Text Classification & Summarization

(Using Natural Language Processing and Machine Learning Techniques)

Ko, Youngjoong

July 5, 2017

Dept. of Computer Engineering,
Dong-A University

Contents

1. Warming Up

2. Text Classification

3. Text Summarization

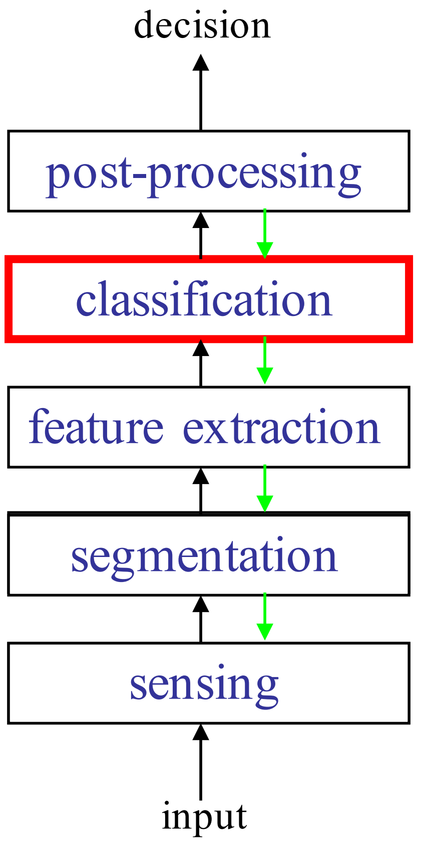
4. Combination of TC and TS

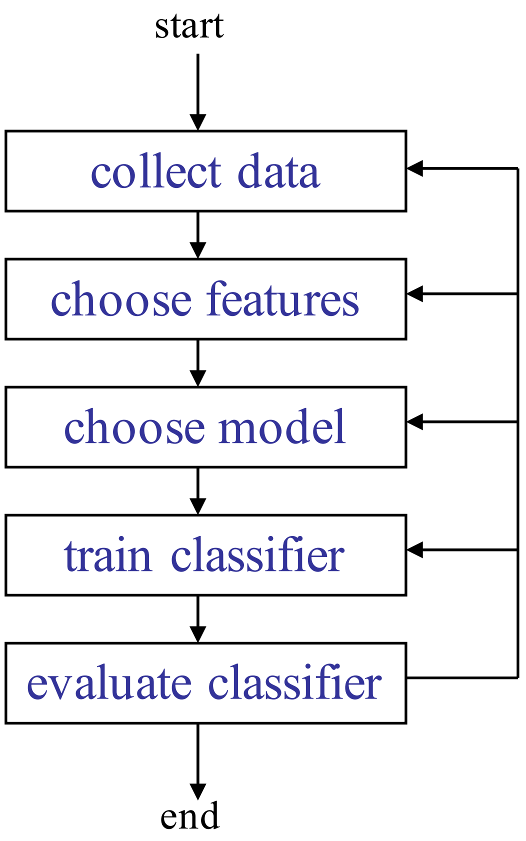
5. Latest Trend of TC and TS

6. References

Warming up!!

- **Pattern classification (Duda & Hart)**



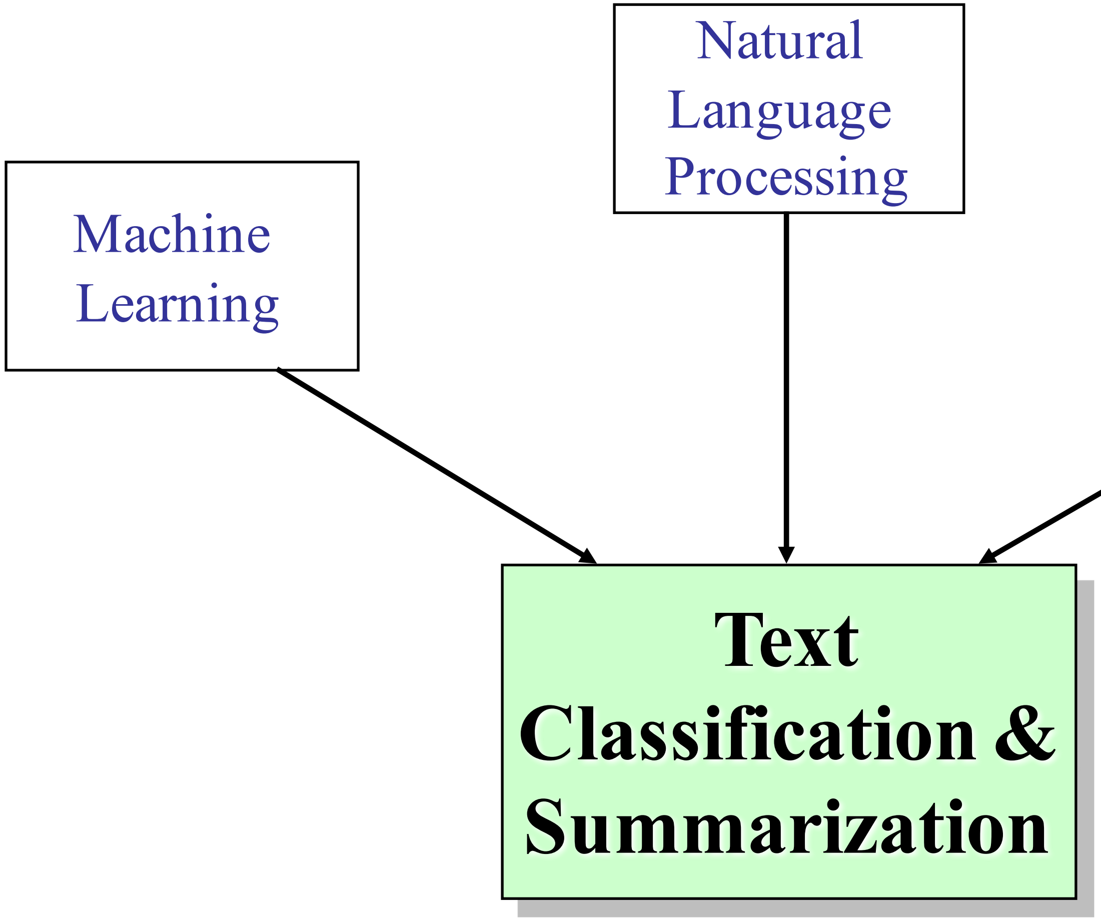


The process of the pattern classification system

The design cycle of the pattern classification system

Richard O. Duda, Peter E. Hart and David G. Stork, *Pattern Classification*, 2nd edition, John Wiley & Sons

Warming up!!


Natural Language Processing

Machine Learning

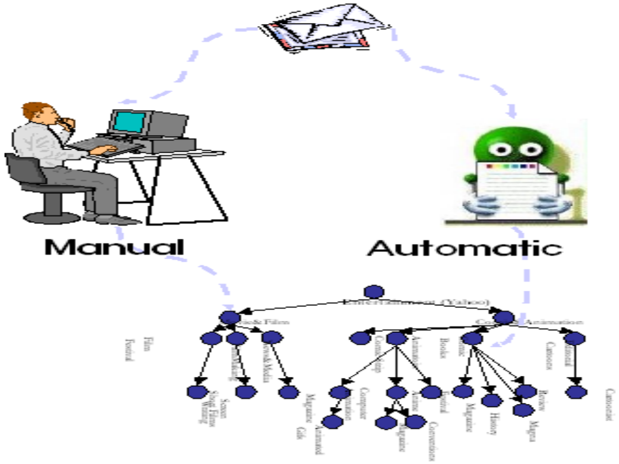
Information Retrieval

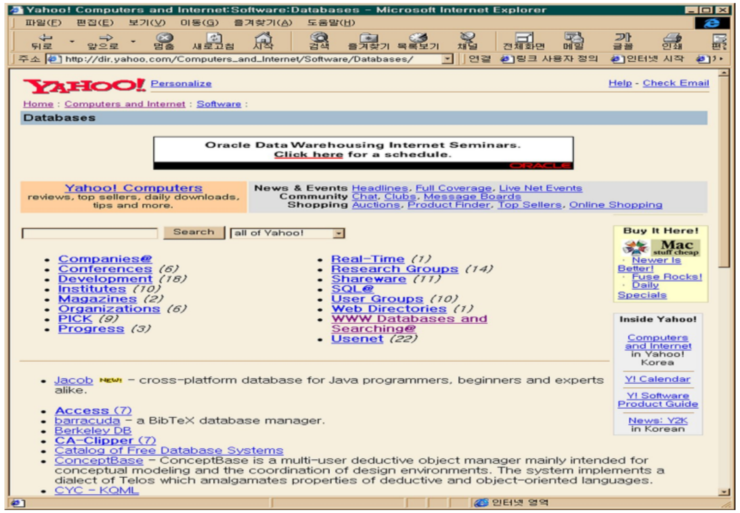
Text Classification & Summarization

Text Classification

Introduction

- **Classify documents into one (or several) of a set of *pre-defined categories (topics of interest)***
- **Prominent status in the information system field**
  - Explosion of electronic texts from article, WWW, e-mail, digital library, CRM, biomedical text etc.
- **The machine learning paradigm**
  - Supervised learning: Find decision rule from an example set of labeled documents




Yahoo! Computers and Internet:Software:Databases - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 이동(G) 즐겨찾기(A) 도움말(H)
뒤로 앞으로 멈춤 새로고침 시작 검색 즐겨찾기 목록보기 채널 전체화면 메일 글꼴 인쇄
주소 http://dir.yahoo.com/Computers_and_Internet/Software/Databases/
연결 링크 사용자 정의 인터넷 시작
YAHOO!
Search
all of Yahoo!
인터넷 영역

Personalize

Help - Check Email

Home : Computers and Internet : Software :

Databases

Oracle Data Warehousing Internet Seminars.
Click here for a schedule.
ORACLE

Yahoo! Computers
reviews, top sellers, daily downloads, tips and more.

News & Events Headlines, Full Coverage, Live Net Events
Community Chat, Clubs, Message Boards
Shopping Auctions, Product Finder, Top Sellers, Online Shopping

- Companies@
- Conferences (6)
- Development (18)
- Institutes (10)
- Magazines (2)
- Organizations (6)
- PICK (9)
- Progress (3)

- Real-Time (1)
- Research Groups (14)
- Shareware (11)
- SQL@
- User Groups (10)
- Web Directories (1)
- WWW Databases and Searching@
- Usenet (22)

Buy It Here!
Mac stuff cheap
- Newer Is Better!
- Fuse Rocks!
- Daily Specials

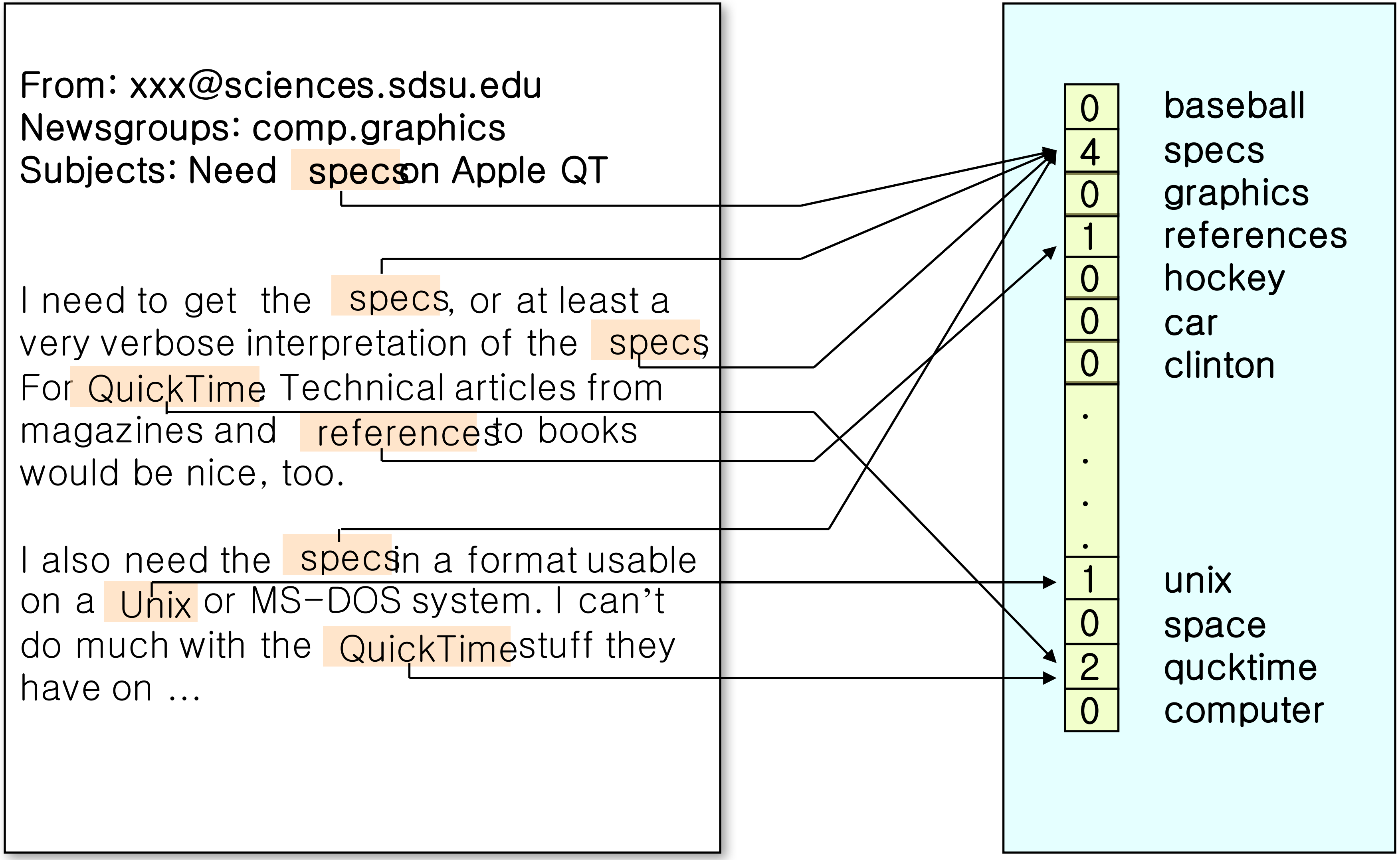
Inside Yahoo!
Computers and Internet in Yahoo! Korea
Y! Calendar
Y! Software Product Guide
News: Y2K in Korean

- Jacob NEW! - cross-platform database for Java programmers, beginners and experts alike.

- Access (7)
- barracuda - a BibTeX database manager.
- Berkeley DB
- CA-Clipper (7)
- Catalog of Free Database Systems
- ConceptBase - ConceptBase is a multi-user deductive object manager mainly intended for conceptual modeling and the coordination of design environments. The system implements a dialect of Telos which amalgamates properties of deductive and object-oriented languages.
- CYC - KQML

Text Classification

Text Representation



From: xxx@sciences.sdsu.edu
Newsgroups: comp.graphics
Subjects: Need specs on Apple QT

I need to get the specs, or at least a very verbose interpretation of the specs For QuickTime Technical articles from magazines and references to books would be nice, too.

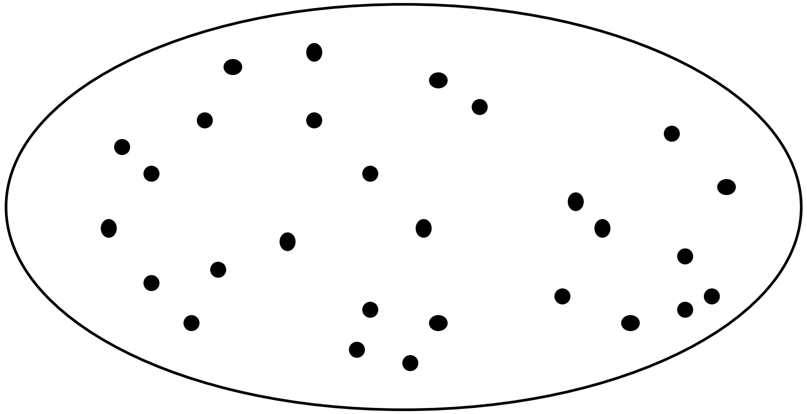
I also need the specs in a format usable on a Unix or MS-DOS system. I can't do much with the QuickTime stuff they have on …

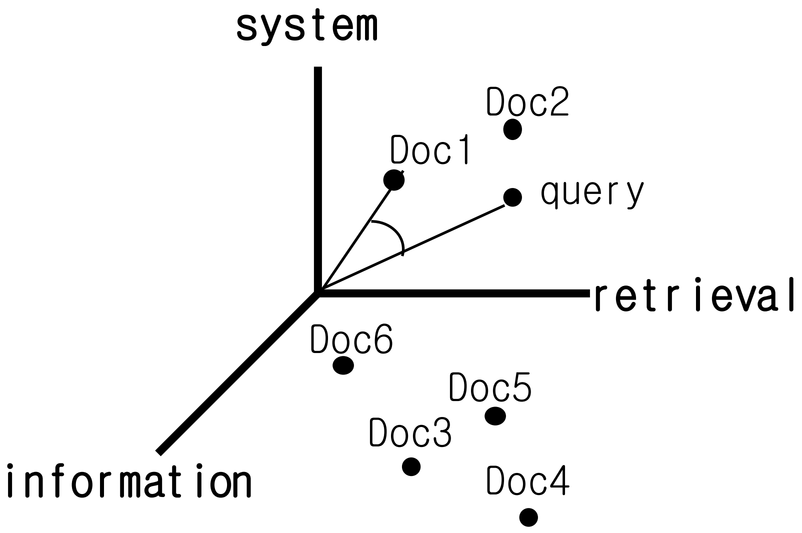
0	baseball
4	specs
0	graphics
1	references
0	hockey
0	car
0	clinton
.	
.	
.	
.	
1	unix
0	space
2	qucktime
0	computer

Text Classification

Vector Space Model

- **Multi-dimensional vector space**
  - A document is represented as a vector in vector space
  - Each dimension
    - Term or concept





Text Classification

Term Weighting Scheme

- **TFIDF term weight**

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#(t_k)}$$

- **Cosine Normalization**
  - The weights resulting from *tfidf*, so as to account for document length

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{r} (tfidf(t_s, d_j))^2}}$$

Text Classification

Semi-supervised Learning Based Text Classification

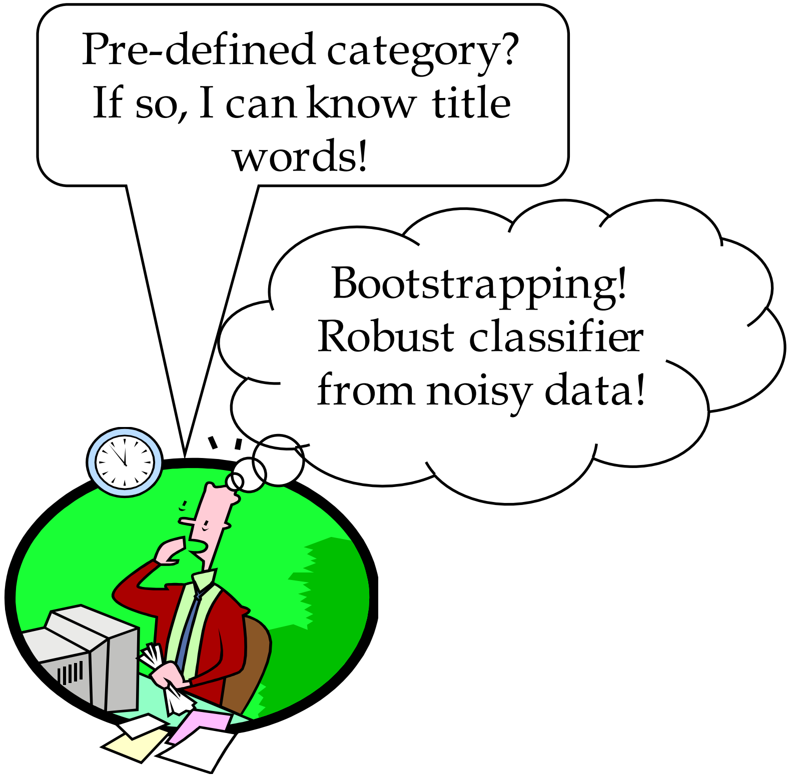
- **Difficulties of supervised learning in TC**
  - Require large, often prohibitive, number of labeled training data

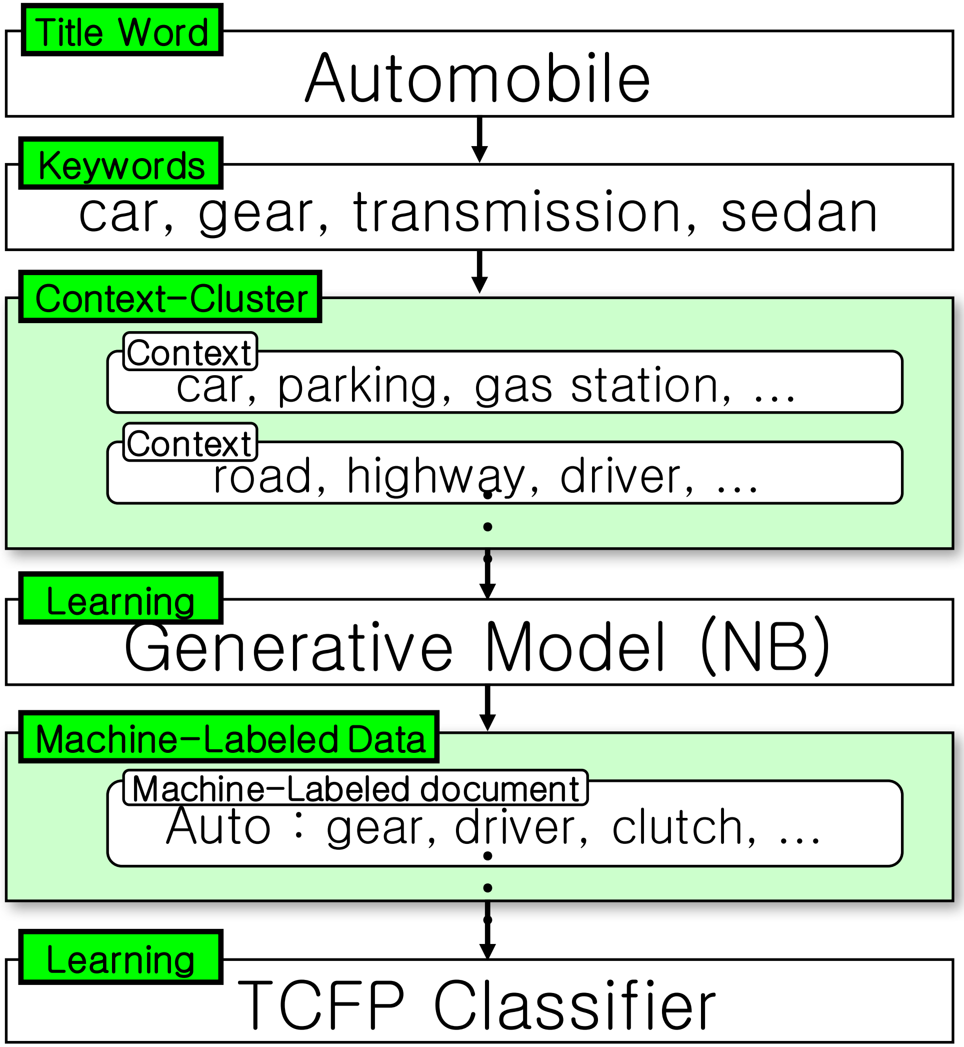
- **Semi-supervised learning in TC**
  - Automatically constructs labeled training data from unlabeled documents and the title word of each category
    - *How can we automatically generate labeled training documents (machine-labeled data) from only title words?*
      - ✓ Bootstrapping Framework
    - *How can we handle incorrectly labeled documents in the machine-labeled data?*
      - ✓ TCFP Classifier

[*Ph.D Dissertation, IPM 2009, IPM 2004, ACL 2004, Coling 2002*]

Text Classification

Semi-supervised Learning Based Text Classification


Pre-defined category? If so, I can know title words!
Bootstrapping! Robust classifier from noisy data!



Title Word

Automobile

Keywords

car, gear, transmission, sedan

Context-Cluster

Context

car, parking, gas station, …

Context

road, highway, driver, …

Learning

Generative Model (NB)

Machine-Labeled Data

Machine-Labeled document

Auto : gear, driver, clutch, …

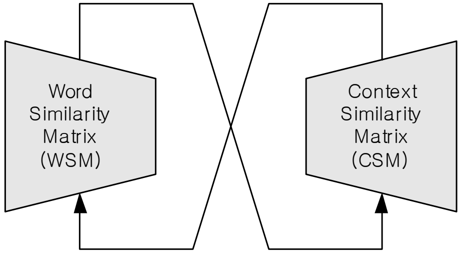
Learning

TCFP Classifier

Text Classification

Semi-supervised Learning Based Text Classification

- **Measuring similarity based on word & context similarity**
  - Two similarity matrices



- **Naïve Bayes with Minor Modification**
  - Kullback-Leibler Divergence

$$P(c_j \mid d_i;\hat{\theta}) = \frac{P(c_j \mid \hat{\theta})P(d_i \mid c_j;\hat{\theta})}{P(d_i \mid \hat{\theta})} \approx P(c_j \mid \hat{\theta})\prod_{t=1}^{|V|} P(w_t \mid c_j;\hat{\theta})^{N(w,d_i)}$$

$$\propto \frac{\log P(c_j;\hat{\theta})}{n} + \sum_{t=1}^{|V|} P(w_t \mid d_i;\hat{\theta})\log\left(\frac{P(w_t \mid c_j;\hat{\theta})}{P(w_t \mid d_i;\hat{\theta})}\right)$$

  - Laplace Parameter Estimation

$$\hat{\theta}_{w_t|c_j} \equiv P(w_t \mid c_j;\hat{\theta}) = \frac{1+N(w_t,G_{c_j})}{|V| + \sum_{i=1}^{|V|} N(w_t,G_{c_j})} \qquad \hat{\theta}_{c_j} \equiv P(c_j \mid \hat{\theta}) = \frac{1+|G_{c_j}|}{|C| + \sum_{c_i}|G_{c_i}|}$$

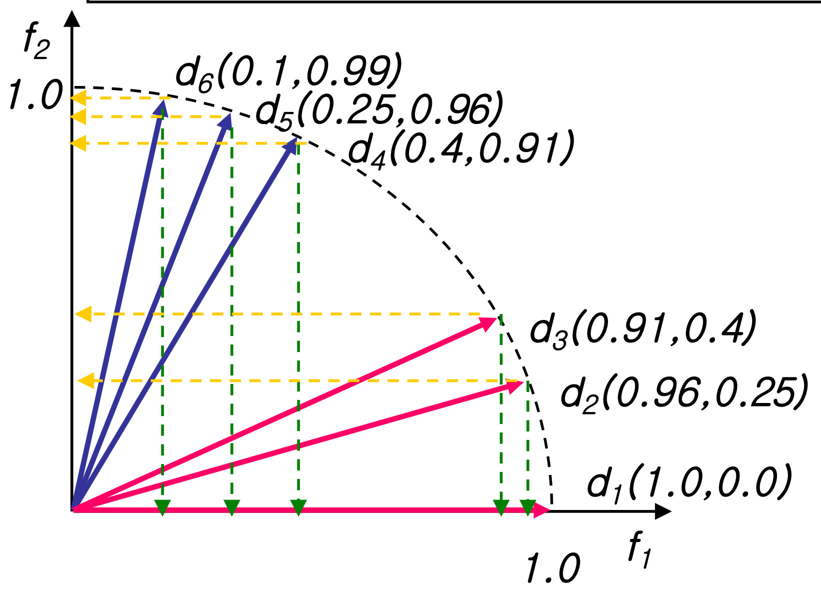
Text Classification

Semi-supervised Learning Based Text Classification

- **An example of feature projections in Text Categorization**

$d=(f_1, f_2),\ \ c_1=\{d_1, d_2, d_3\},\ \ \ c_2=\{d_4, d_5, d_6\}$

w: weight, *d*: document, *c*: category



w, d, c
1.0, d_1, c_1
0.96, d_2, c_1
0.91, d_3, c_1
0.4, d_4, c_2
0.25, d_5, c_2
0.1, d_6, c_2

feature projections on feature f_1

w, d, c
0.99, d_6, c_2
0.96, d_5, c_2
0.91, d_4, c_2
0.4, d_3, c_1
0.25, d_2, c_1

feature projections on feature f_2

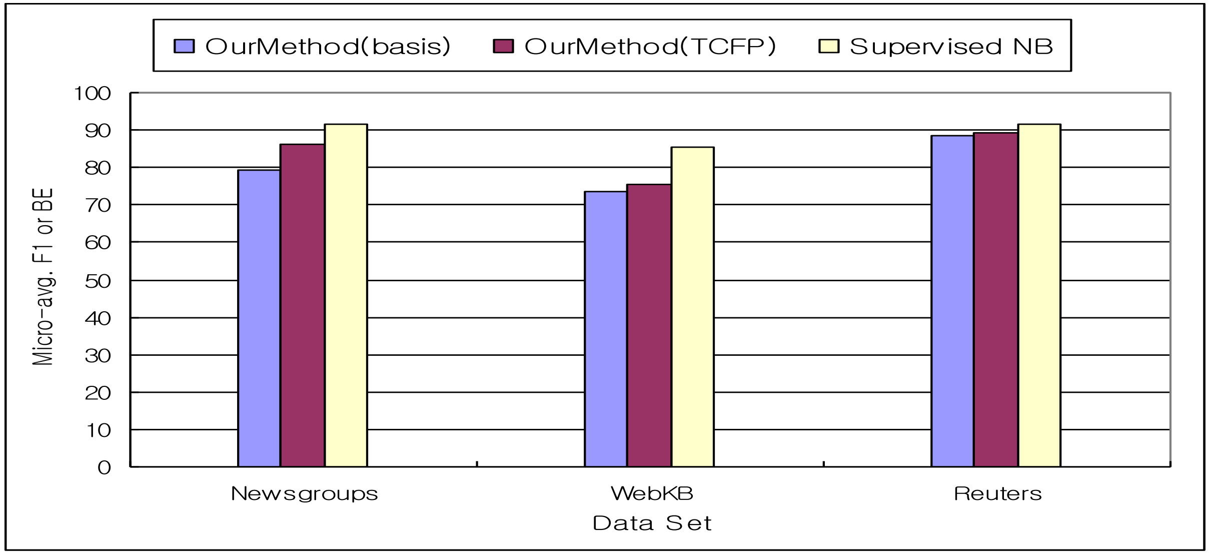
Document representation in a conventional vector space

Feature representation in feature projections

Text Classification

Semi-supervised Learning Based Text Classification

❖ **Final Results**



Data Set	***OurMethod (basis)***	***OurMethod (NB)***	***OurMethod (Rocchio)***	***OurMethod (k-NN)***	***OurMethod (SVM)***	***OurMethod (TCFP)***	***Supervised NB***
Newsgroups	***79.36***	*83.46*	*83*	*79.95*	*82.49*	***86.19***	***91.72***
WebKB	***73.63***	*73.22*	*75.28*	*68.04*	*73.74*	***75.47***	***85.29***
Reuters	***88.62***	*88.23*	*86.26*	*85.65*	*87.41*	***89.09***	***91.64***

Text Classification

Text Classification Using Revised EM Algorithm

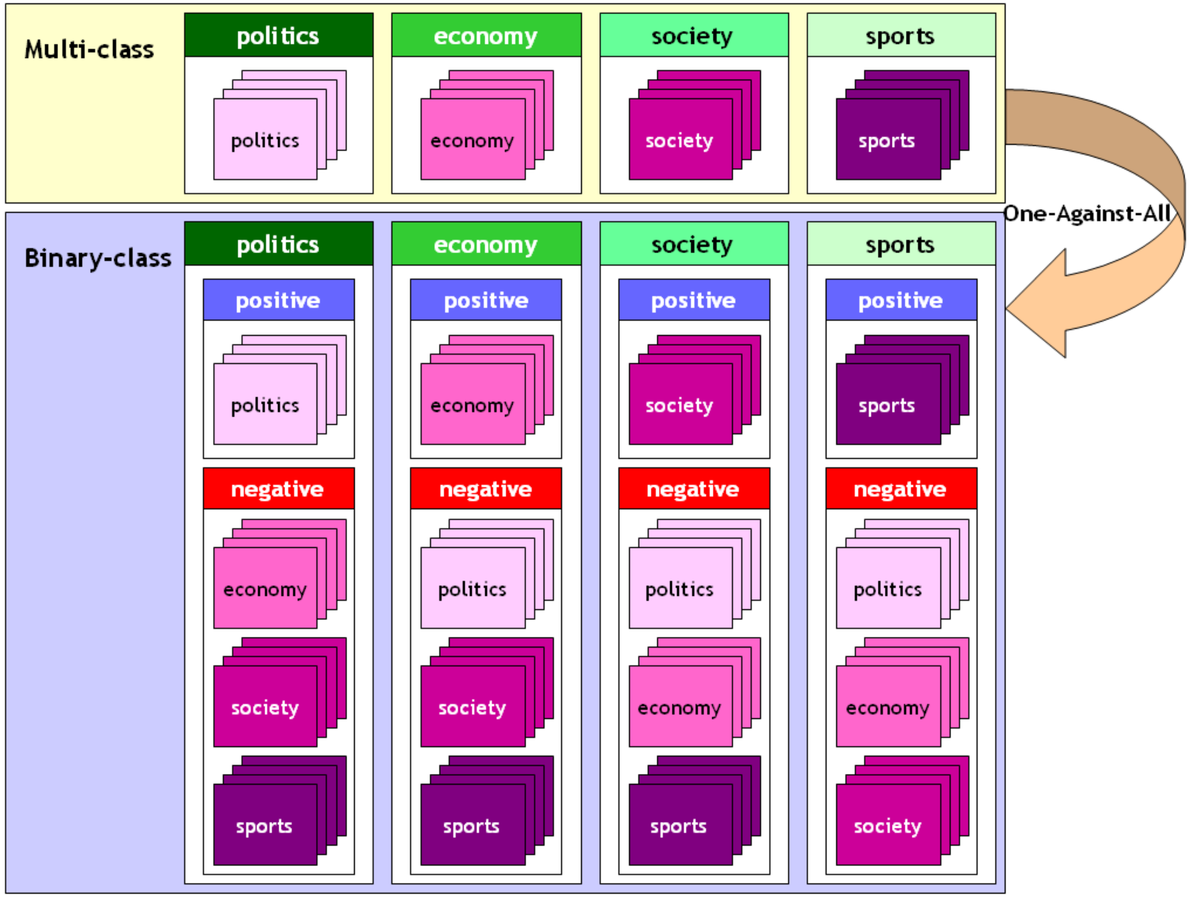
- **Problem of the one-against-the-rest method in TC**
  - Negative examples in the one-against-the-rest method have noisy examples

- **Solutions**
  - Automatically removing noisy examples by the sliding window technique and the revised EM (Expectation Maximization) algorithm
    - *How can we find a boundary area containing many noisy documents?*
      - ✓ Sliding window technique
    - *How can we deal with noisy documents found from the boundary?*
      - ✓ Revised EM algorithm

[*IPM 2007*, *AIRS 2004*]

Text Classification

- The multi-class setting with four categories changed into the binary setting using the One-Against-the-Rest method.



Multi-class

politics: politics

economy: economy

society: society

sports: sports

One-Against-All

Binary-class

politics: positive: politics; negative: economy, society, sports

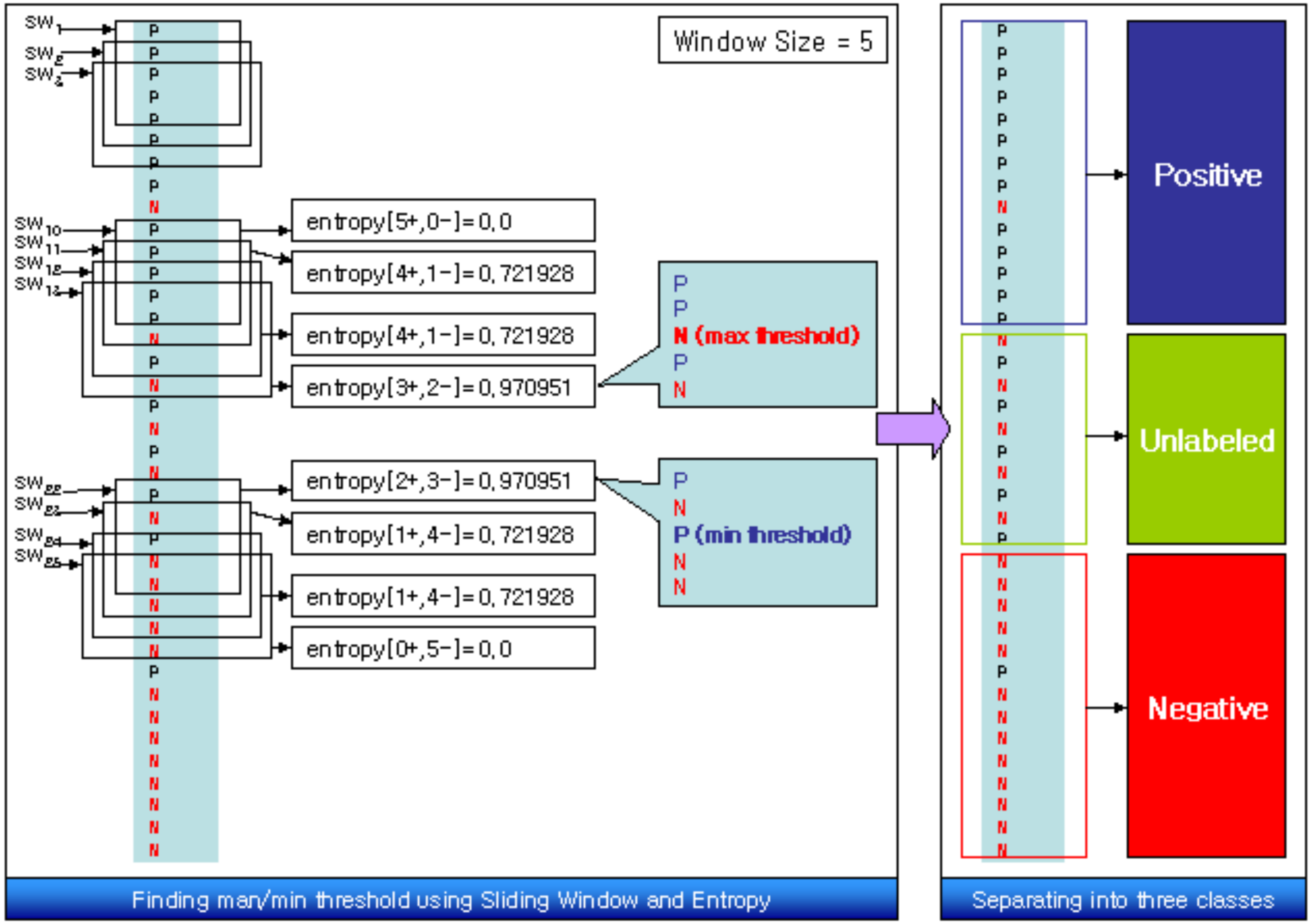
economy: positive: economy; negative: politics, society, sports

society: positive: society; negative: politics, economy, sports

sports: positive: sports; negative: politics, economy, society

Text Classification

Text Classification Using Revised EM Algorithm



Finding max/min threshold using Sliding Window and Entropy

Separating into three classes

Text Classification

Text Classification Using Revised EM Algorithm

❖ **Final Results**

Data Set	*k-NN (origin)*	*k-NN (proposed)*	*NB (origin)*	*NB (proposed)*	*Rocchio (origin)*	*Rocchio (proposed)*	*SVM (origin)*	*SVM (proposed)*
Newsgroups *(**micro**-avg.)*	86.07	***87.96*** ***(+2.19)***	83.17	***84.86*** ***(+2.03)***	82.84	***84.48*** ***(+1.98)***	88.34	***89.08*** ***(+0.84)***
Newsgroups *(**macro**-avg.)*	84.58	***87.03*** ***(+2.89)***	82.87	***84.55*** ***(+2.03)***	81.5	***83.57*** ***(+2.54)***	87.73	***89.08*** ***(+1.53)***
WebKB *(**micro**-avg.)*	84.97	***86.74*** ***(+2.08)***	85.67	***87.21*** ***(+1.8)***	86.52	***88.26*** ***(+2.01)***	92.12	***92.64*** ***(+0.56)***
WebKB *(**macro**-avg.)*	82.13	***85.55*** ***(+4.16)***	83.58	***86.53*** ***(+3.55)***	83.71	***87.03*** ***(+3.96)***	91.52	***92.17*** ***(+0.71)***
Reuters *(**micro**-avg.)*	91.47	***94.27*** ***(+3.06)***	90.80	***93.86*** ***(+3.37)***	89.24	***91.80*** ***(+2.86)***	94.66	***95.52*** ***(+0.91)***
Reuters *(**macro**-avg.)*	82.66	***85.43*** ***(+3.34)***	81.26	***86.38*** ***(+6.31)***	77.56	***83.55*** ***(+7.71)***	89.86	***90.72*** ***(+0.96)***

Text Classification

Improving Indexing Technique in Text Classification

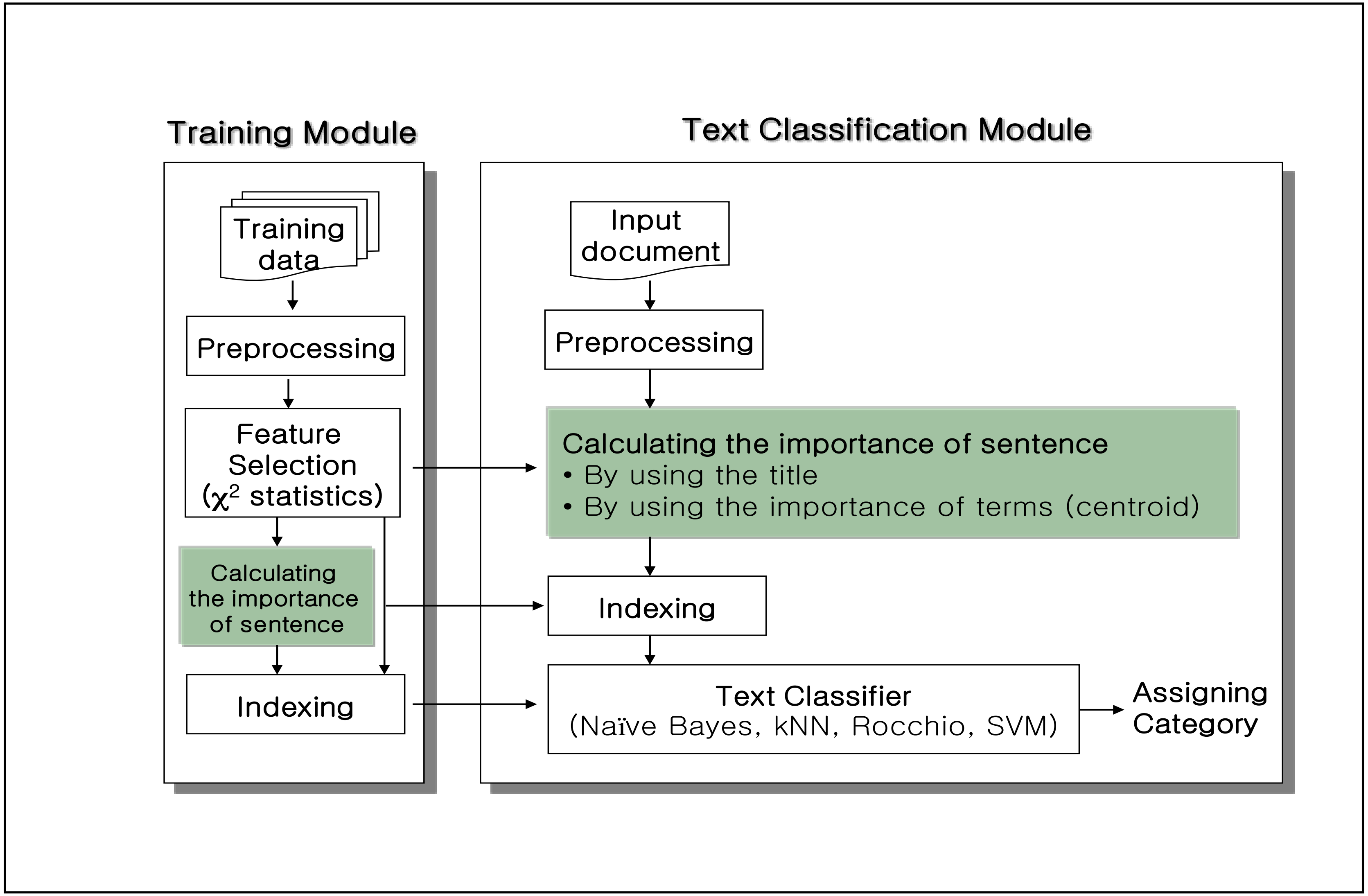
- **The conventional indexing technique in TC**
  - Vector space model in TFIDF

- **Improvement point**
  - Each sentence has different importance for identifying the content of document
    - *Text summarization techniques*
      - ✓ *Measuring similarity between the title and each sentence*
    - *Revised term weight*
      - ✓ Modified by the sentence importance value

[*IPM 2004*, *Coling 2004*]

Text Classification

Improving Indexing Technique in Text Classification



Training Module
Text Classification Module
Training data
Preprocessing
Feature Selection (χ^2 statistics)
Calculating the importance of sentence
Indexing
Input document
Preprocessing
Calculating the importance of sentence
• By using the title
• By using the importance of terms (centroid)
Indexing
Text Classifier
(Naïve Bayes, kNN, Rocchio, SVM)
Assigning Category

Text Classification

Improving Indexing Technique in Text Classification

- **Results in English Newsgroup data set**

	Naïve Bayes	
	Basis system	*Proposed system*
macro-avg F_1	*83.2*	***84.4***
micro-avg F_1	*82.9*	***84.3***

	Rocchio	
	Basis system	*Proposed system*
macro-avg F_1	*79.8*	***80.5***
micro-avg F_1	*79.4*	***80.3***

	k-NN	
	Basis system	*Proposed system*
macro-avg F_1	*81.3*	***82.7***
micro-avg F_1	*81.1*	***82.5***

	SVM	
	Basis system	*Proposed system*
macro-avg F_1	*85.8*	***86.4***
micro-avg F_1	*85.8*	***86.3***

Text Classification

Improving Indexing Technique in Text Classification

- **Results in Korean Newsgroup data set**

	Naïve Bayes	
	Basis system	Proposed system
macro-avg F_1	78.4	**80.8**
micro-avg F_1	79.1	**81.3**

	Rocchio	
	Basis system	Proposed system
macro-avg F_1	77.8	**79.2**
micro-avg F_1	78.7	**80.1**

	k-NN	
	Basis system	Proposed system
macro-avg F_1	78.6	**80.6**
micro-avg F_1	79.9	**81.3**

	SVM	
	Basis system	Proposed system
macro-avg F_1	84.8	**85.5**
micro-avg F_1	86.0	**86.5**

Text Classification

Term Weighting Scheme Using Class Information

- **Can we develop a novel term-weighting scheme for specialized text classification better than those used in information retrieval?**

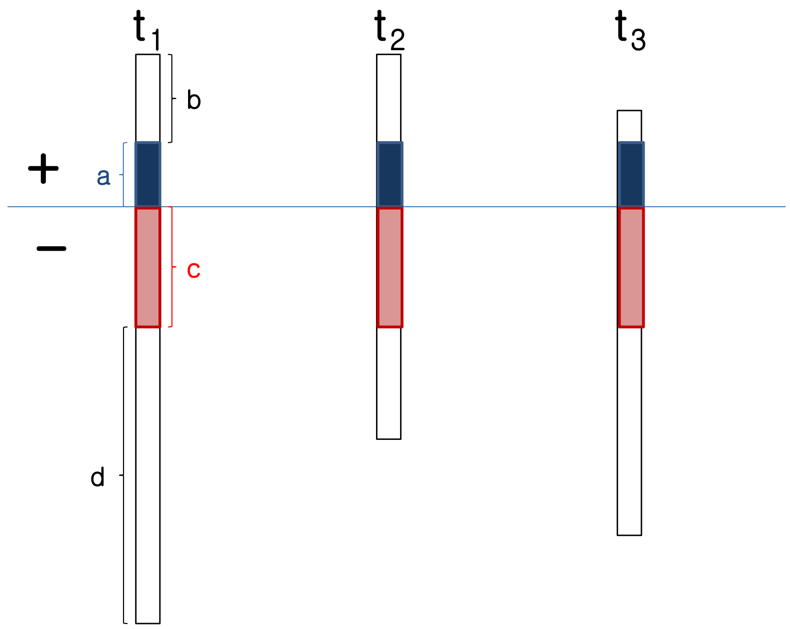
- **Class Information**
  - Supervised-learning based text classification has the training data labeled as either positive or negative for each class
    - *Novel term weighting scheme*
      - ✓ *Substitution for idf*

[*IPM 2017, JASIST 2015, PRL 2015, SIGIR 2012*]

Text Classification

Term Weighting Scheme Using Class Information

- Different distributions of positive and negative classes in a whole collection



Text Classification

Term Weighting Scheme Using Class Information

- **New Term Weighting Scheme Using Term Relevance Ratio (TRR)**

$$log\ \mathrm{tf.TRR} = (\log tf_{t_i} + 1) \cdot \log \frac{P(t_i|cl)}{P(t_i|\overline{cl})}$$

- **TRR Estimation**
  - Maximum Likelihood Estimation (MLE)

$$\mathrm{P}(t_i|\mathrm{cl}) = \frac{\sum_{j=1}^{|T_{cl}|} tf_{t_i,j}}{\sum_{k=1}^{|V|} \sum_{j=1}^{|T_{cl}|} tf_{t_k,j}},\ \mathrm{P}(t_i|\overline{cl}) = \frac{\sum_{j=1}^{|T_{\overline{cl}}|} tf_{t_i,j}}{\sum_{k=1}^{|V|} \sum_{j=1}^{|T_{\overline{cl}}|} tf_{t_k,j}}$$

  - Reformation of the MLE

$$\mathrm{P}(t_i|\mathrm{cl}) = \sum_{k=1}^{|T_{cl}|} P(t_i|d_k) \cdot P(d_k|cl)$$

$$\mathrm{P}(t_i|\overline{cl}) = \sum_{k=1}^{|T_{\overline{cl}}|} P(t_i|d_k) \cdot P(d_k|\overline{cl})$$

Text Classification

Term Weighting Scheme Using Class Information

❖ Final Results

			tf.idf	log tf.idf	log tf.chi	tf.rf	Delta tf.idf	log tf.TRR
Reuters (BEP)	kNN	macro-averaging	85.66	86.95	83.52	85.52	86.94	**90.48 (+4.06)**
		micro-averaging	92.46	93.29	89.99	92.68	92.46	**94.90 (+1.72)**
	SVM	macro-averaging	89.89	90.51	89.89	90.12	88.13	**91.68 (+1.29)**
		micro-averaging	94.87	94.86	93.83	94.76	93.00	**95.30 (+0.45)**
NG (F_1)	kNN	macro-averaging	81.13	85.13	71.53	82.96	86.22	**87.78 (+1.81)**
		micro-averaging	81.20	85.17	71.55	82.79	86.01	**87.75 (+2.02)**
	SVM	macro-averaging	86.78	87.46	80.54	87.13	86.35	**88.09 (+0.72)**
		micro-averaging	87.07	87.74	81.41	87.37	86.64	**88.43 (+0.78)**
KNG (F_1)	kNN	macro-averaging	77.92	79.34	63.34	69.9	79.73	**83.24 (+4.40)**
		micro-averaging	79.69	80.66	66.77	74.54	81.69	**84.36 (+3.27)**
	SVM	macro-averaging	83.64	84.14	74.84	81.61	80.41	**84.67 (+0.62)**
		micro-averaging	85.12	85.47	76.84	83.34	81.83	**85.70 (+0.27)**

Text Classification

Selected Published Papers

[International Journal Papers]

Youngjoong Ko (2017), “How to Use Negative Class Information for Naive Bayes Classification**”, Information Processing and Management**, Pergamon-Elsevier Science, 2017. **[SSCI, SCIE, Top 20%]**

Youngjoong Ko (2015), “A New Term Weighting Scheme for Text Classification Using the Odds of Positive and Negative Class Probabilities”, *Journal of the Association for Information Science and Technology*, Wiley-Blackwell, Vol. 66, No. 12, pp. 2553-2565, 2015. **[SSCI, SCIE, Top 10%]**

Youngjoong Ko (2015), ”New Feature Weighting Approaches for Speech-act Classification”, *Pattern Recognition Letters, Elsevier Science*, Vol. 51, pp. 107-111, January 2015. **[SCIE]**

Youngjoong Ko and Jungyun Seo (2009), “Text Classification from Unlabeled Documents with Bootstrapping and Feature Projection Techniques”, *Information Processing and Management*, Pergamon-Elsevier Science, Vol. 45, No. 1, pp. 70-83, 2009. **[SSCI, SCIE, Top 20%]**

Hyoungdong Han, **Youngjoong Ko** and Jungyun Seo (2007), “Using the Revised EM Algorithm to Remove Noisy Data for Improving the One-against-the-rest Method in Binary Text Classification”, *Information Processing and Management*, Pergamon-Elsevier Science, Vol. 43, No. 5, pp. 1281-1293, 2007. **[SSCI, SCIE, Top 20%]**

Youngjoong Ko and Jungyun Seo (2004), “Using the Feature Projection Technique Based on a Normalized Voting Method for Text Classification”, *Information Processing and Managemen*t, Pergamon-Elsevier Science, Vol. 40, No. 2, pp.191-208, 2004. **[SSCI, SCIE, Top 20%]**

Youngjoong Ko, Jinwoo Park and Jungyun Seo (2004), “Improving Text Categorization Using the Importance of Sentences”, *Information Processing and Management*, Pergamon-Elsevier Science, Vol. 40, No. 1, pp.65-79, 2004. **[SSCI, SCIE, Top 20%]**

Text Classification

Selected Published Papers

[International Conference Papers]

Kyoungman Bae and **Youngjoong Ko** (2014), "An Effective Category Classification Method Based on a Language Model for Question Category Recommendation on a cQA service", *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, pp. 2255-2258, in Maui Hawaii, USA, 2012. **[SCOPUS]**

Youngjoong Ko (2012), "A Study of Term Weighting Schemes Using Class Information for Text Classification", *Proceedings of the 35th Annual International ACM SIGIR Conference (SIGIR 2012)*, pp. 1029-1030, in Portland Oregon, USA, 2012. **[SCOPUS]**

Youngjoong Ko (2004), "Improving Binary Text Classification Using the EM Algorithm", *Proceedings of Asian Information Retrieval Symposium (AIRS 2004)*, pp. 325-328, in Beijing, China, 2004. **[SCOPUS]**

Youngjoong Ko and Jungyun Seo (2004), "Learning with Unlabeled Data for Text Categorization Using a Bootstrapping and a Feature Projection Technique", *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pp. 255-262, in Barcelona, Spain, 2004 **[SCOPUS]**

Youngjoong Ko and Jungyun Seo (2002), "Text Categorization Using Feature Projections", *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp. 467-473, in Taipei, Taiwan, 2002. **[SCOPUS]**

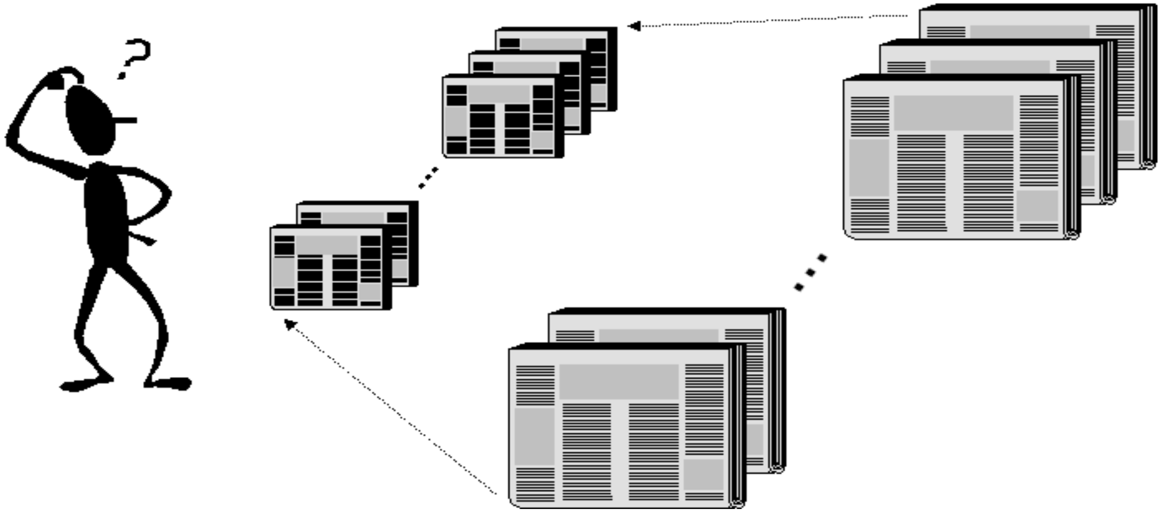
Youngjoong Ko, Jinwoo Park and Jungyun Seo (2002), "Automatic Text Categorization using the Importance of Sentences", *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pp. 474-480, in Taipei, Taiwan, 2002. **[SCOPUS]**

Youngjoong Ko and Jungyun Seo (2000), "Automatic Text Categorization by Unsupervised Learning", *Proceedings of The 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 453-459, in Saarbrucken, Germany, 2000. **[SCOPUS]**

Text Summarization

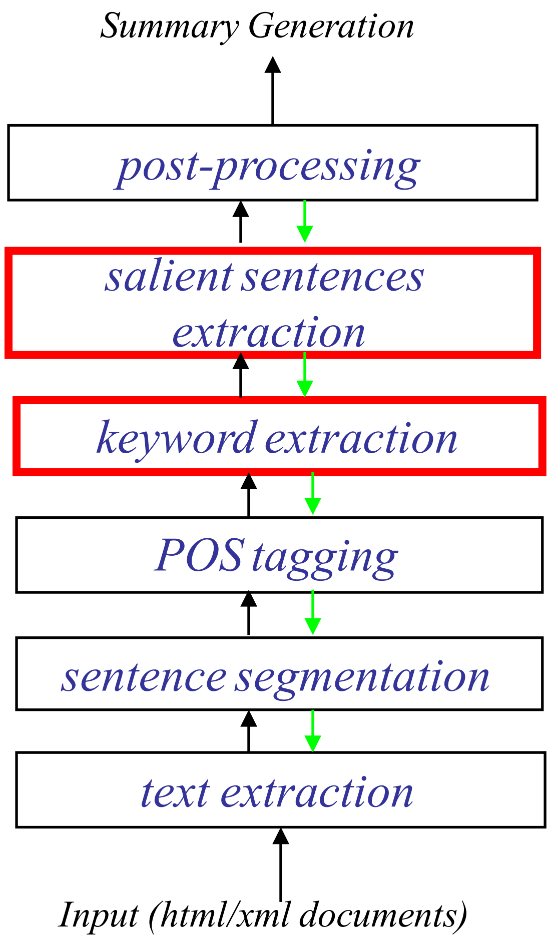
Introduction

- **Reducing the size of a document while preserving its content**
  - To extract content and present the most important content to a user in a condensed form
- **Text summarization (TS) system**
  - Identify the most salient information in a document
  - The most widespread summarization strategy: sentence extraction



Text Summarization

Fundamental Processes



Text Summarization

- Generic Summarization
  - Based on the content of a given text, ATS systems often produce generic summaries that highlight the most salient points of a given text
- Query-based Summarization
  - Most summaries of Web search engines such as Google are based on the query terms from the user's search: Snippet

Text Summarization

Fundamental Processes

- **Linguistic approaches**
  - High performance
  - Require high quality linguistic analysis tools (parser etc.) and linguistic resources (WordNet etc.)
- **Statistical approaches**
  - Easy to understand and implement, low cost
  - Generally low performance

Text Summarization

General Statistical Methods

- **Title Method**
  - How many words are commonly used between the sentence and title
  - Boolean weighted vector space model

$$Score(S_i) = sim(S_i, Q) \qquad sim(S_i, Q) = \sum_{k=1}^{n} w_{ik} w_{jk}$$

- **Location Method**
  - Leading several sentences of an article are important and a good summary

$$Score(S_i) = 1 - \frac{i-1}{N}$$

Text Summarization

General Statistical Methods

- **Aggregation Similarity Method**
  - The sum of similarity with other all sentence vectors

$$sim(S_i, S_j) = \sum_{k=1}^{n} w_{ik} w_{jk} \qquad asim(S_i) = \sum_{j=1, j \neq i}^{n} sim(S_i, S_j)$$

- **Frequency Method**
  - The sum of *tf-idf* term weights of words in each sentence

$$Score(S_i) = \sum_{k=1}^{n} (tf_i \times \log \frac{N}{df_i})$$

Text Summarization

General Statistical Methods

TF-based query method

- For no-title cases
- High frequent words as keywords instead of title

$$sim(S_i, TfQ) = \sum_{k=1}^{n} w_{ik} w_{TFQk}$$

Text Summarization

The Performance of Generic Statistical Methods

Methods	30%	10%
Title	**48.8**	**43.5**
Location	**49.4**	**46.6**
TF based query	**45.6**	**46.5**
Aggregation Similarity	**40.6**	**23.9**
Frequency	**35.2**	**13.0**

Text Summarization

Using Lexical Clustering in Text Summarization

- **Topic keyword identification using lexical clustering**
  - Automatic detection of topic words is very useful in Text Summarization

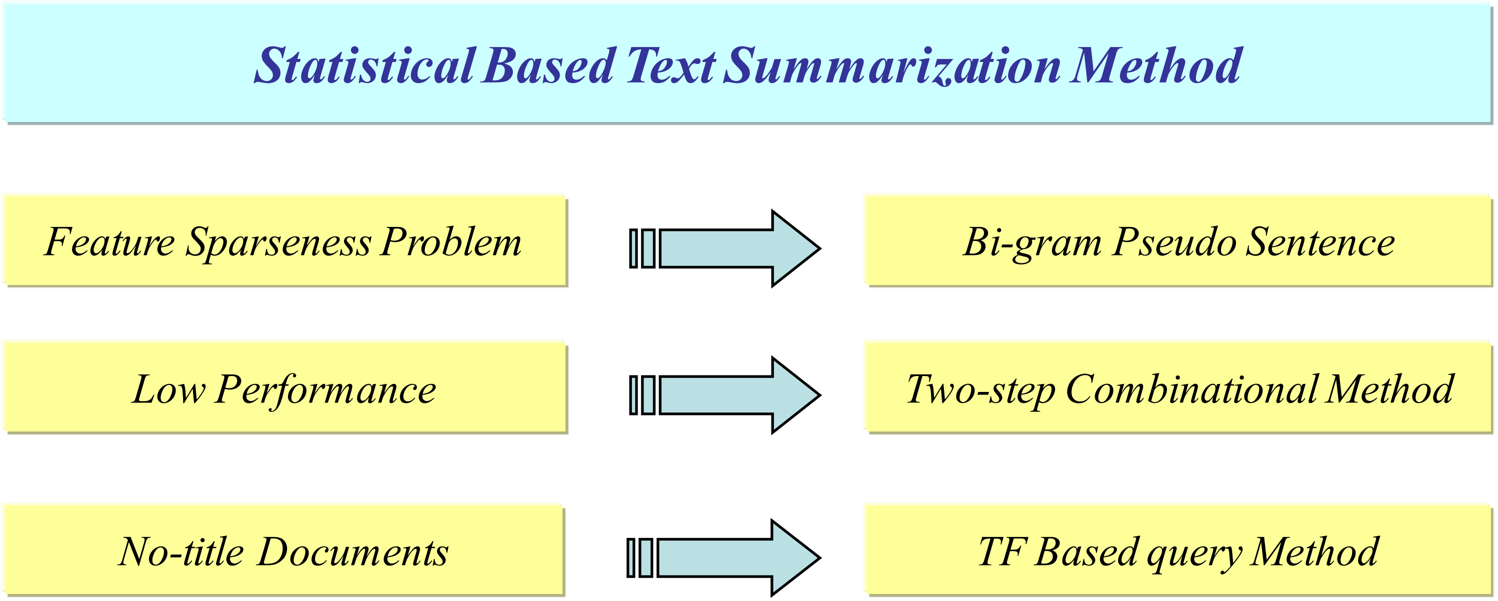
- **The proposed method**
  - Keyword identification for text summarization
    - *Using co-occurrence statistics*
      - ✓ Context vector space
    - *Using context vector space and k-means algorithm*
      - ✓ Lexical clustering

[*IEICE 2003*]

Text Summarization

Two-step Text Summarization

- High performance summarization method to efficiently combine statistical approaches



Statistical Based Text Summarization Method

Feature Sparseness Problem → *Bi-gram Pseudo Sentence*

Low Performance → *Two-step Combinational Method*

No-title Documents → *TF Based query Method*

[*PRL 2008*, *AIRS 2004*]

Text Summarization

First Step: Removing Noisy Sentences

- **The goal of the first step**
  - Not extract salient sentences but reduce noisy sentences
- **Bi-gram pseudo sentences**
  - Solve the feature sparseness problem
    - Can get few feature information from only single sentence
  - A new meaning unit
    - Two adjacent sentences
- **The linear combination method in the first step**
  - Title and Location methods
  - Remove about 50% noisy bi-gram pseudo sentences

$$\boxed{Score(S_i) = sim(S_i, Q) + (1 - \frac{i-1}{N})}$$

Text Summarization

Second Step: Extracting Summary

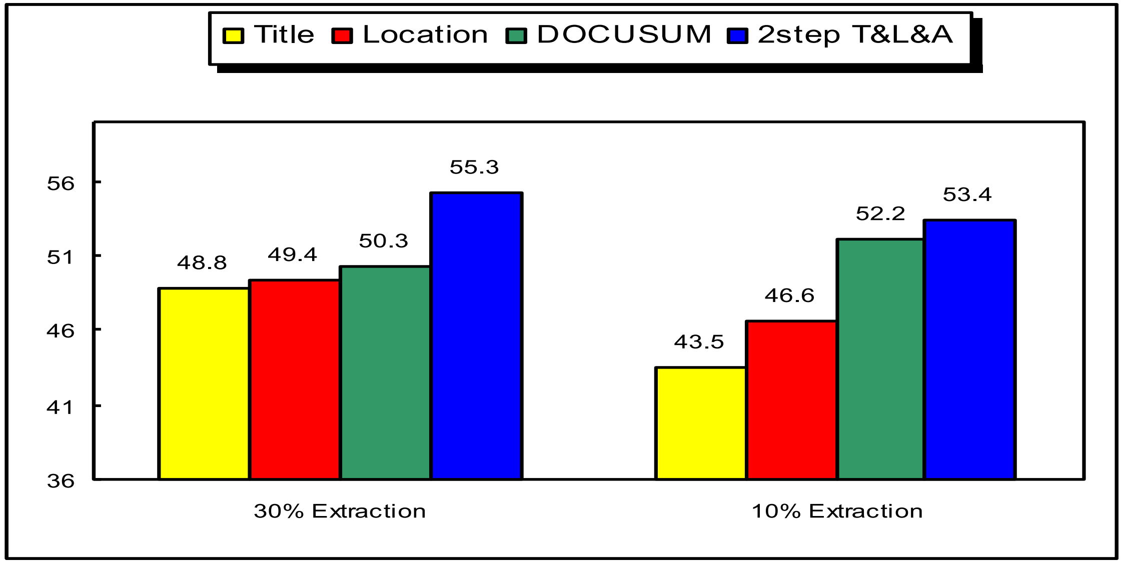
- **The goal of the second step**
  - Generate summary from extracting the salient original single sentences
- **Separate the remaining bi-gram pseudo sentences into original single sentences**
- **Aggregation Similarity Method**
  - Since noisy sentences are eliminated in the first step, we can get important score with good quality.

$$\boxed{Score(S_i) = sim(S_i, Q) + (1 - \frac{i-1}{N}) + w_a asim(S_i)}$$

Text Summarization

Evaluation Results

- **Comparing with other summarization methods with title**
  - Title, Location, DOCUSUM



	DOCUSUM	Two-step	Improvement
10%	**52.2**	**53.4**	*+1.2*
30%	**50.3**	**55.3**	*+5.0*

Text Summarization

Query-based Summarization

- **Using Relevance Feedback Technique**

  - *Each sentence is segmented.*
  - *Relevant and non-relevant sentences are separated by whether or not each sentence includes a query term.*
  - *The relevance weight of candidate terms from relevant sentences is estimated by the statistical weighting function and the initial query is expanded by using the candidate terms with the high relevance weight (TSV).*
  - *The important score of each sentence is estimated by using TSV and the location information of expanded query terms.*
  - *Finally, a snippet is generated by sentences with a high important score*

[*IPL 2008*, *SIGIR 2007*]

Text Summarization

Query-based Summarization

- **Term Selection Value (TSV)**

$$w_t = TSV_t = \log \frac{p(1 - q)}{q(1 - p)} = \log \frac{(r + 0.5)(S - s + 0.5)}{(R - r + 0.5)(s + 0.5)}$$

- **The Importance Score Estimation using TSVs**

$$Score(S_i) = \alpha \left(\frac{RWscore(S)}{RWscoreMax} \right) + (1 - \alpha)\left(1 - \frac{i - 1}{N}\right)$$

Text Summarization

Evaluation Results

	Title Method	Proposed Method	Search Engine
Naver Data Set	**56.6%**	**67.1%**	**20.4%**
Google Data Set	**57.4%**	**68.7%**	**59.5%**

Text Summarization

Selected Published Papers

[International Journal Papers]

Hyoungil Jeong , **Youngjoong Ko** and Jungyun Seo (2015), "Efficient Keyword Extraction and Text Summarization for Reading Articles on a Smart Phone", *Computing and Informatics*, Vol. 34, No. 4, pp. 779-794, 2015. **[SCIE]**

Youngjoong Ko, Hongkuk An and Jungyun Seo (2008), "Pseudo-Relevance Feedback and Statistical Query Expansion for Web Snippet Generation", *Information Processing Letters*, Elsevier Science, Vol. 109, No. 1, pp. 18-22, 2008. **[SCIE]**

Youngjoong Ko and Jungyun Seo (2008), "An Effective Sentence-Extraction Technique Using Contextual Information and Statistical Approaches for Text Summarizatione", *Pattern Recognition Letters*, Elsevier Science, Vol. 29, No. 9, pp. 1366-1371, 2008. **[SCIE]**

Youngjoong Ko, Kono Kim and Jungyun Seo (2003), "Topic Keyword Identification for Text Summarization Using Lexical Clustering", *IEICE Transactions on Information and System*, Vol. E86-D, No. 9, pp.1695-1701, 2003. **[SCIE]**

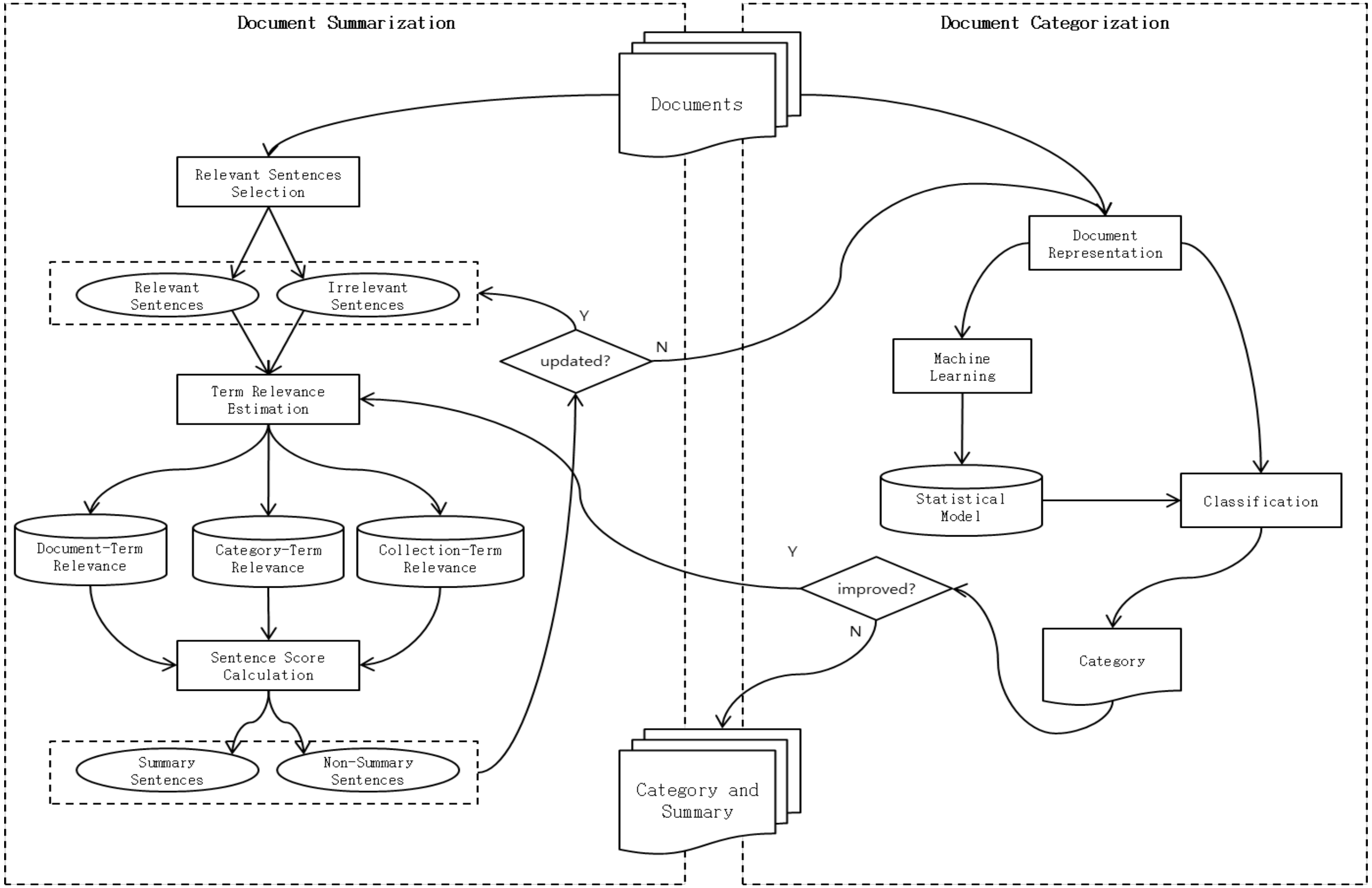
[International Conference Papers]

Youngjoong Ko, Hongkuk An and Jungyun Seo (2007), "An Effective Snippet Generation Method Using the Pseudo Relevance Feedback Technique", *Proceedings of the 30th Annual International ACM SIGIR Conference (SIGIR 2007)*, pp. 711-712, in Amsterdam, The Netherlands, July 2007. **[SCOPUS]**

Wooncheol Jung, **Youngjoong Ko** and Jungyun Seo (2004), "Automatic Text Summarization Using Two-step Sentence Extraction", *Proceedings of Asian Information Retrieval Symposium (AIRS 2004)*, in Beijing, China, pp.43-48, Oct, 2004.

Combination of TC and TS

Interactive Framework of the Summarization and Categorization



Hyoungil Jeong, **Youngjoong Ko** and Jungyun Seo (2016), "How to Improve Text Summarization and Classification by Mutual Cooperation on an Integrated Framework", ***Expert Systems with Applications***, Vol. 60, pp. 222-233, 2016. **[SCIE, Top 10%]**

Combination of TC and TS

Enhancing Summarization

- **Importance score of sentence using Category-based Language Model**

$$
\begin{aligned}
Score(S_i) = \; & \alpha \times \sigma(Score_d(S_i)) + && \text{Intra-document Information} \\
& (1.0-\alpha)\times\beta \times \sigma(Score_c(S_i)) + && \text{Intra-Category Information} \\
& (1.0-\alpha)\times(1.0-\beta)\times\sigma(Score_{COL}(S_i)) && \text{Global Information}
\end{aligned}
$$

$$
,where \quad \alpha = \frac{\overset{\text{Size of document}}{|d|}}{|d| + \underset{\text{Average size of document}}{\mu_1}} \;, \quad \beta = \frac{\overset{\text{Size of category}}{|c|}}{|c| + \underset{\text{Average size of category}}{\mu_2}}
$$

$$
\sigma(x) = \frac{1.0}{1.0 + exp(-x)}
$$

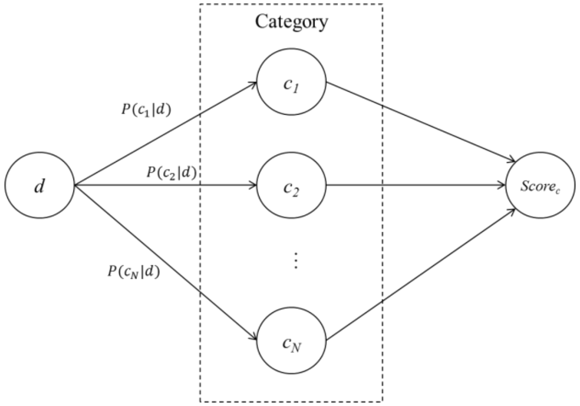
Combination of TC and TS

Enhancing Summarization

- **Applying Probabilistic Latent Class Model**
  - Similar to Soft Clustering or Topic Modeling
  - Deal with the error propagation problem from the classifier

$$Score(S_i) = \begin{matrix} \alpha & \times\sigma(Score_d(S_i))+ \\ (1.0-\alpha)\beta & \times\sum_c P(c|d)\times\sigma(Score_c(S_i))+ \\ (1.0-\alpha)(1.0-\beta) & \times\sigma(Score_{COL}(S_i)) \end{matrix}$$

$$,where \quad \sum_c P(c|d)=1$$



Combination of TC and TS

Enhancing Classification

Background

- Previous classification methods use only frequencies of each own term
- The term-frequency does not reflect the importance of sentence

Proposition

- Apply the term relevance to the document representation of classification

$$w'_j = w_j \times (0.5 + \sigma(TR_s(t_j)))$$

$$\sigma(x) = \frac{1.0}{1.0 + exp(-x)}$$

$$TR_s(t) = log\{\frac{(r_s + 0.5)(S_s - s_s + 0.5)}{(R_s - r_s + 0.5)(s_s + 0.5)}\}$$

$$, where\ R_s + S_s = |d|$$

w_s: the typical feature weight of j-th term t_j
w'_s: the proposed feature weight of t_j

r_s: the # of summary sentences that include t_j in document d
s_s: the # of non-summary sentences that include t_j in d

R_s: the # of relevant sentences in d
S_s: the # of irrelevant sentences in d

Summary of Evaluations

Evaluation Results

❖ In Document Summarization

Summarization Methods		KORDIC	AbleNews
Without Category Information	Proposed system	**0.567**	**0.612**
	Previous best system (Contextual Information)	0.553	0.564
With Category Information	Proposed system	**0.614**	**0.664**
	Previous best system (Contextual Information)	0.588	0.614

❖ In Document Classification

Classification Methods		KORDIC	AbleNews
Categorization	Proposed system	**0.784**	**0.890**
	Previous best system (SVM with TF-IDF)	0.760	0.852
Clustering	Proposed system	**0.545**	**0.598**
	Previous best system (LDA with TF)	0.518	0.549

Latest Trend of TC and TS

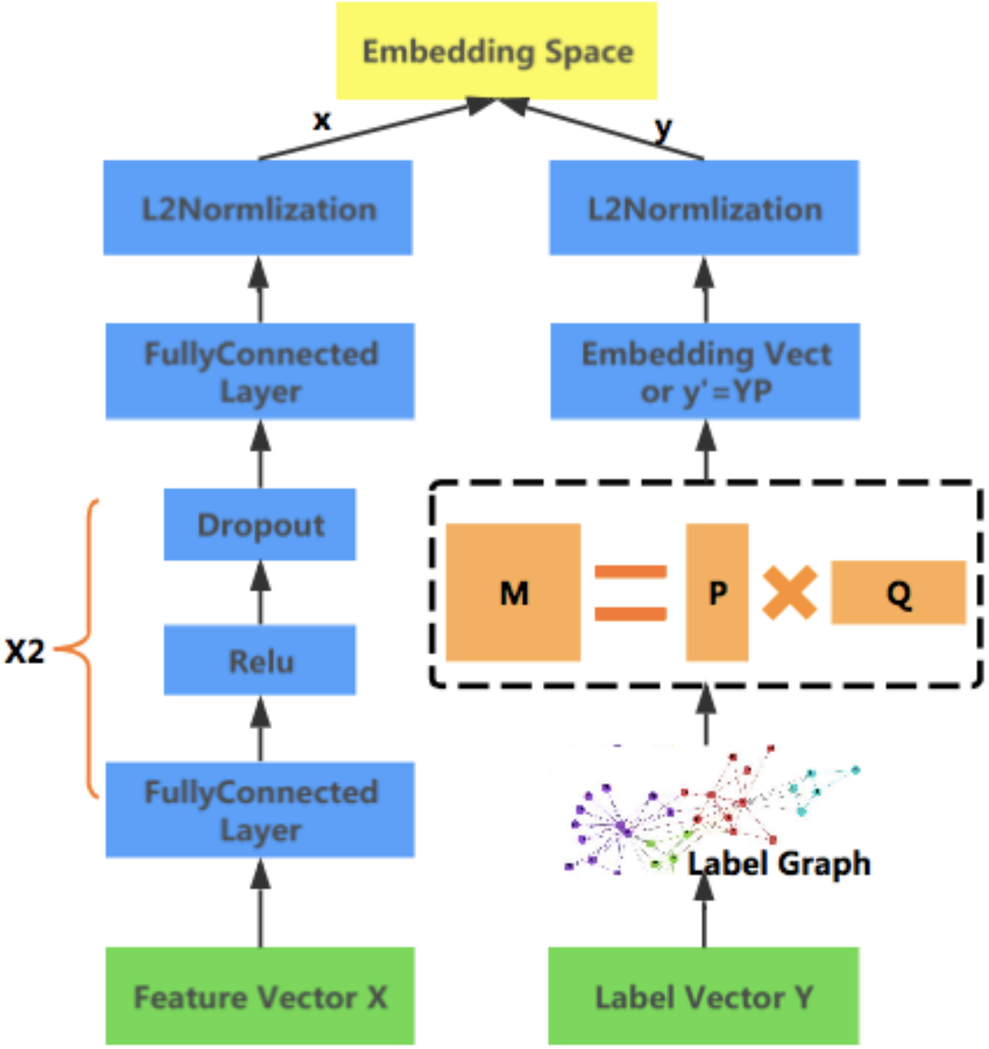
Deep eXtreme Multi-label Learning (XML)

- **XML focuses on tackling the problem of extremely high input dimensions for both input feature dimension and label dimension.**
  - Allows for the co-existence of more than one labels for a single data sample
  - One-against-the-rest classifiers are not feasible since it will be almost computationally intractable to train a massive number of one million classifiers.
  - Tree based method and embedding based method
  - **Embedding based method**: Projecting the high dimensional label vectors onto a low dimension linear subspace

Wenjie Zhang, Liwei Wang, Junchi Yan, Xiangfeng Wang and Hongyuan Zha (2017), "Deep Extreme Multi-label Learning", CoRR abs/1704.03718.

Latest Trend of TC and TS

Deep eXtreme Multi-label Learning (XML)



Latest Trend of TC and TS

Extractive Summarization Using a Neural Attention Model

- A neural model, CRSum, to take a sentence's contextual relations with its surrounding sentences into consideration for extractive summarization

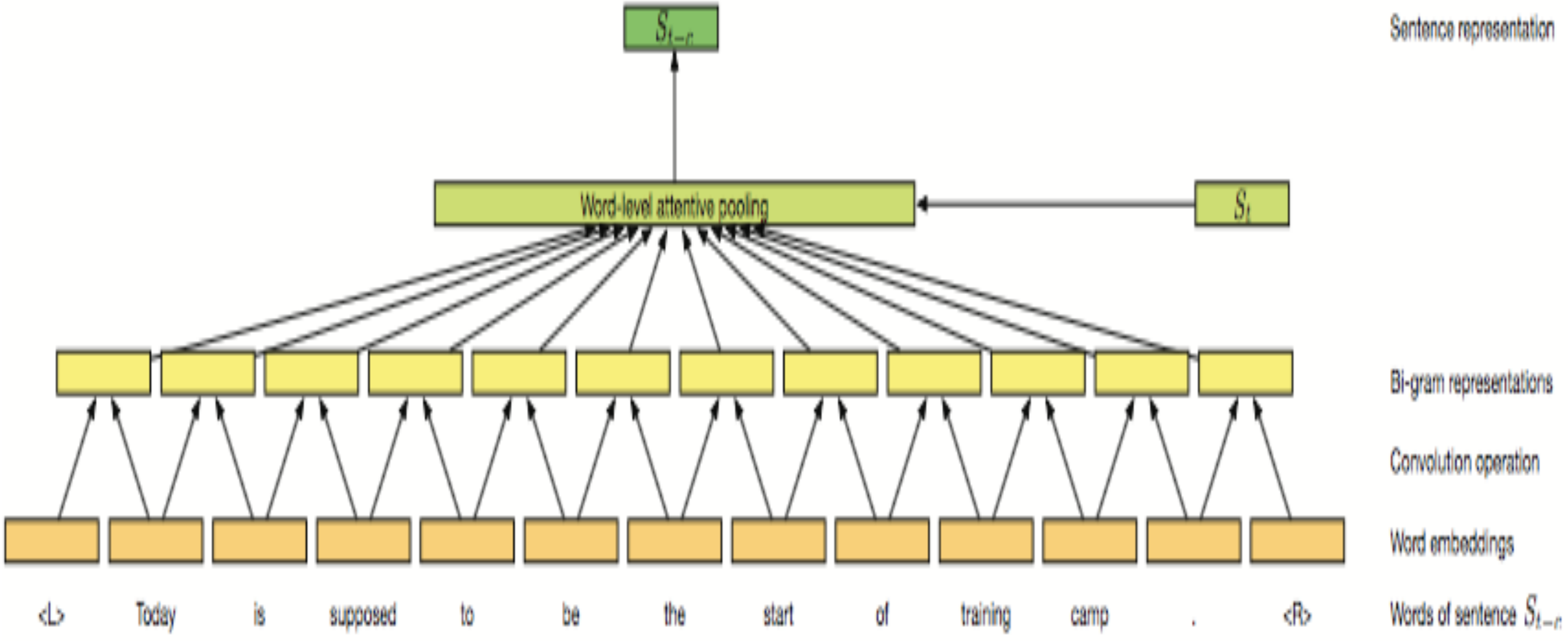
- Contextual relations with a two-level attention mechanism in CRSum

- CNN (Convolutional Neural Network), RNN (Rucurrent Neural Network) and Attention Mechanism

Pengjie Ren et al. (2017), "Leveraging Contextual Sentence Relations for Extractive Summarization Using a Neural Attention Model," *Sigir 2017*.

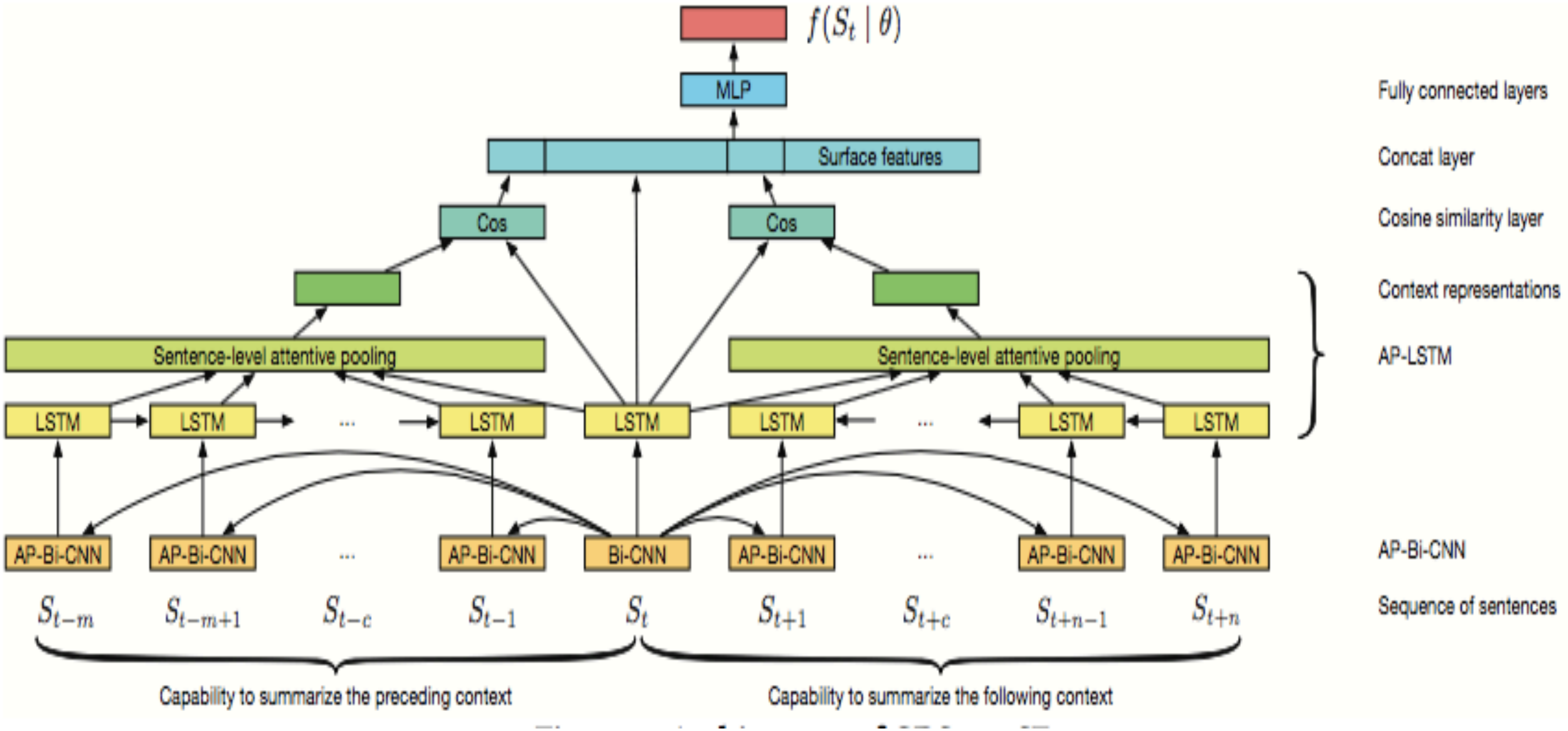
Latest Trend of TC and TS

Extractive Summarization Using a Neural Attention Model



The Latest Trend of TC and TS

Extractive Summarization Using a Neural Attention Model



Thank you for your attention!

고 영 중 (Ko, Youngjoong)

web.donga.ac.kr/yjko