SUBWORD LATENT SEMANTIC ANALYSIS FOR TEXTTILING-BASED AUTOMATIC STORY SEGMENTATION OF CHINESE BROADCAST NEWS

Yulian Yang, Lei Xie

A , S L P G , S f C S ,
N P U , X
@ . . . , @ - .

ABSTRACT

T f (LSA) / - f - (ASR) , LSA, f T T f - f C . LSA f f . W LSA - f - T T - - . S - - , - f- (OOV) , - C ASR . T f LSA f . E - TDT2 M LSA- T T ff - f . C - -LSA- T T F1- f 0.6598 f 17.4% - T T 6.5% - - T T .

Index Terms— , T T , - , , , , ,

1. INTRODUCTION

A , , . I f , , f , , f - . S fi f (BN) , / f f . T f , / 1 , 2 / - . M - 3 , f 4 5 . L - - f - , . . , , - , T T 6 - - BN 7

T N N S F - f C (60802085), R F f D P f H E C (20070699015), N S B R - P f S P (2007F15), NPU A S P (07XE0150), NPU F f F R (W018103).

8 f ffi . T - ff f ff f , f f . I fi . C *et. al.* (LSA) 9 f f - - . C / , LSA f - *concept matching*. LSA- - f (PCA) ff 10 . P - , (. . BN) f f f (LVCSR). S , f . T f - f- - f , f , 30% f E - 40% f M TRECVID 2006. B , - f- (OOV) f . f f . T OOV f C f f C . C OOV , - f BN . M f f 11 , C ( ) T T - - f C BN 7 . S f , . E - f C , - f f C - 7 . I , LSA, f . W ff f . W f f f - LSA .

2. CORPUS

W TDT2 M BN C . 53 f VOA M C BN. T 177 .

. TDT2 D LVCSR , 37%, 20% 15%, . W - P . W - : 90 (1390) 87 (1400) . A TDT2, 15- -

3. LATENT SEMANTIC ANALYSIS

L (LSA) 12 . H , - . A , . H . LSA - PCA 10 . I LSA, - , , .

G $\Gamma = \{t_1, t_2, t_3 \ldots, t_J\}$ $\{w_1, w_2, w_3 \ldots, w_I\}$, LSA $I \times J$ **M**, M_{ij} w_i t_j. S (SVD) 10 **M**

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T. \quad (1)$$

T **U** **V** $\mathbf{MM}^T$ $\mathbf{M}^T\mathbf{M}$, , Σ - . $\mathbf{MM}^T$ - w_i - $\{w_1, w_2, \ldots, w_I\}$. T *K* **U**, Λ_K, *K* ( ) $\mathbf{MM}^T$ *K*- . T $I-K$. A , w_i , . .,

$$w_i \rightarrow \Lambda_K(i) \quad (2)$$

$\Lambda_K(i)$ *i* Λ_K.

4. MOTIVATIONS OF USE OF SUBWORDS

L . H , - ASR - . W - , C . T . S . T , LSA . . H C . T ASR - .

4.1. Robustness to Flexibility in Word Segmentation

C E . A C , ,

Table 1. S TDT2. E -

O	ASR	B
阿尔及利亚 (*Algeria*)	鲍尔 激励 要 (*Bauer drive want*)	*a er ji li ya* *bao er ji li yao*
奥尔布莱特 (*Albright*)	二 步 莱特 (*two step Wright*)	*ao er bu lai te* *er bu lai te*
互联网 (*internet*)	互 连网 (*mutual connection*)	*hu lian wang*
赈济 (*relieve*)	震级 (*quake magnitude*)	*zhen ji*
过失 (*defect*)	国事 (*national affair*)	*guo shi*

. I , C C . A , - . T . F , 北韩 (N K) 北 (N) 韩 (K) TDT2 M . I , . T C - . H , .

4.2. Robustness to Speech Recognition Errors

E C . S . I M , 1200 6500 C . W , 400, . T *homophones* . T - C . T C ASR , - , . T 1 TDT2 ASR . F , 阿尔及利亚 (A) /鲍尔 激励 要/ (B) . F 奥尔布赖特 (A) - /二步 莱特/ (W). R . H , . F , 阿尔及利亚 /鲍尔 激励 要/, *bu lai te* 赖特 奥尔布赖特 /二 步 莱特/,

4.3. Robustness to OOV Words

T C - C . H C . C , OOV C ASR , - BN . N (NE) OOV , . 10% BN . A OOV - ( )

Table 2. S f OOV v. f TDT2. E - .

C		B
OOV v.	: 王行才 (*a Chinese name*)	*wang you cai*
ASR	当 行 财 (*when have money*)	*dang you cai*
	王 油菜 (*king rape*)	*wang you cai*
	邦友 才 (*national friendship talent*)	*bang you cai*
OOV v.	: 莱温斯基 (*Lewinsky*)	*lai wen si ji*
ASR	来 文 斯 基 (*come article this base*)	*lai wen si ji*
	来 问 司机 (*come ask driver*)	*lai wen si ji*
	来 的 司机 (*show-up driver*)	*lai de si ji*
OOV v.	: 科索沃 (*Kosovo*)	*ke suo wo*
ASR	克 祖国 (*gram motherland*)	*ke zu guo*
	客座 我 (*guest me*)	*ke zuo wo*

. F , f OOV v. C C - (. .). A , ff - v. , . T 2 v. f - OOV v. f TDT2. F , OOV f ff 莱温斯基 (L v.) , . ., /来 问 司机/, /来 文 斯 基/ /来 的 司机/, v. f . M - - OOV v. v. f ASR . ( ).

5. WORD AND SUBWORD LSA IN TEXTTILING

5.1. TextTiling-based Story Segmentation

T T T 7 : , - fi . A , (. . v.). S ASR TDT2 v. , f v. . T f v. S 5.3.

I , fi f - . T - *g* *lexical score*:

$$(g) = \cos(\mathbf{v}_s, \mathbf{v}_{s+1}) = \frac{\sum_{i=1}^{I} v_{s,i} v_{s+1,i}}{\sqrt{\sum_{i=1}^{I} v_{s,i} v_{s,i} \times \sum_{i=1}^{I} v_{s+1,i} v_{s+1,i}}} \quad (3)$$

v. $\mathbf{v}_s, \mathbf{v}_{s+1}$ f f v. - s $s+1$ g, $v_{s,i}$ i f $\mathbf{v}_s$, . ., f f w_i s. I , v. v. v. - fi fi f (T). T : 1) - ASR f f f ; 2) f v. v. v. v. 7 . S - , v. . L $\{T, T+\Delta, T+2\Delta ...\}$, v. Δ $\Delta \leq T$. T T *depth score* . fi . D :

$$(u) = (\quad (p_l) - \quad (u)) + (\quad (p_r) - \quad (u)) \quad (4)$$

v. u , p_l p_r f u, . T . F , fi f , v. v. - fi θ .

5.2. Applying LSA to TextTiling

I , LSA f f 12 . A S 5.1, T T . T LSA f T - T f f . A f , ASR - , v. ASR - (0.8) LSA f f - - . T LSA - fi (S 5.1). T f f v. v. . 1) S - . 1 . 2) C - fi v. v. v. f v. . T v. f- f - . W LSA (S 3) - SVDLIBC [1]. A , w_i LSA f $\Lambda_K(i)$ f (2) v. . W $\Lambda_K(i)$ T T . T f - (fi S 5.1) *K*-

$$\hat{\mathbf{v}}_s = \sum_{i=1}^{I} v_{i,s} \times \Lambda_K(i), \quad (5)$$

v. $v_{i,s}$ f f w_i. T f , - g -f .

$$(g) = \cos(\hat{\mathbf{v}}_s, \hat{\mathbf{v}}_{s+1}) = \frac{\sum_{i=1}^{K} \hat{v}_{s,i} \hat{v}_{s+1,i}}{\sqrt{\sum_{i=1}^{K} \hat{v}_{s,i} \hat{v}_{s,i} \times \sum_{i=1}^{K} \hat{v}_{s+1,i} \hat{v}_{s+1,i}}}, \quad (6)$$

v. $\hat{v}_{s,i}$ f $\hat{\mathbf{v}}_s$. S - fi f E . (6). S fi , v. w_i v. ASR , . v. -LSA- T T .

5.3. Subword-LSA-based TextTiling

W f LSA ff C v. , . ., - . F - f v. $\{w_1 w_2 w_3 \dots w_Q\}$, - fi f (-) ( ), . .,$\{c_1 c_2 c_3 \dots c_L\}$. T - v. f

$$: \{c_1 c_2 \ c_2 c_3 \ c_3 c_4 \cdots c_{L-1} c_L\}, \quad (7)$$

$$: \{c_1 c_2 c_3 \ c_2 c_3 c_4 \ c_3 c_4 c_5 \cdots c_{L-2} c_{L-1} c_L\}, \quad (8)$$

. H v. - f . T f - f f v. , v. v. .

[1] :// . . / / /

A S 5.2, LSA -
f - f -
, w_i -
. L E . (6) -
.

6. EXPERIMENTS

W (1)
- T T 6 , (2) .
T T 7 , (3) -LSA- T T (4)
-LSA- T T . E
fi f
F1- f . P
f $T = 50$ $\Delta = 20$
f - T T . T
T Δ fi f
θ LSA K f .
[2]. E
f F1- F . 1.
R LSA T -
T f C
BN LSA f
T T .
LSA. B -LSA- T T f .
C -LSA F1- f 0.6598
f 17.4%
T T (f) 6.5%
- - T T . T f f .
LSA f f -
(2957 f 395 f).
T (LSA f) f /
. . T f
LSA f f .
C - f .
.
f
-LSA-

T S
T T . T
-
. A

T T
. T
(LSA)
SA
-
LSA,
-
- -
0.6598
- .
.