Large Scale Chinese News Categorization

--based on Improved Feature Selection Method

Peng Wang

Joint work with H. Zhang, B. Xu, H.W. Hao

Computational-Brain Research Center
Institute of Automation, Chinese Academy of Sciences

CASIA

Outline

- **Introduction**
- **Our Framework**
  - Preprocessing
  - Feature selection
  - Machine learning methods
  - Measurements for evaluation
- **Experiments**
- **Conclusions**

Outline

- Introduction
- Our Framework
  - Preprocessing
  - Feature selection
  - Machine learning methods
  - Measurements for evaluation
- Experiments
- Conclusions

Introduction

- Task Definition
  - given a news document and a predefined hierarchy of categories with a depth of 2.
  - the Classification and Code of News in Chinese (CCNC) as the predefined hierarchy of categories.
    - Some samples from CCNC:

01 政治	02 法制	03 外交・国际关系
01001 国家（地区）概况	02001 法制建设	03001 外交政策
01002 国家元首	02002 法学研究	03002 对外关系
01003 权力机构	02003 法律服务	03003 外交事务
01004 行政机构	02004 知识产权保护	03004 国际关系
01005 中国政府行政管理	02005 消费者权益保护	03005 国际问题

    - We are required to provide the IDs of the categories which this document belongs to.

Introduction (contd.)

- About categories,
  - This hierarchy of categories consists of at most 2 levels of subdivisions, specifically, which includes 24 main entries and 367 entries in the first and the second levels.
- Text corpus
  - includes about 30,000 news articles.
  - provided by courtesy of the Xinhua News Agency.
  - category annotation in XML format is:

<doc id="1">
<title>博尔特、纳达尔等体坛名将获劳伦斯奖提名</title>
<content >新华网吉隆坡 2 月 26 日体育专电（记者赵博超）经全球媒体提名投票，博尔特、纳达尔、小威廉姆斯、老虎·伍兹等体坛名将获 2014 年劳伦斯世界体育奖提名。其中，博尔特和小威廉姆斯已经赢得过 3 次劳伦斯奖，F1 冠军维特尔是第五次获得该奖的提名，而老虎·伍兹则在 2000 年就获得过首届劳伦斯奖。另外，此次纳达尔和伊辛巴耶娃则在劳伦斯奖下的两个分奖项均获得了提名。...... </content>
<ccnc_cat id ="1">39.14</ccnc_cat>
<ccnc_label id ="1">体育|体育奖</ccnc_label>
</doc>

  - may have more than one category ID;
  - with up to 2 category IDs;
  - Required to sort multiple IDs in descending order with respect to their confidence scores.

Outline

- **Introduction**
- **Our Framework**
  - Preprocessing
  - Feature selection
  - Machine learning methods
  - Measurements for evaluation
- **Experiments**
- **Conclusions**

CASIA

Our Framework

- Our framework based on Feature selection,

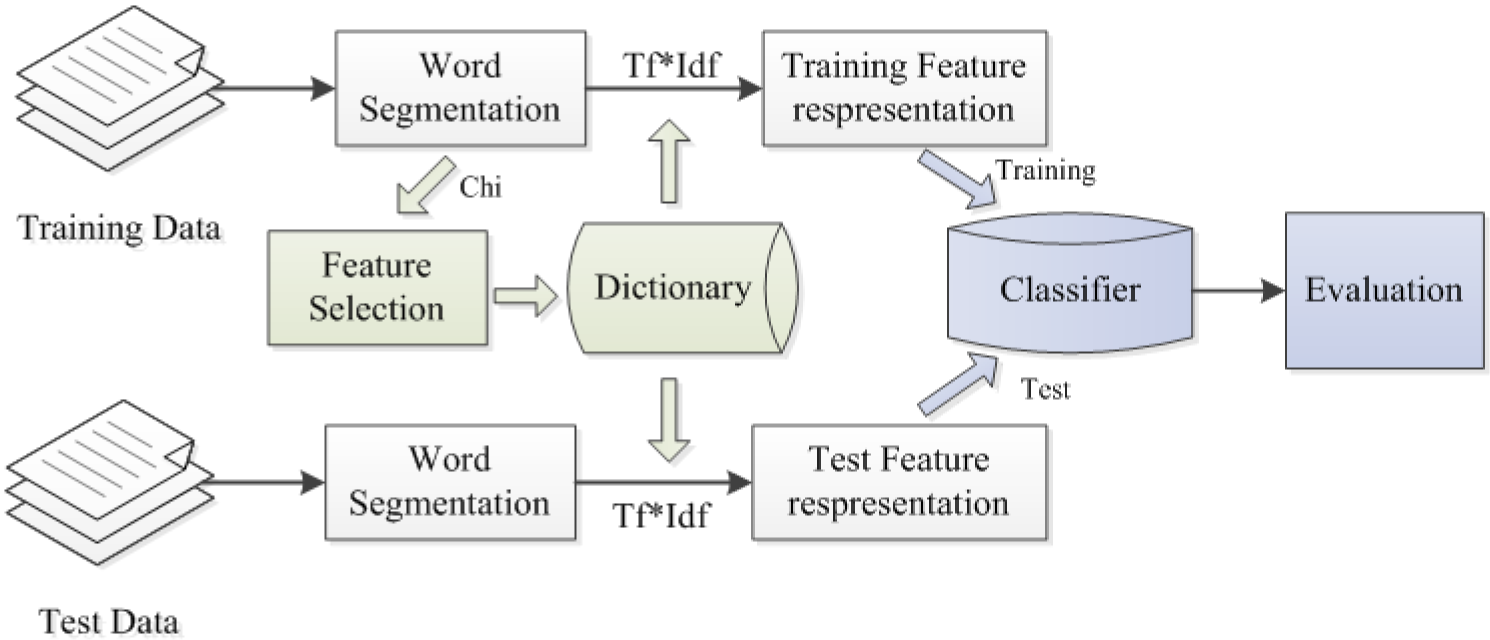

Figure 1. The framework of our method

Our Framework

- Our framework based on Feature selection,

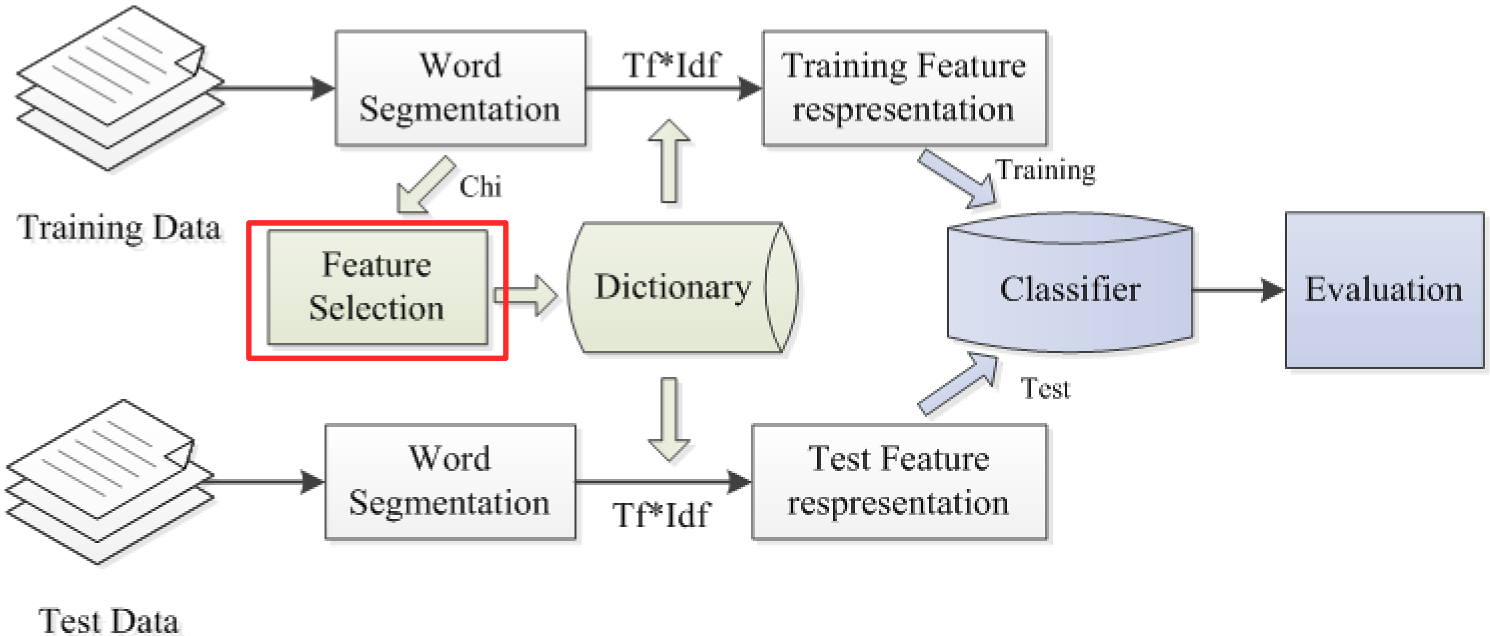

Figure 1. The framework of our method

Our Framework

- Preprocessing,
  - word segmentation or stemming;
  - Removing stop-words,
    - prepositions, conjunctions and pronouns;
    - occur in many documents and hold very high DF scores;
    - Contain little useful information for feature representation.
- Feature selection,
  - Bag of words (BOW) leads to a high dimensional feature space;
  - selects a specific subset of the terms from original feature;
  - remove these irrelevant and redundant words;
  - CHI statistic is employed.

Our Framework

- Feature selection using CHI,
  - each term is assigned with a score according to CHI function;
  - with higher scores are selected;
  - measures the lack of independence between term and the class, defined as Equation (1),

$$\chi^2(t,c) = \frac{N \times (AD - BC)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}, \qquad (1)$$

Table 1 Definitions of notions used in χ^2 statistic

Notations	Definitions
c_i	Label of category i
A	Number of texts that contain the term t and also belong to C_i
B	Number of texts that contain the term t but do not belong to C_i
C	Number of texts that do not contain the term t but belong to C_i
D	Number of texts that neither contain the term t nor belong to C_i
N	Total number of all documents in the training data

Our Framework

- From equ.(1),
  - if term *t* and class *c* are independent, the value of it is zero.
  - Otherwise, the larger indicate that the term *t* is more related to category *c*.
- From Table 1,
  - shortcoming of the CHI is that they just count whether a term and a special category co-occurrence in each document,
  - instead of the frequency.
  - the native score may magnify the contributions of terms with low-frequency in feature representation;
  - propose a measurement of term-goodness for feature selection in Equ.(2),

$$FS(t) = \log(\mathrm{tf}(t)) \sum_{i=1}^{m} p(c_i)\chi^2(t, c_i), \qquad (2)$$

Our Framework

- For construct feature dictionary, how many terms reserved ?

$$l_* = \lfloor L * \sigma \rfloor, \quad (3)$$

  - Where L is total number of terms, σ reserving ratio.

- Advantages
  - reduce the dimensionality;
  - removing noisy features;
  - avoid over-fitting

Our Framework

- Feature weight,
  - Tf-idf
- Machine learning methods,
  - In this task, each text may have more than one category, but the concrete number of category is indeterminate.
  - In this evaluation, we choose softmax regression model to predict a confidence score.
  - Generalized version of logistic regression for probabilistic multiclass problems.

$$hf(x^{(i)},\boldsymbol{\theta})=\begin{pmatrix} p(y^{(i)}=1\mid x^{(i)},\boldsymbol{\theta}) \\ p(y^{(i)}=2\mid x^{(i)},\boldsymbol{\theta}) \\ \vdots \\ p(y^{(i)}=k\mid x^{(i)},\boldsymbol{\theta}) \end{pmatrix}=\frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}\begin{pmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{pmatrix}, \qquad (4)$$

Our Framework

- Estimate the parameter $\boldsymbol{\theta}$,
  - the cost function,

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\left(\sum_{i=1}^{m}\sum_{j=1}^{k} 1(y^{(i)} = j)\log p(y^{(i)} = j \mid x^{(i)}, \boldsymbol{\theta})\right) \qquad (5)$$

Table 2. The steps of parameters estimation for softmax model

Step1. Initialize vector $\boldsymbol{\theta}$ and learning rate $\boldsymbol{\lambda}$;
Step2. Compute $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$, then $\boldsymbol{\theta}^{*} = \boldsymbol{\theta} - \boldsymbol{\lambda}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$;
Step3. If $\left\Vert J(\boldsymbol{\theta}) - J(\boldsymbol{\theta}^{*})\right\Vert < \varepsilon$, go to Step5, otherwise go to Step4;
Step4. Update $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^{*}$, and go to Step2;
Step5. Converge to an optimal solution $\boldsymbol{\theta}^{*}$.

Measurements for evaluation

- Measurements,

$$\text{Precision}_{macro} = \frac{1}{k}\sum_{i=1}^{k} \frac{\text{\# samples whose human label match model's in } c_i}{\text{\# samples labled as } c_i \text{ by model}},$$

$$\text{Recall}_{macro} = \frac{1}{k}\sum_{i=1}^{k} \frac{\text{\# samples whose human lable match model's in } c_i}{\text{\# samples labled as } c_i \text{ by human}},$$

$$\text{F1}_{macro} = \frac{2\text{Precision}_{macro} * \text{Recall}_{macro}}{\text{Precision}_{macro} + \text{Recall}_{macro}},$$

$$\text{Precision}_{micro} = \frac{\text{\# samples whose human label match model's}}{\text{\# all samples}},$$

$$\text{Recall}_{micro} = \text{F1}_{micro} = \text{Precision}_{micro},$$

CASIA

Experiments

- *Experimental data*
  - The Chinese News articles;
  - 20Newsgroup.
- *Experimental results,*
  - The definitions of hierarchical category indicate that the second level category information can deduce that of first level.
  - For concision, we classify the test samples directly at the second level using our framework in this evaluation.

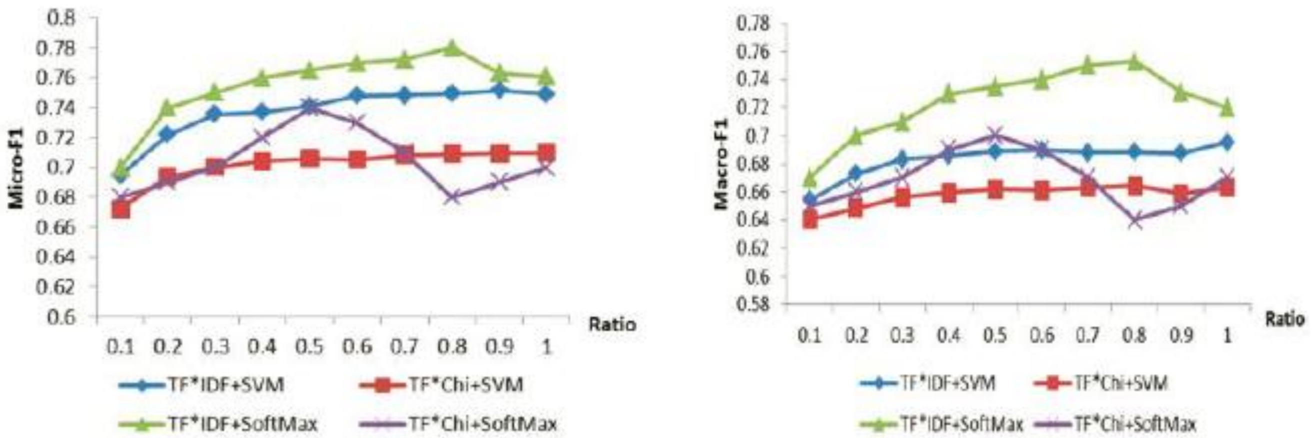

Figure 2. The F-measurement vary with feature ratio on Chinese News articles

Evaluation

❖ Experimental Results on 20NGs,

Table 3. Accuracy vary with topic numbers on 20NG

	SVMs		SoftMax	
ratio	Tf*idf	Tf*chi	Tf*idf	Tf*chi
0.1	0.7386	0.7141	0.7212	0.6807
0.2	0.7827	0.7455	0.7922	0.6450
0.3	0.8014	0.7520	0.8145	0.6931
0.4	0.7967	0.7632	0.8032	0.7027
0.5	0.8162	0.7701	0.8008	0.7215
0.6	0.8356	0.7780	0.8253	0.7360
0.7	0.8378	0.7827	0.8274	0.7372
0.8	0.8204	0.7731	0.8138	0.7451
0.9	0.8317	0.7842	0.8213	0.7543
1.0	0.8367	0.7617	0.8169	0.7574

- From Table 3,
  - the terms weighting method *tf *idf* is more robust than *tf *Chi*.
  - However, the softmax when the category number is small have little merits compared with SVMs.

Thanks For Your Time！