Highly Restricted Keyword Selection Based on Sparse Analysis for Uyghur Text Categorization

Dong Wang, Rayilam Parhat

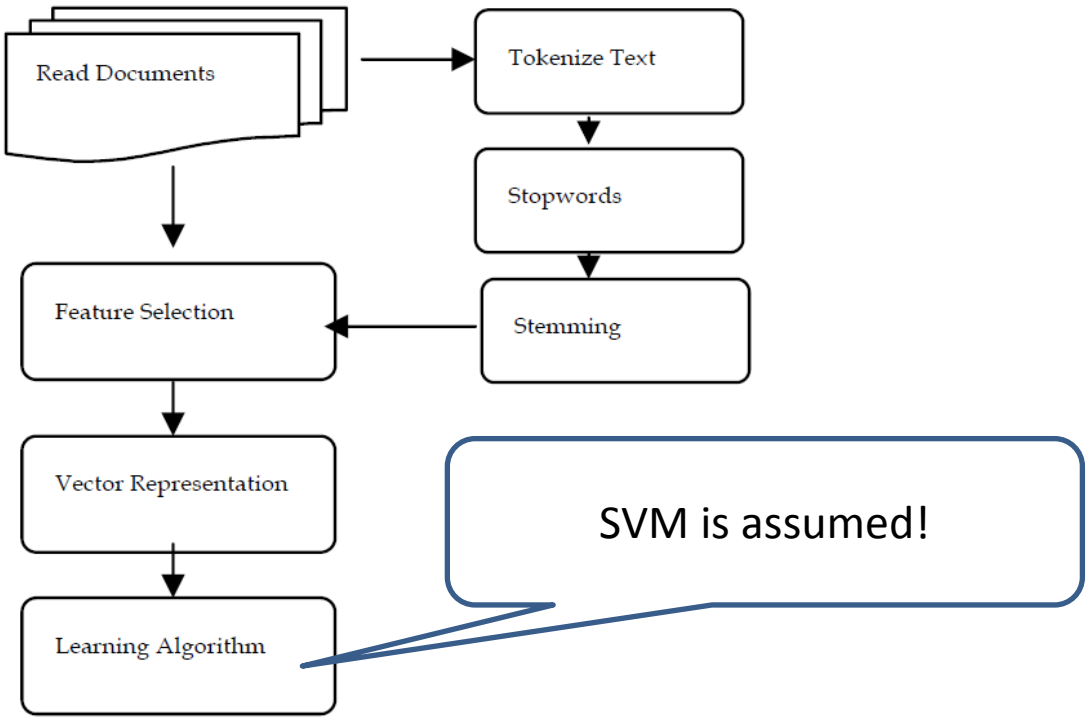
2014/11/17

Contents

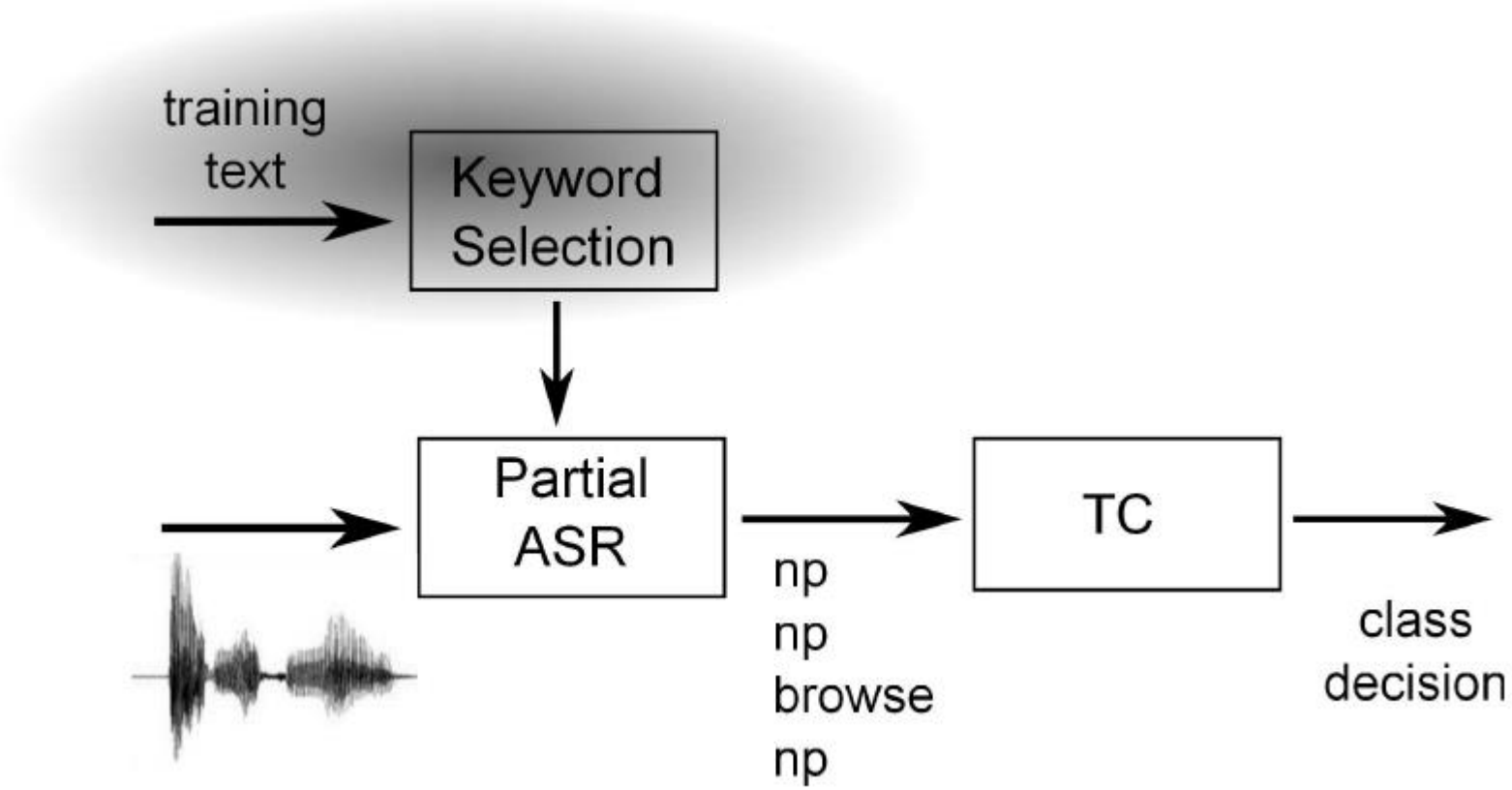
- Highly restricted Text categorization (TC)
- Keyword selection with sparse analysis
- Experiments

Text categorization

- Classify texts to pre-defined categories.
- No category learning.



Highly restricted TC content-aware PTN



Difficulties in highly restricted TC

- We can afford a very small set of keywords
- We need the best keywords to reserve the performance
- We need online decision (not resolved yet)

Keyword selection in TC (1)

- Keyword selection based on intermediate scores: Gini index, information Gain, mutual information, X^2 test, class discriminating measure (CDM), weight of evidence for text, odds ratio, expected cross entropy, DF, TF…

Keyword selection in TC (2)

Gain Ration

$$GR(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(t, c) \log \frac{P(t, c)}{P(t)P(c)}}{-\sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log P(c)}$$

Informational Gain(IG)

$$IG(w) = -\sum_{j=1}^{K} P(c_j) \log P(c_j) + P(w) \sum_{j=1}^{K} P(c_j \mid w) \log P(c_j \mid w) + P(\bar{w}) \sum_{j=1}^{K} P(c_j \mid \bar{w}) \log P(c_j \mid \bar{w})$$

$$= H(samples) - H(samples \mid w)$$

Chi Square

$$\chi^2(f_i, c_j) = \frac{|D| \times (\#(c_j, f_i)\#(\bar{c}_j, \bar{f}_i) - \#(c_j, \bar{f}_i).\#(\bar{c}_j, f_i))^2}{(\#(\#(c_j, f_i) + \#(c_j, \bar{f}_i)) \times (\#(\bar{c}_j, f_i) + \#(\bar{c}_j, \bar{f}_i)) \times ((c_j, f_i) + \#(\bar{c}_j, f_i)) \times (\#(c_j, \bar{f}_i) + \#(\bar{c}_j, \bar{f}_i))}$$

Conditional mutual Information

$$CMI(C \mid S) = H(C) - H(C \mid S_1, S_2, ..., S_n)$$

Document Frequency(DF)

$$DF(t_k) = P(t_k)$$

Term Frequency(TF)

$$tf(f_i, d_j) = \frac{freq_{ij}}{\max_k freq_{kj}}$$

Inverse Document Frequency(IDF)

$$|idf| = \log \frac{|D|}{|\#(f1)|}$$

Term

$$s(t) = P(t \in y \mid t \in x)$$

Weighted Ration

$$WOddsRation(w) = P(w) \times OddsRatio(w)$$

Odd Ration

$$OddsRatio(f_i, c_j) = \log \frac{P(f_i \mid c_j)(1 - P(f_i \mid \neg c_j))}{(1 - P(f_i \mid c_j))(P(f_i \mid \neg c_j))}$$

A Review of Machine Learning Algorithms for Text-Documents Classification, Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee*, Khairullah khan

Keyword selection in TC (3)

- Keyword selection based on keyword/non-keyword classification
  - Word position, POS tag, DF …
  - SVM, MLP, NB

Keyword selection in TC (4)

- Joint dimension selection and classifier optimization. For example, keep only prominent dimensions by checking the regression coefficients in a linear model.
- Evolution approach, e.g., ant colony optimization.
- Graph based keyword selection, e.g., the rank method

Keyword selection in TC (5)

- Dimension reduction by linear transform. E.g., LDA, SVD, NMF
- Dimension reduction by semantic representation. E.g., LSI, PLSA, LDA

Sparse analysis

- Involving L1 generally leads to sparse solutions for both regression and classification models.
- We want to employ sparse analysis to select the prominent dimensions.
  - Joint optimization for keyword selection and model training
  - The optimization function is more related to the task goal (TC).

From LDA to SDA

- LDA can be cast to an optimization problem as follows(optimal score criterion)

$$min_{\beta_k,\theta_k}\{||Y\theta_k - X\beta_k||_2^2\} \quad s.t. \quad \frac{1}{N}\theta_k^T Y^T Y \theta_k = 1, \;\; \theta_k^T Y^T Y \theta_l = 0 \;\; \forall l < k$$

- Involving L1 leads to the sparse version of LDA.

$$min_{\beta_k,\theta_k}\{||Y\theta_k - X\beta_k||_2^2 + \gamma\beta_k^T\Omega\beta_k + \lambda||\beta_k||_1\}$$
$$s.t. \quad \frac{1}{N}\theta_k^T Y^T Y \theta_k = 1, \;\; \theta_k^T Y^T Y \theta_l = 0 \;\; \forall l < k$$

For two-class problems

$$min_{\beta,\theta}\{||Y\theta - X\beta||_2^2 + \gamma\beta^T\Omega\beta + \lambda||\beta||_1\}$$

$$s.t. \quad \frac{1}{N}\theta^T Y^T Y\theta = 1.$$

Sparse!

$$min_{\beta}\{||\hat{Y} - X\beta||_2^2 + \gamma\beta^T\Omega\beta + \lambda||\beta||_1\}$$

$$\hat{Y}_{n,k} = \sqrt{\frac{N}{N_k}}$$

From SVM to sparse SVM

$$y_n = w^T x_n + b$$

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2} ||w||_2^2 \quad s.t. \quad t_n y_n > 1 - \xi_n$$

Sparse!

$$\sum_{n=1}^{N} \xi_n + \gamma ||w||_2^2 + \lambda ||w||_1, \quad s.t. \quad t_n y_n > 1 - \xi_n$$

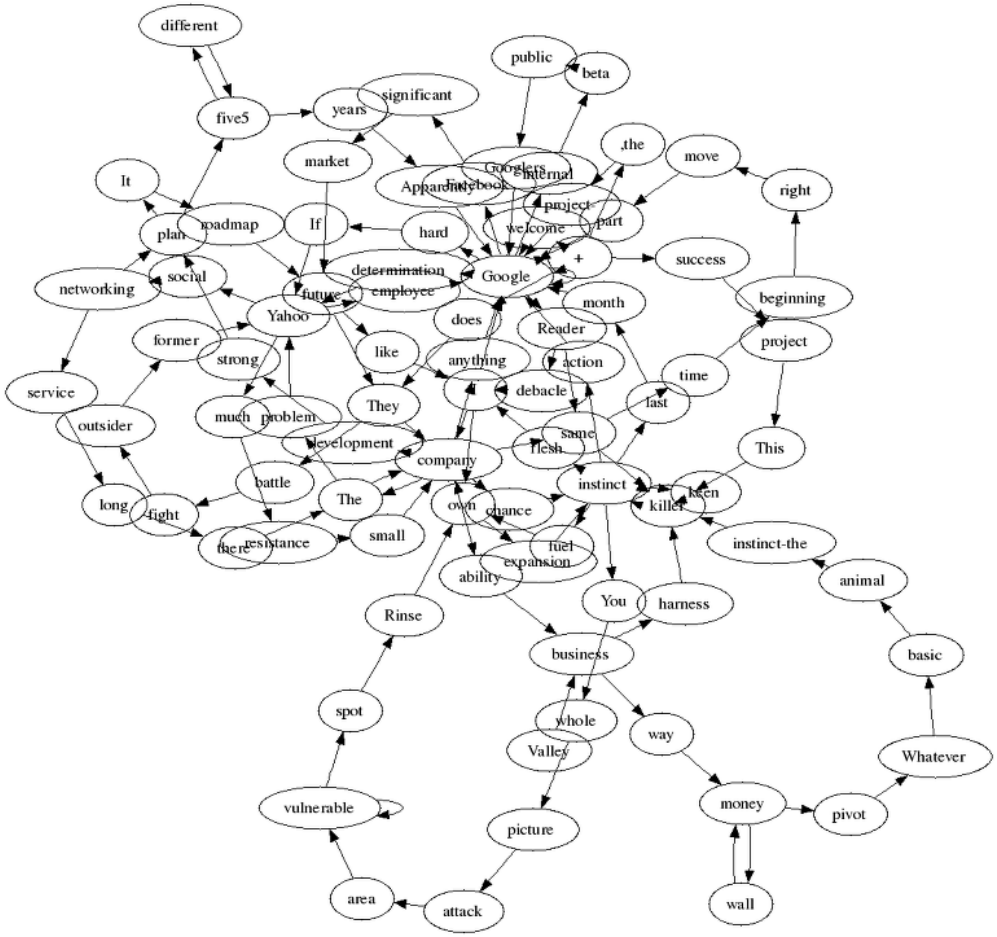
Experiments

- Simulation by text!
- Data profile
  - 1000 Uyghur documents. 500 health, 500 non-health
  - 70% for training, 10% for dev, 20% for evaluation
- Text pre-processing
  - Character purging
  - Latinization
  - Stop words removal

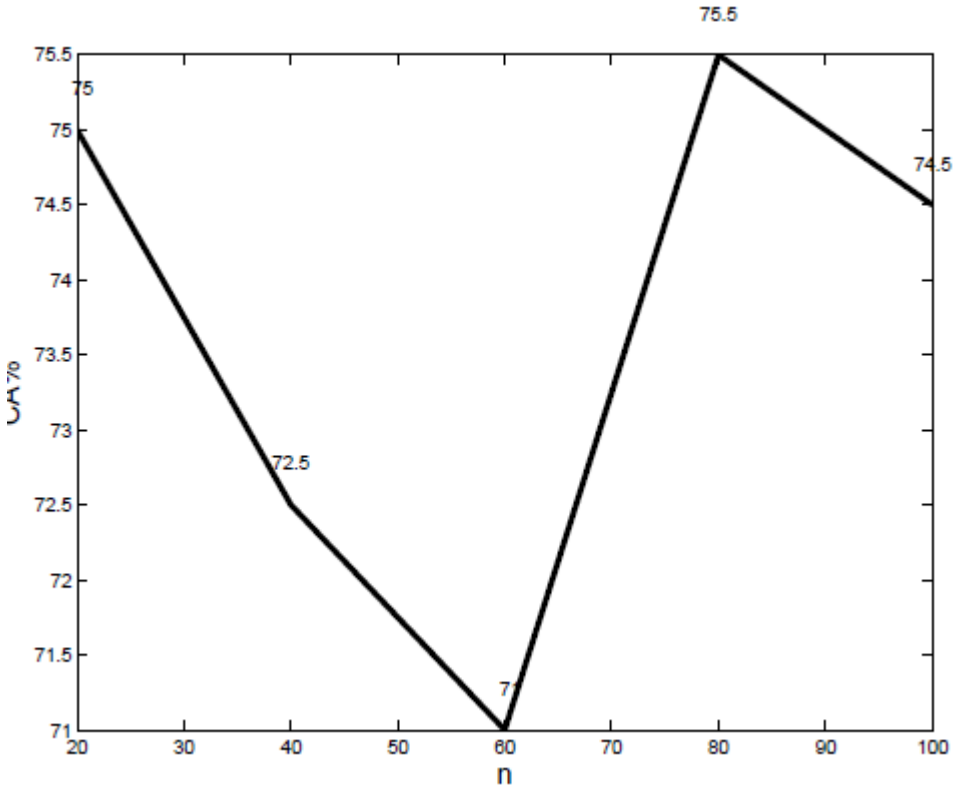
Latinization

ي	y	ت	t	ر	r	گ	g
ا	E	و	o	ن	n	ف	f
ل	I	پ	p	ك	N	ۋ	w
غ	G	م	m	چ	c	ۆ	O
ۇ	u	س	s	ې	e	ژ	J
ز	z	ب	b	ق	q	ج	j
ك	k	د	d	خ	H	ئ	,
ش	x	ۇ	E	ۇ	U	؛	;
ى	i	ۈ	v	ھ	h	؟	?

Reference system: Textrank



Reference system: Textrank

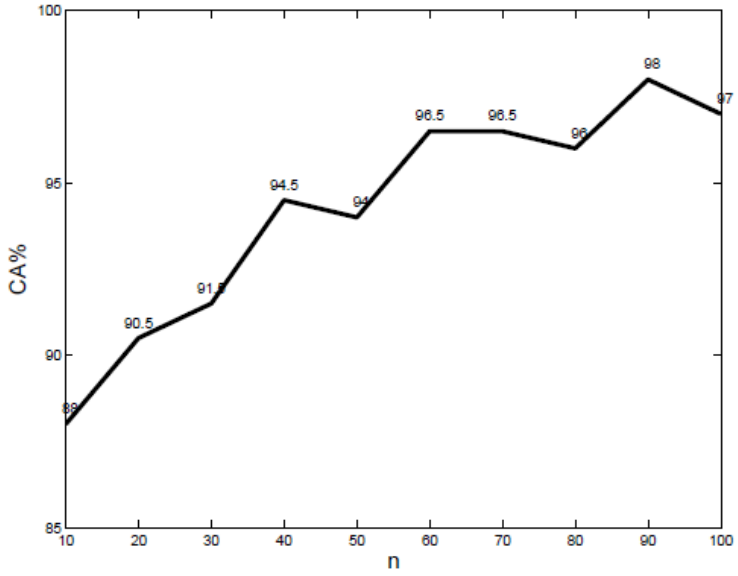


Reference system: document statistics

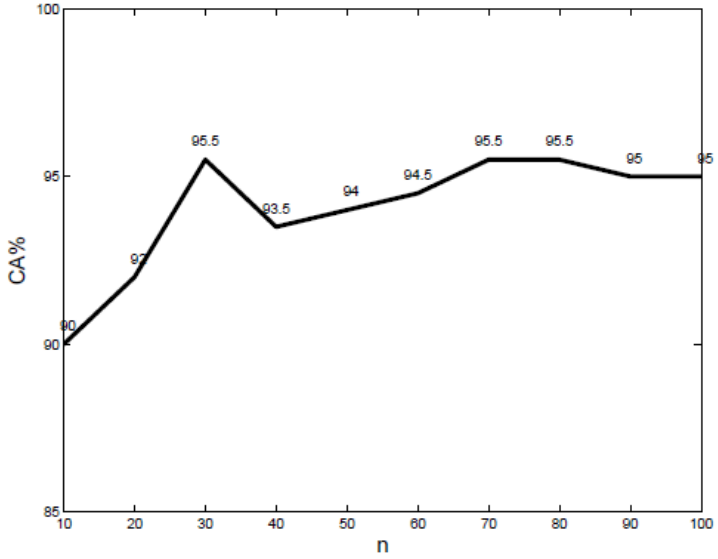
	CA%				
n	20	40	60	80	100
DF_h	88.0	87.5	88.5	89.5	89.0
TF_h	90.5	92.5	89.5	89.5	90.0
$TF_h * DF_h$	91.0	91.5	92.5	90.0	93.0
$TF_h * DF_h * IDF_{h+n}$	93.0	92.5	90.0	90.5	92.5

	CA%				
n	20	40	60	80	100
$DF_h - DF_n$	91.0	93.5	91.5	93.5	92.0
$TF_h - TF_n$	93.5	92.5	90.0	92.5	92.0
$TF_h * DF_h - TF_n * DF_n$	91.5	91.0	95.0	94.0	91.5
$TF_h * DF_h * IDF_{h+n} - TF_n * DF_n * IDF_{h+n}$	90.5	90.5	89.5	91.0	94.0

Sparse systems



SDA



Sparse SVM

Extracted keywords

TextRank		TF_h-TF_n		SDA		Sparse SVM	
original	ئەسلىدىكى	blood	قان	blood	قان	tooth	چىش
certain	بەرەر	benefit	پايدا	benefit	پايدا	blood	قان
done	قىلىشنىك	heart	يۈرەك	traffic	قاتناش	cold	زۇكام
age	ياشنىك	can	بولىدۇ	heart	يۈرەك	heart	يۈرەك
hand	قولغا	disease	كېسەللىك	can	بولىدۇ	diabetes	دىئابېت
mother	ئانىلار	more	كۆپ	disease	كېسەللىك	liver	جىگەر
liver	جىگەر	induce	كەلتۈرۈپ	more	كۆپ	joint	بوغۇم
oneself	ئۆزىگە	cure	داۋالاش	induce	كەلتۈرۈپ	fever	قىزىتما
child	بالىلارنى	body	بەدەن	cure	داۋالاش	smoke	تاماكا
infection	يالليۇغىدىن	property	خاراكتېرلىك	body	بەدەن	cancer	راك

Conclusions

- Keyword selection based on sparse analysis is theoretically sound and experimentally works well.
- Sparse SVM obtains better performance than SDA with very limited keywords
- How about in real ASR and with online decision?