基于类别层次结构的多层文本分类样本扩展策略

Expanding Training Dataset with Class Hierarchy in Hierarchical Text Categorization

李保利（Baoli LI）

河南工业大学

Henan University of Technology

Outline

- Introduction:
  - Hierarchical Text Classification
  - NLPCC-2014 Shared Task: Large Scale Chinese News Categorization
  - Our focus in this shared task
- Strategies for building our NLPCC-2014 system
  - Flat Classification Based Solution
  - Expanding Training Dataset with Class Hierarchy
- Experiments and Discussion
- Conclusions and Future Work

Introduction

- Hierarchical Text Classification:
  - Thousands of Classes organized into a class hierarchy
  - Very Challenging: LSHTC 1-4, exact accuracy < 50%
  - Approaches:
    - Local classifier approaches
      - A series of classifiers
      - Top-down, along the class hierarchy
      - At each node: Binary classification or multi-class classification
    - Global classification approaches:
      - A single classifier
      - A special one: Flat approach
  - Hard to get a reasonable training data: size and distribution

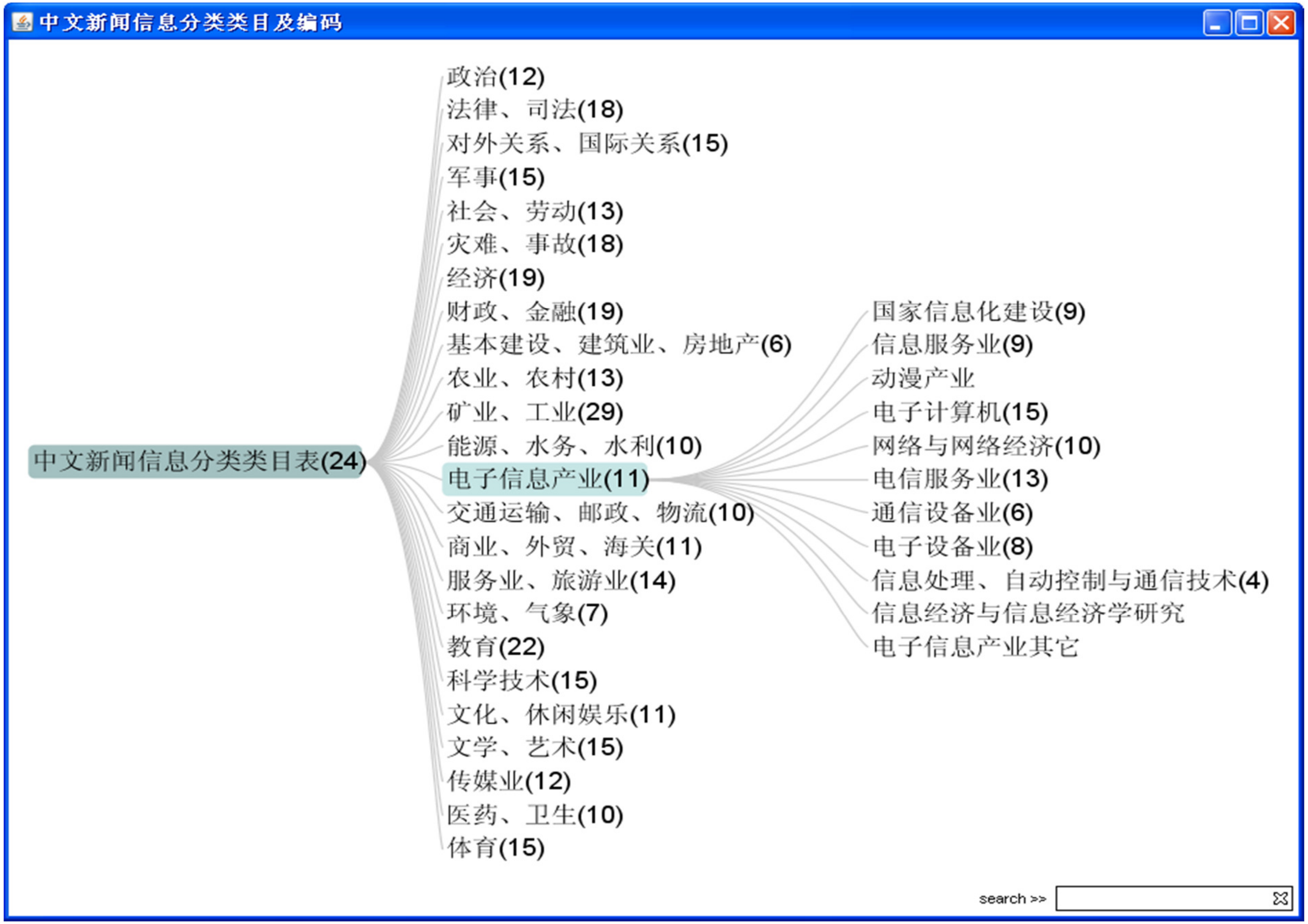
Introduction

- NLPCC-2014 Shared Task: Large Scale Chinese News Categorization
  - 1st large scale open evaluation on Hierarchical Chinese Text Classification
  - the Classification and Code of News in Chinese (CCNC)
    - 5 levels, 6200+ categories
    - Level one 24 categories, Level two 340.
  - Main Characteristics

Introduction

- NLPCC-2014 Shared Task: Large Scale Chinese News Categorization
  - Main Characteristics:
    - Consider only 247 classes on the second level
    - Single label
    - Closed or open: not specified
    - Consistent distribution across training and test data (unknown before submission)
    - Evaluation metrics: macro mean of precision, recall, F1 on the first and second level

Class Hierarchy of Chinese News

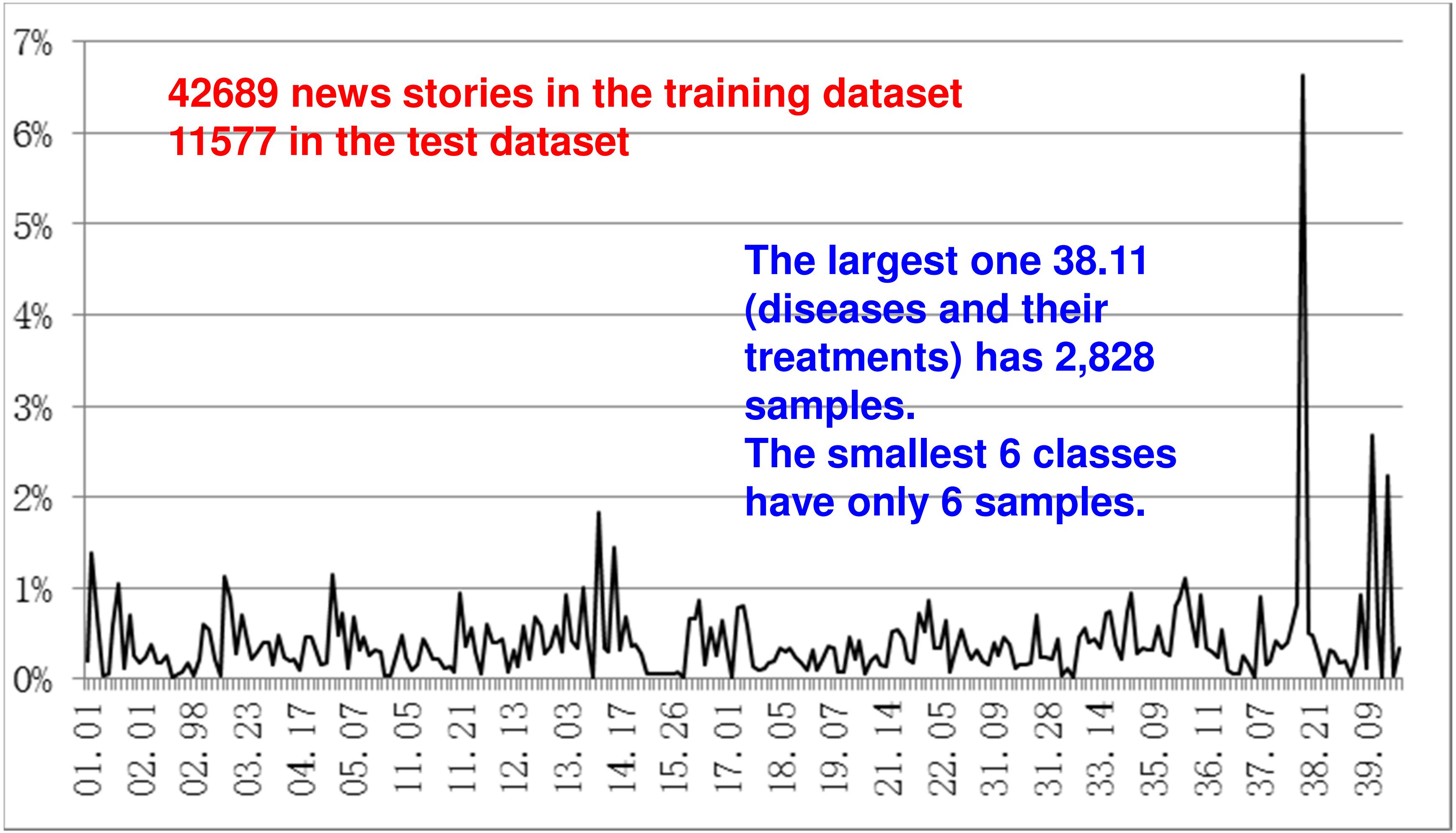


Given Information of each class

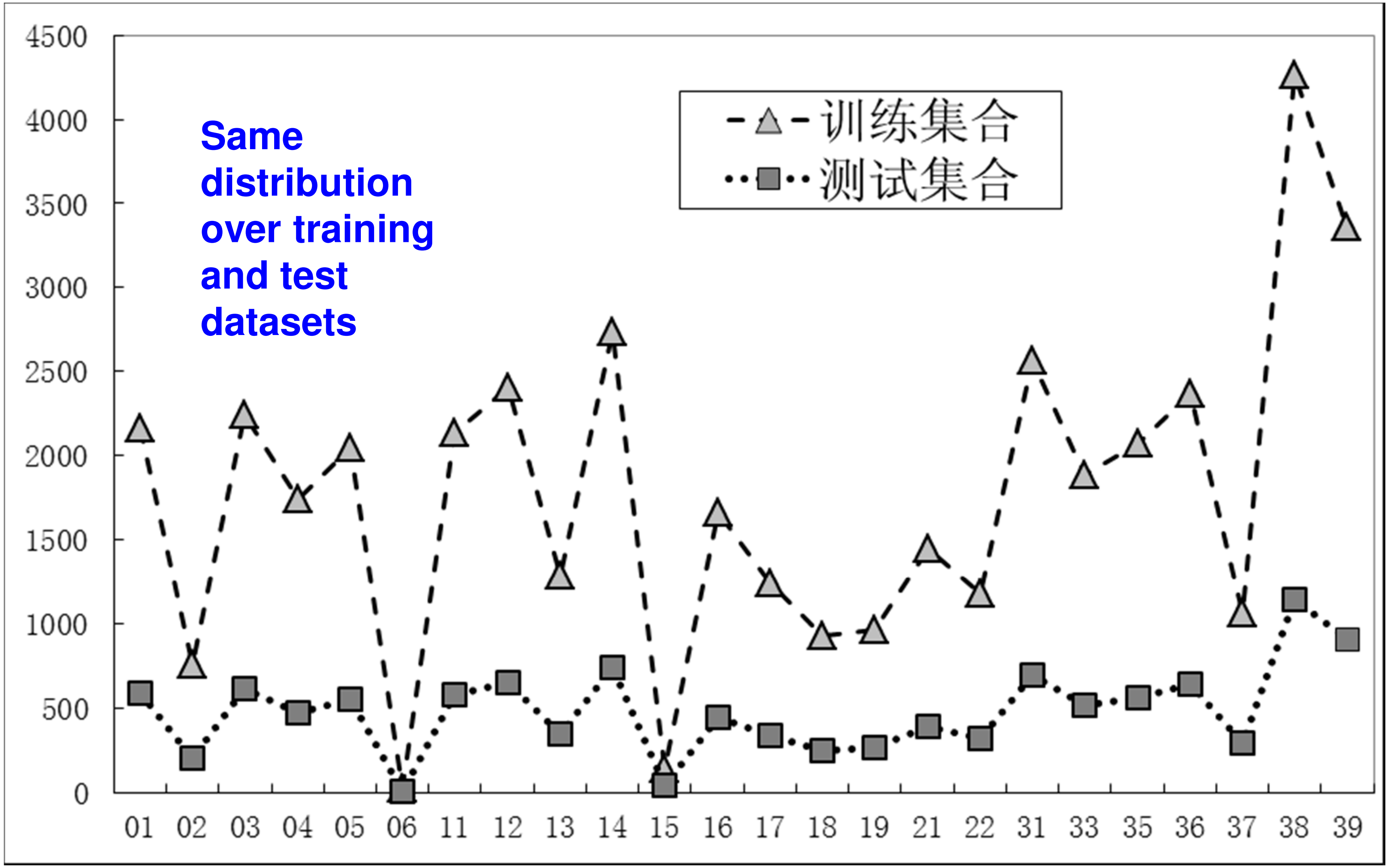
表 3　中文新闻信息分类详表

代　码	类 目 名 称	说　明
01 Class code	**政治** Class name	Class description
01．01	国家概况、地区概况	
01．01．01	政治体制	简称政体。政治体制改革、党政关系改革入此
01．01．03	国庆、区庆	国庆黄金周旅游报道同时入 21．51 旅游业
01．01．05	首都、首府	宜同时入 33．19．12．08．23 城市地理
01．01．07	国家、地区标志	国旗、国徽、国歌、国花、国树、国石、国鸟，以及省、市、区的旗、徽、花、树、城标雕塑、城市精神等（包括评选活动）入此
01．01．09	国力、竞争力	综合国力、国家硬实力和软实力、国际竞争力（全球竞争力）、经济竞争力、科技竞争力、城市竞争力、竞争力排名等入此

Sample Distribution of level 2 classes in the training dataset



Sample Distribution of level 1 classes



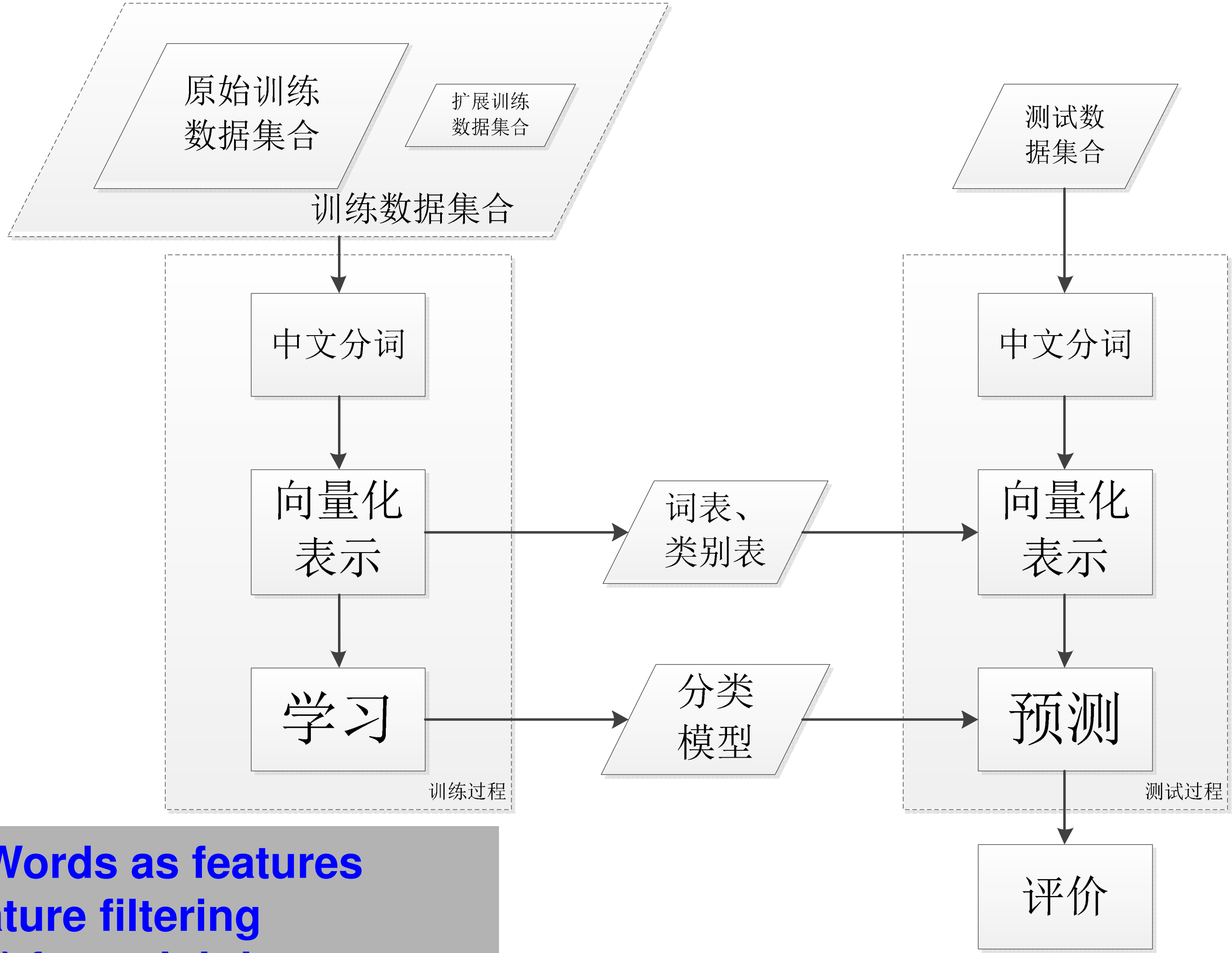
Our Goals in Participating in this shared task

1. Gain knowledge and experience to deal with this challenging problem
2. Explore to study some key issues in hierarchical text classification
   - Focus on training data expansion

Key Strategies for Building our System

- **Flat Classification approach**
  - To gain global optimization
  - Each story can be assigned to one class of the 340 second level categories
- **Expanding training data with the class hierarchy**
  - To classify stories into those classes without any samples (Macro-Average Metrics)
  - Closed: not use external resources, e.g. search engines
  - The hierarchy does have some useful information.

Flow Chart of our System

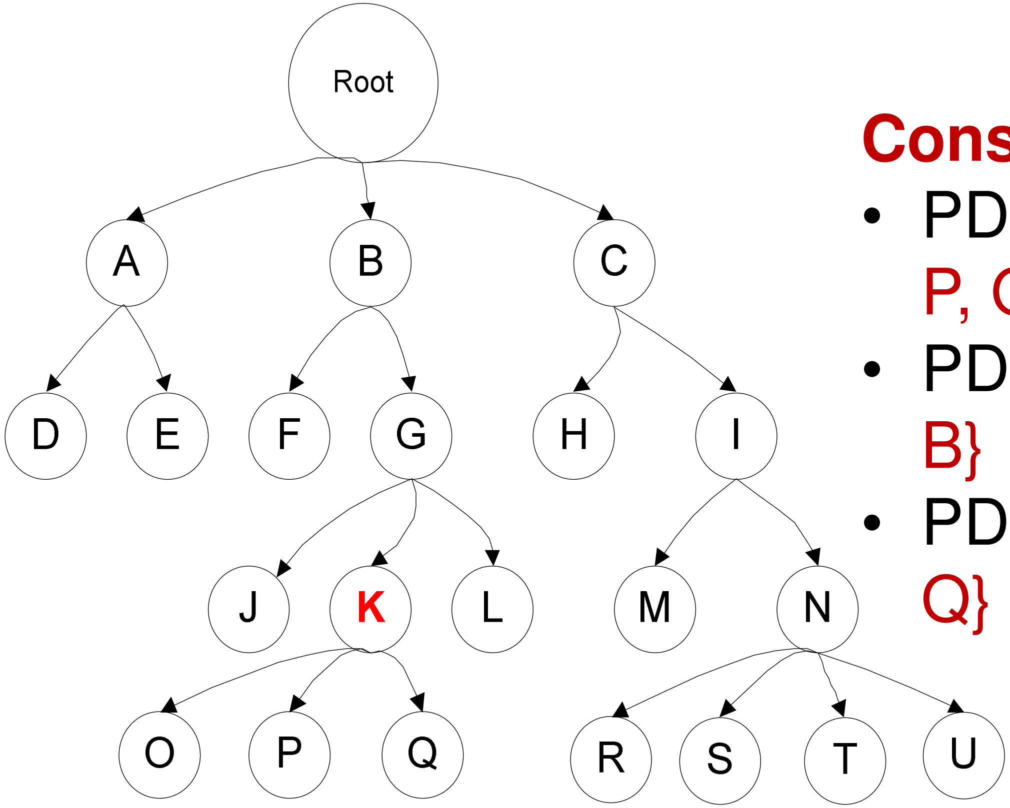


- Chinese Words as features
- DF for feature filtering
- TFIDF (ltc) for weighting

Strategies for expanding the training dataset

- Generate a pseudo sample from a class itself: class name + its short description
- Further, generate a pseudo sample based on class hierarchy:
  - Connotation-based: include classes' names and their descriptions of those direct ancestor classes (those on the path from root to it) (PDT_ANCESTOR)
  - Extension-based: include classes' names and their descriptions of its all offspring classes (PDT_OFFSPRING)
  - Connotation and extension based: combine the above two (PDT_ALL)

An example



Consider the class K:

- PDT_OFFSPRING: {K, O, P, Q}
- PDT_ANCESTOR: {K, G, B}
- PDT_ALL: {B, G, K, O, P, Q}

Strategies for expanding the training dataset

- Other variants:
  - When generating pseudo samples, only consider classes of level 2 and 3 dependent on the task itself (Localized version)
    - PDT_ANCESTOR_V1
    - **PDT_OFFSPRING_V1**
    - PDT_ALL_V1
  - Name only
  - Give different weights to class name and its description
  - …

Experiments and Discussion

- Comparison of Different Classification Algorithms
- Comparison of Different Pseudo Sample Generation Strategies
- NLPCC-2014 LSCNC Official Results
- Comparison of Flat approach and Top-down approach

Experiments and Discussion

- Comparison of different classification algorithms

Algorithm	Level 1				Level 2			
	MacroP	MacroR	MacroF1	Acc.	MacroP	MacroR	MacroF1	Acc.
CB	0.7705	0.7741	0.7723	0.7995	0.6565	0.6000	0.6270	0.6782
NBB	0.3234	0.0805	0.1289	0.1626	0.0531	0.0122	0.0198	0.1047
NBM	0.7058	0.5274	0.6037	0.6375	0.4546	0.2365	0.3112	0.4677
kNN(k=60)	0.7635	0.7664	0.7649	0.8025	0.6172	0.6106	0.6139	0.6901
SVM	0.8323	0.7468	0.7873	0.8087	0.6947	0.5439	0.6101	0.7039
LINEAR	0.8532	0.8256	0.8392	0.8586	0.7503	0.6616	0.7032	0.7656

Experiments and Discussion

- Comparison of different pseudo sample generation strategies

- use pseudo samples only for training, test dataset for testing
- No feature filtering

Method	Level 1				Level 2			
	MacroP	MacroR	MacroF1	Accuracy	MacroP	MacroR	MacroF1	Accuracy
PDT_OFFSPRING	0.5416	0.5185	0.5298	0.5552	0.3782	0.2988	0.3338	0.3257
PDT_OFFSPRING_V1	0.5673	0.5474	0.5572	0.5692	0.4214	0.3136	0.3596	0.3491
PDT_ANCESTOR	0.5163	0.5023	0.5092	0.5026	0.3725	0.3082	0.3373	0.2964
PDT_ANCESTOR_V1	0.4955	0.4656	0.4800	0.4732	0.3487	0.2888	0.3159	0.2592
PDT_ALL	0.5360	0.5227	0.5293	0.5225	0.3884	0.3210	0.3515	0.3161
PDT_ALL_V1	0.5433	0.5336	0.5384	0.5473	0.4063	0.3202	0.3582	0.3350

Experiments and Discussion

- Official Results

Rank	System #	Level 1				Level 2			
		MacroP	MacroR	MacroF1	Accuracy	MacroP	MacroR	MacroF1	Accuracy
1	9	0.8725	0.8633	0.8679	0.8848	0.7772	0.7726	0.7749	0.8161
2	2	0.8513	0.8315	0.8413	0.8604	0.7487	0.6822	0.7139	0.7720
3	10	0.7422	0.7770	0.7592	0.7904	0.5646	0.6238	0.5927	0.6294
4	5	0.7336	0.7076	0.7204	0.7507	0.6024	0.5240	0.5604	0.6249
5	4	0.7260	0.7023	0.7140	0.7450	0.5922	0.5203	0.5539	0.6185
6	8	0.6536	0.6428	0.6481	0.7197	0.5073	0.4711	0.4885	0.5874
7	6	0.5817	0.4576	0.5123	0.5363	0.4577	0.2430	0.3174	0.3658
8	3	0.7389	0.6616	0.6981	0.7339	0.1352	0.1336	0.1344	0.1664
9	1	0.3758	0.2453	0.2969	0.2856	0.0761	0.0867	0.0892	0.0761
10	7*	0	0	0	0	0	0	0	0

Experiments and Discussion

- Comparison of flat and top-down approach

Top-down approach

Algorithm	Level 1				Level 2			
	MacroP	MacroR	MacroF1	Accuracy	MacroP	MacroR	MacroF1	Accuracy
CB	0.7712	0.7740	0.7726	0.7994	0.6569	0.5994	0.6282	0.6776
LINEAR	0.8534	0.8256	0.8393	0.8587	0.7515	0.6616	0.7037	0.7657

Flat approach

Algorithm	Level 1					Level 2		
	MacroP	MacroR	MacroF1	Acc.	MacroP	MacroR	MacroF1	Acc.
CB	0.7705	0.7741	0.7723	0.7995	0.6565	0.6000	0.6270	0.6782
LINEAR	0.8532	0.8256	0.8392	0.8586	0.7503	0.6616	0.7032	0.7656

Conclusions and Future Work

- Class hierarchy can be used to derive some new pseudo training samples, and these pseudo samples can help to improve system's performance.
- Among the proposed strategies, the localized expansion strategy based on class extensions performs better.
- NLPCC-2014 LSCNC shared task actually is not a typical hierarchical classification problem.

Conclusions and Future Work

- Explore other strategies to expand the training dataset: e.g. give different weights for class name and class description, remove noises in the descriptions;
- Explore how to build an ideal training dataset: size
- with the datasets, explore other possible hierarchical text classification algorithms

Thanks for your attention!

Questions & Discussion