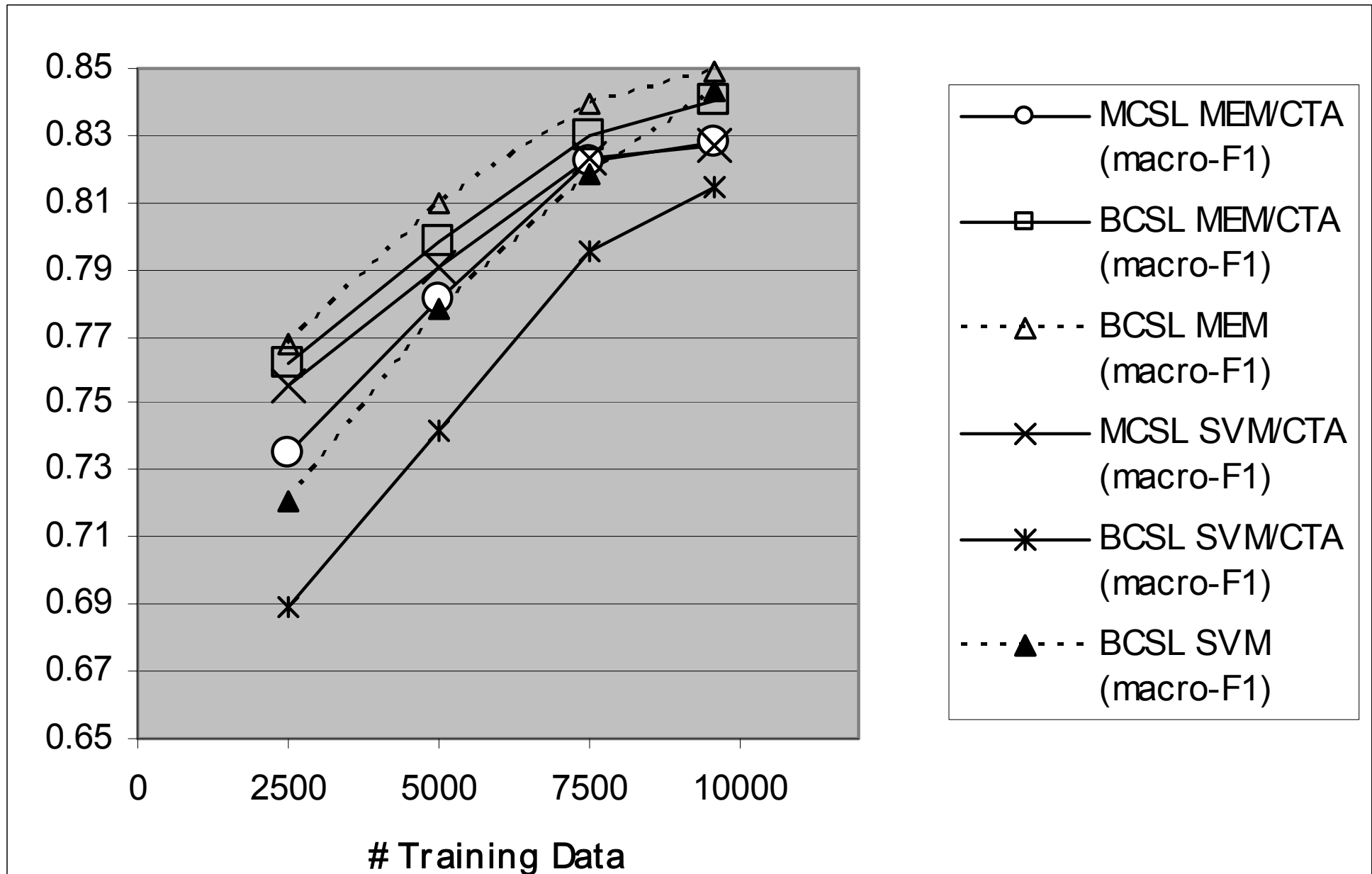
Multi-topic accuracy (Reuters)



Micro-average F1 (Reuters)



Macro-average F1 (Reuters)



References

Rule-based vs. Machine Learning Based Text Classification

Robert H. Creecy, Brij M. Masand, Stephen J. Smith, David L. Walt, Trading MIPS and memory for knowledge engineering, Communications of the ACM, Vol. 35, Issue 8, pp. 48-64, 1992.

Review paper on Text Classification

Fabrizio Sebastiani, Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No.1, pp.1-47, 2002.

CMC Medical NLP Challenge 2007

<http://www.computationalmedicine.org/challenge/index.php>

Clinical Text Classification

Yutaka Sasaki, Brian Rea, Sophia Ananiadou, Multi-Topic Aspects in Clinical Text Classification, IEEE International Conference on Bioinformatics and Biomedicine 2007 (IEEE BIBM-07), Silicon Valley, Nov. 2-7, 2007.

Selected papers on Text Classification

S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami, Inductive Learning Algorithms and Representations for Text Categorization, Proc. CIKM '98, pp.148-155, 1998.

Thorsten Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proc. of 10th European Conference on Machine Learning (ECML-98)}, pp.137-142, 1998.

A. McCallum, Multi-label Text Classification with a Mixture Model Trained by EM, AAAI-99 Workshop on Text Learning, 1999.

K. Nigam, J. Lafferty, A. McCallum, Using Maximum Entropy for Text Classification, IJCAI-99 Workshop on Machine Learning for Information Filtering, pp.61-67, 1999.

John C. Platt, Nello Cristianini, John Shawe-Taylor, Large Margin DAGs for Multiclass Classification, Proc. of NIPS-1999, pp. 547-553, 1999.

RE Schapire and Y Singer, BoosTexter: A Boosting-based System for Text Categorization, Machine Learning, Springer, Vol. 39, pp.135-168, 2000.

Thank you