中国民生银行
CHINA MINSHENG BANKING CORP.,LTD.

UNIVERSITY OF LIVERPOOL

A Hybrid Statistical Data Pre-processing Approach for Language-independent Text Classification

Yanbo J. Wang[1], Frans Coenen[2], Robert Sanderson[2]

[1] Information Management Center,
China Minsheng Banking Corporation Ltd., China

[2] Department of Computer Science,
University of Liverpool, UK

The International Conference on Advanced Data Mining And Applications
ADMA 2009
BeiJing,China,August 17-19,2009

Outline

- **Background**
  - Text Classification
  - Textual Data Pre-processing
  - Classification
  - Summary of Background
- **Motivation**
  - Language-independent Text Classification
  - Availability of Language-independent "Bag of Phrases"
  - Summary of Motivation
- **Language-independent Feature Selection**
  - Previous Studies
  - Proposed "Hybrid DIAAF/GSSC" Approach
- **Experimental Results**
  - Text Collections
  - Setting of Experiments
  - Classification Accuracy
- **Conclusions & Future Work**

Background

Text Classification

- What is Text Classification (TC)?
  - TC is the task of assigning one or more predefined categories to natural language text documents, based on their contents.
  - Early studies of TC can be dated back to the early 1960s.
  - Broadly speaking, TC studies can be separated into two divisions: *single-label* vs. *multi-label*.
  - With regard to the single-label TC, three distinct approaches can be identified: *one-class TC*, *binary TC* & *multi-class TC*.
  - **Our study is concerned with the *single-label multi-class* TC.**
  - The overall TC process can be divided into two stages: *textual data pre-processing* & *classification*.

Textual Data Pre-processing

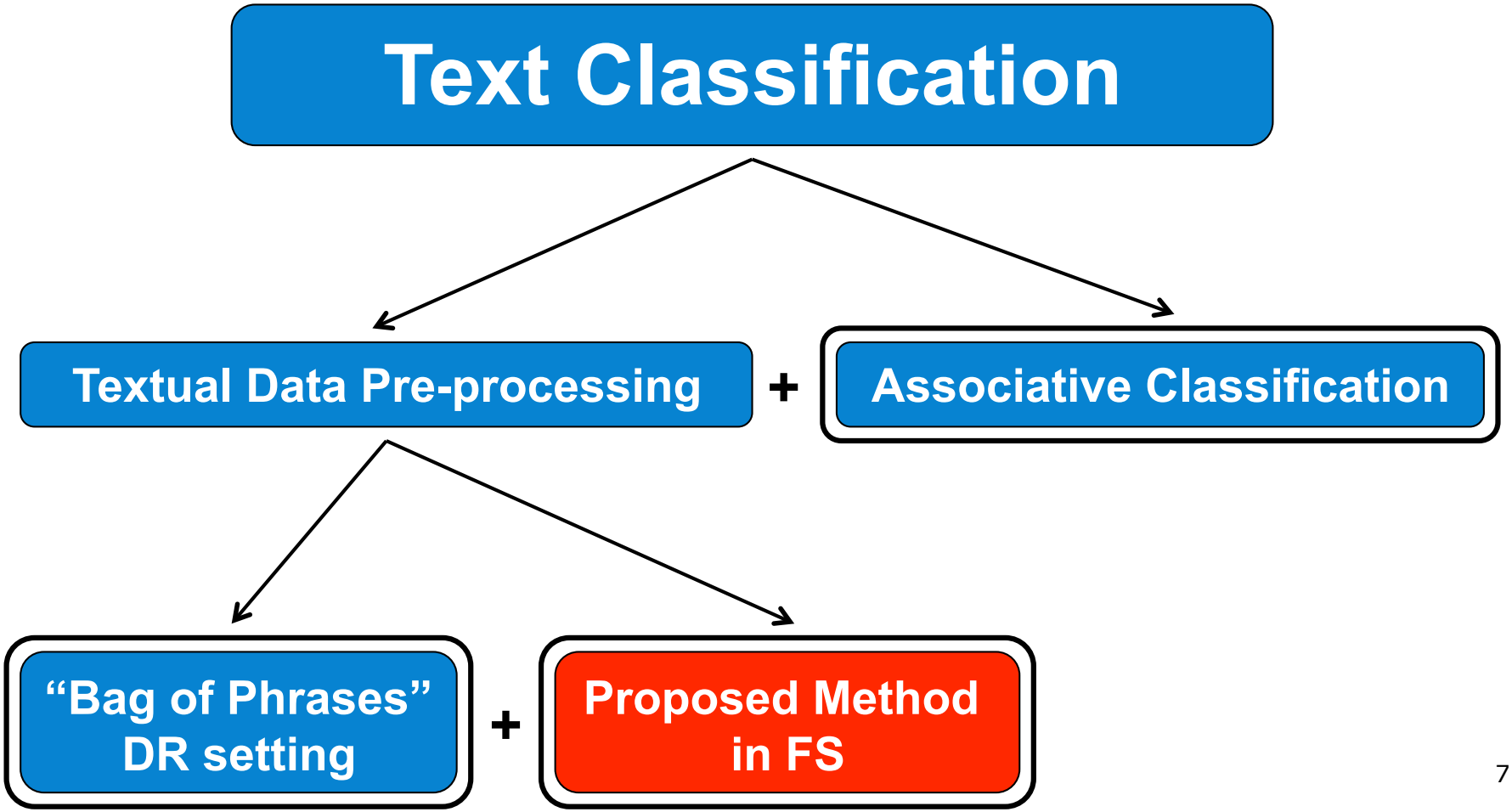
- Textual data pre-processing comprises: *Document-base Representation (DR)* & *Feature Selection (FS)*.
  - **DR** aims to design an application oriented data stucture that precisely interprets a given document-base in an explicit and structured manner.
  - In DR, the *"bag of *"* approach or *vector space model* is considered appropriate for many text mining applications, especially when dealing with TC problems.
  - The *"*"* sign stands for the type of text-units, i.e. words, word-sets, phrases, concepts, etc.
  - **In our study, we select to use the *"bag of phrases"* DR setting.**
  - **FS** aims to identify the most significant text-features (i.e. *key* words) in the document-base that can be used to generate *key* text-units, based on DR.
  - **In this study, we aim to develop an improved *statistical* FS method.**

Classification

- Mechanisms on which classification algorithms have been based include: *decision trees*, *naive bayes*, *k-nearest neighbour*, *support vector machine*, *association rules*, *genetic algorithm*, *neural networks*, etc.
- Previous studies indicate that in many cases classification based on *association rules* (i.e. ***associative classification***) offers greater classification accuracy than other classification approaches.
- In the past decade, associative classfication has been proposed for application in TC with the following advantages:
  - An associative text classifier is fast during both training and categorisation phases, especially when handling large document-bases; and
  - Such text classifiers can be read, understood and modified by humans, so that users are able to see why the classification predictions have been made.
- **Thus, associative classification approach is adopted in our study.**

Summary of Background

In our study:



Motivation

Language-independent TC

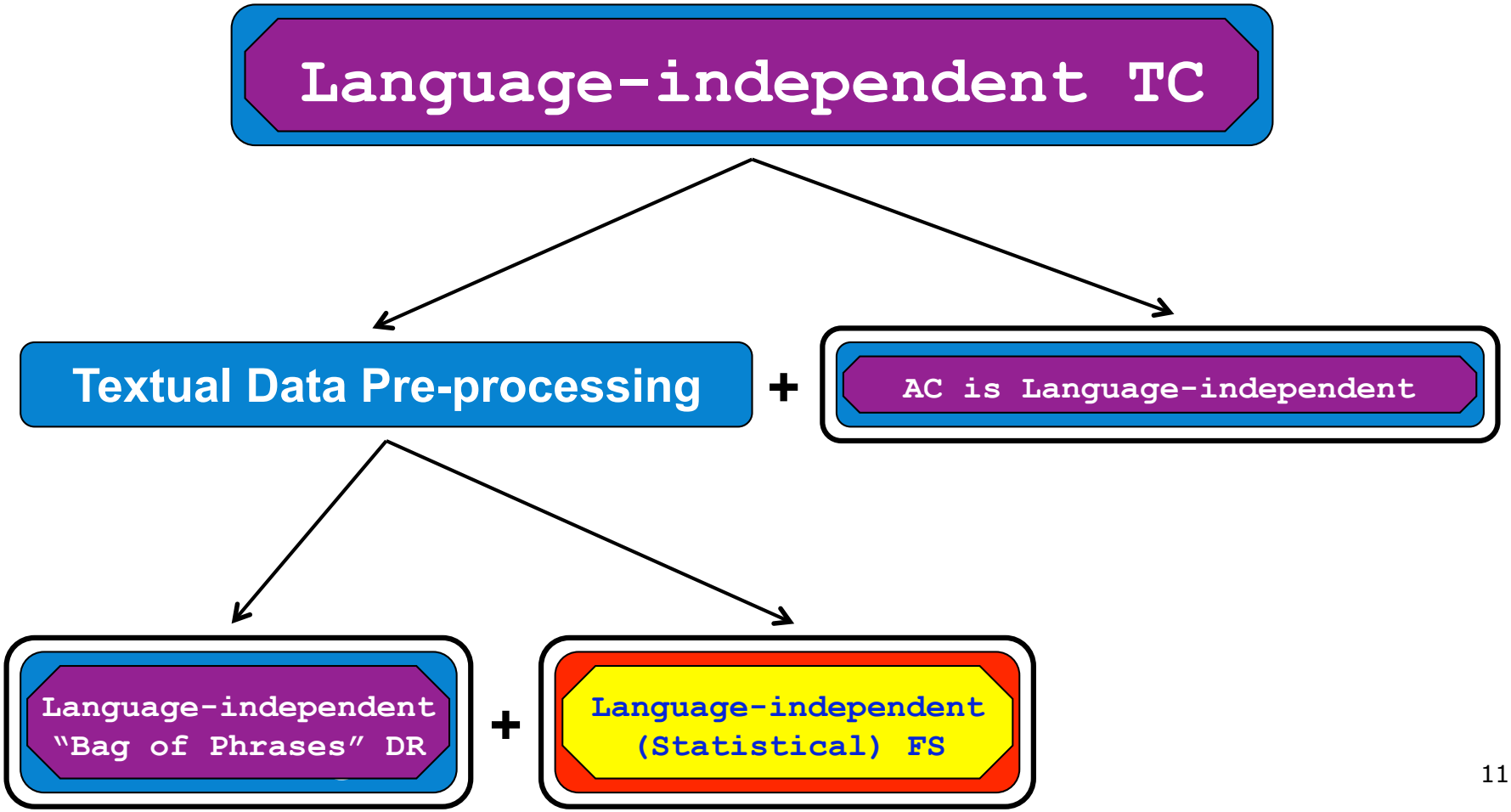
- Many textual data pre-processing mechanisms use language-dependent ideas to identify *key* words and phrases (e.g. *stop word lists*, *synonym lists*, *stemming*, *part-of-speech tagging*, *word sense disambiguation*, etc).
- These techniques operate well but are designed with particular target languages in mind.
- They are therefore not generally applicable to all languages (e.g. *Chinese*, *Arabic*, *Spanish*, etc).
- **We are interested in *language-independent TC*, which aims to address above disadvantage.**
- Such text classifier can also be applied to *cross-lingual*, *multi-lingual* and/or *unknown lingual* textual data collections.

Availability of Language-independent "Bag of Phrases"

- Some definitions
  - **Words:** Words in a document-base are defined as *continuous sequences of alphabetic characters* delimited by non-alphabetic characters.
  - **Noise Words (N):** *Common* and *rare* words are collectively defined to be *noise* words in a document-base.
  - **Potential Significant Words:** A potential significant word, also referred to as a *key* word/ feature, is a non-noise identified in the ***Feature Selection*** stage.
  - **Significant Words (G):** *The first K words* (i.e. the first *k* words for each predefined class) that are selected from the ordered list of potential significant words are defined to be significant words.
  - **Ordinary Words (O):** Other non-noise words that have not been selected as significant words.
  - **Stop Marks (S):** Not actual words but six key punctuations marks (, . : ; ! *and* ?). All other non-alphabetic characters are ignored.
- Language-independent "Bag of Phrases" Generation
  - This approach is named as *DelSNcontGO*: **phrases are *Del*imited by stop marks (*S*) or noise words (*N*), and (as phrase *cont*ents) made up of sequences of one or more significant words (*G*) and ordinary words (*O*); sequences of ordinary words delimited by stop marks or noise words that do not include at least one significant word (in the contents) are ignored.**

Summary of Motivation

In our study:



Language-independent Feature Selection

Previous Studies

- Previous language-independent (statistical) FS mechanisms are described as follows. Each is applied to calculate how significantly a word/feature (u_h) determines a predefined text-category (C_i) in document-base (D_R).

Name	Probabilistic Form	Calculation	Description
DIA (Darmstadt Indexing Approach) Association Factor (**DIAAF**)	$\text{diaaf\_score}(u_h, C_i) = \mathbf{P}(C_i \mid u_h)$	$\frac{\text{count}(u_h \in C_i)}{\text{count}(u_h \in D_R)}$	This score expresses the proportion of the word's occurrence in the given class divided by the word's document-base occurrence.
Galavotti·Sebastiani·Simi Coefficient (**GSSC**)	$\text{gssc\_score}(u_h, C_i) = \mathbf{P}(u_h, C_i) \times \mathbf{P}(\neg u_h, \neg C_i) - \mathbf{P}(u_h, \neg C_i) \times \mathbf{P}(\neg u_h, C_i)$	$\text{count}(u_h \in C_i) \times \text{count}(\neg u_h \in (D_R - C_i)) - \text{count}(u_h \in (D_R - C_i)) \times \text{count}(\neg u_h \in C_i)$	This score expresses the subtraction of two multiplications: (i) the word's occurrence in the given class multiplied by the word's non-occurrence in the complement of the class; and (ii) the word's occurrence in the complement of the class multiplied by the word's non-occurrence in the given class.
Mutual Information (**MI**)	$\text{mi\_score}(u_h, C_i) = \log(\mathbf{P}(u_h \mid C_i) / \mathbf{P}(u_h))$	$\log\left[\frac{\frac{\text{count}(u_h \in C_i)}{\lvert C_i \rvert}}{\frac{\text{count}(u_h \in D_R)}{\lvert D_R \rvert}}\right]$	This score expresses the proportion (in a logarithmic term) of the frequency with which the word occurs in documents of the given class divided by the word's document-base frequency.
Relevancy Score (**RS**)	$\text{rs\_score}(u_h, C_i) = \log((\mathbf{P}(u_h \mid C_i) + d) / (\mathbf{P}(u_h \mid \neg C_i) + d))$	$\log\left[\frac{\frac{\text{count}(u_h \in C_i)}{\lvert C_i \rvert} + d}{\frac{\text{count}(u_h \in (D_R - C_i))}{\lvert D_R - C_i \rvert} + d}\right]$ where d is a constant damping factor	This score expresses the proportion (in a logarithmic term) of the frequency with which the word occurs in documents of the given class divided by the word's frequency in the complement of the class.

Hybrid DIAAF/GSSC

- In this study, we propose a hybrid statistical FS approach that integrates the DIAAF and the GSSC mechanisms, namely "Hybrid DIAAF/GSSC".
- The intuition of the "Hybrid DIAAF/GSSC" approach is:
  - The score tends to be high if the ratio of the class based word count to the document-base word count is high;
  - The score tends to be high if the ratio of the class-complement based word count of non-appearance to the document-base word count of non-appearance is high;
  - The score tends to be high if the ratio of the class-complement based word count to the document-base word count is low; and
  - The score tends to be high if the ratio of the class based word count of non-appearance to the document-base word count of non-appearance is low.
- The calculation of this proposed approach can be shown as follows.

Hybrid DIA Association Factor based Galavotti•Sebastiani•Simi Coefficient (**DIAAF-GSSC**)	$\text{diaaf-gssc\_score}(u_h, C_i) = \mathbf{P}(C_i \mid u_h) \times \mathbf{P}(\neg C_i \mid \neg u_h) - \mathbf{P}(\neg C_i \mid u_h) \times \mathbf{P}(C_i \mid \neg u_h)$	$\left[\frac{\text{count}(u_h \in C_i)}{\text{count}(u_h \in D_R)} \times \frac{\text{count}(\neg u_h \in (D_R - C_i))}{\text{count}(\neg u_h \in D_R)}\right] - \left[\frac{\text{count}(u_h \in (D_R - C_i))}{\text{count}(u_h \in D_R)} \times \frac{\text{count}(\neg u_h \in C_i)}{\text{count}(\neg u_h \in D_R)}\right]$

Experimental Results

Text Collections

- We evaluate the proposed "Hybrid DIAAF/GSSC" approach with respect to the accuracy of classification, using three well-known text collections:
  - Usenet Articles (20 Newsgroups);
  - Reuters-21578; and
  - MedLine-OHSUMED.
- In our experiments, four individual document-bases (textual datasets) are prepared/extracted from above text collections.
  - **20NG.D10000.C10:** This document-base randomly picks up 10 groups of documents (resulting *10,000 documents in 10 classes*) from the 20 Newsgroups collection.
  - **20NG.9997.C10:** This document-base comprises the rest part of the 20 Newsgroups collection (having *9,997 documents in 10 classes*).
  - **Reuters.D6643.C8:** We first of all select the top-10 populous classes from Reuters-21578. Then we remove those multi-labelled and/or non-text documents from each class. As a consequence, 2 of the 10 classes contain no more documents. So, the Reuters.D6643.C8 document-base comprises *6,643 documents in 8 classes*.
  - **OHSUMED.D6855.C10:** First, we select the top-100 most populous classes from this collection. We then simply pick up 10 target-classes from these 100 classes by hand, so as to exclude obvious super-and-sub class-relationships. Finally, we remove such documents either multi-labelled or without a proper text-content from each target-class. The document-base created here comprises *6,855 documents in 10 classes*.

Setting of Experiments

- The experiments are run on a 1.87 GHz Intel(R) Core(TM)2 CPU with 2.00 GB of RAM running under the Windows Command Processor.
- Our experiments are designed to evaluate the proposed "Hybrid DIAAF/GSSC" FS approach, in comparison with previous mechanisms (i.e. DIAAF, GSSC, MI and RS), with regard to the ***DelSNcontGO*** (language-independent) "bag of phrases" DR setting.
- All evaluations described there are conducted using the ***TFPC (Total From Partial Classification)*** associative classifier although any other associative classifier generator could equally well have been used. (Note: the TFPC software can be downloaded from **http://www.csc.liv.ac.uk/~frans/KDD/Software/Apriori-TFPC/aprioriTFPC.html**)
- Accuracy figures, describing the proportion of correctly classified "unseen" documents, are decided using the ***Ten-fold Cross Validation (TCV)***.
- A ***support threshold value*** of 0.1% (for AC approach) is used.
- A ***confidence threshold value*** of 50% (for AC approach) is used.
- A ***lower noise threshold*** value of 0.2% (for too *rare* words) is used.
- A ***upper noise threshold*** value of 20% (for too *common* words) is used.
- The parameter *K* (***maximum number of selected final key features***) is chosen to be 1,000. (Note: the value of *K* is changed to be 900 for OHSUMED.D6855.C10 document-base since 1,000 *key* features will cause more than 2^{15} *key* phrases to be generated; and, for reasons of computational efficiency, the TFPC associative classifier limits the total number of identified attributes (significant phrases) to 2^{15}.

Classification Accuracy

	DIAAF	GSSC	MI	RS	Hybrid DIAAF/GSSC
20NG.D10000.C10	76.36	0	76.36	76.36	76.43
20NG.D9997.C10	81.45	0	81.45	81.45	81.62
Reuters.D6643.C8	87.57	0	87.79	87.79	88.23
OHSUMED.D6855.C10	78.83	0	79.53	79.64	79.74
Average Accuracy	81.05	0	81.28	81.31	81.51
# of Best Accuracies	0	0	0	0	4

- The column of GSSC is shown with value '0' for all the records. The reason of this is that when applying the GSSC FS technique, with the TFPC associative text classifier, too many rules will be generated thus causing computational difficulty and consequently no results are obtained.

Classification Accuracy (continue...)

- The number of instances of best classification accuracies obtained throughout the 4 document-bases can be ranked in order as follows.

1	Hybrid DIAAF/GSSC	All cases
2	DIAAF	None of any case
2	GSSC	None of any case
2	MI	None of any case
2	RS	None of any case

- The average accuracy of classification throughout the 4 document-bases can be ranked in order as follows.

1	Hybrid DIAAF/GSSC	81.51
2	RS	81.31
3	MI	81.28
4	DIAAF	81.05
5	GSSC	0

- These show the good performance and the stability of the good performance for "Hybrid DIAAF/GSSC".

Conclusions & Future Work

Conclusions & Future Work

- In this study, we introduce a new language-independent FS technique, namely "Hybrid DIAAF/GSSC", which integrates the ideas of DIAAF and GSSC.
- From the experimental results, it can be seen that the proposed "Hybrid DIAAF/GSSC" approach outperforms existing mechanisms regarding the *DelSNcontGO* language-independent "bag of phrases" DR setting and the TFPC associative text classifier.
- Our study in turn improves the performance of language-independent TC.
- The results presented in this study corroborate that the traditional TC problem can be solved, with good classification accuracy, in a language-independent manner.
- Further research is suggested to identify the improved language-independent textual data pre-processing approach, and improve the performance of language-independent TC.

The End

Thank You!