Using Accessor Variety Features of Source Graphemes in Machine Transliteration of English to Chinese

Mike Tian-Jian Jiang

Department of Computer Science, National Tsing Hua University

Chan-Hung Kuo and Wen-Lian Hsu

Institute of Information Science, Academia Sinica

November 12, 2012

Introduction to Machine Transliteration

* What is machine transliteration ?
  * Subfield of computation linguistics
  * Proper nouns and technical terms across languages
* Transliteration modeling approaches are as follow:
  * Phoneme-based
  * Grapheme-based, which is also known as direct orthographical mapping (DOM)
  * Hybrid of phoneme and grapheme

Proposed Approach

* Grapheme-based approach of English-to-Chinese (E2C) transliteration
  * Many-to-many alignment (M2M aligner)
  * Conditional Random Field (CRF)
  * Feature based on source graphemes
    * Accessor Variety (AV)
* Adopt the same definition of transliteration as during the NEWS 2009 workshop at ACL-IJCNLP 2009

Concept of M2M-aligner

* Many-to-Many alignment
  * Different length between letter and phoneme strings
  * Training data lacks explicit alignment
  * Accurate grapheme-to-phoneme relationships

A	BE	RT
阿	贝	特

* The M2M-aligner
  * Aligns between substrings of various lengths (based on EM)
  * Unsupervised method for generating alignment without null graphemes

Concept of Accessor Variety

* Accessor Variety (AV)
  * Evaluating the likelihood that a character substring is a Chinese word
  * Determination is related to a perspective of ***n-gram*** and information theory of cross entropy
* The AV of a string s is defined as :

$$AV(s) = min\{L_{av}(s), R_{av}(s)\}$$

Transliteration Using EM and CRF

* Previous works of CRF-based transliteration
  * Report only one configuration of CRF
  * Alignments of name pairs were prepared by GIZA++ or by human annotators
* This study proposed
  * Different feature sets and context depths
  * Automatic procedure using EM-based M2M-aligner

Example of M2M-aligner

* M2M-aligner
  * Maximize the likelihood of the observed word pairs by using the EM algorithm
  * To obtain better alignment results, the parameters was set
    * MaxX = 8 (Source Side), MaxY = 1 (Target Side)

Source	Target	M2M-Aligner Result	
RANARD	拉纳德	R:A|N:A:R|D|	拉|纳|德

* CRF Toolkit
  * Wapiti

CRF Alignment Labeling

* CRF alignment labeling

Character (Grapheme)	Label
R	*B* 拉
A	*I*
N	*B* 纳
A	*I*
R	*I*
D	*B* 德

  * *B* an *I* indicate whether or not the character is in the starting position of the chunk

CRF Labeling Scheme

* CRF labeling scheme
  * Context depths(template) : one or two characters
  * AV feature
  * Label
    * Tag : BI or BIE
    * Chinese char position : only B or all of positions

Example of CRF Labeling Scheme

Feature Template	AV	Tag	Chinese Char
C_0, C_{-1}, C_1 C_{-2}, C_2 $C_0C_1, C_{-1}C_0$ $C_{-2}C_1, C_1C_2$	No	B, I	B and I

Grapheme	Label
R (C_{-2})	B 拉
A (C_{-1})	I 拉
N (C_0)	B 纳
A (C_1)	I 纳
R (C_2)	I 纳
D	B 德

CRF with AV Feature

* Why AV ?
  * The standard runs of NEWS is only using the data
  * Unsupervised feature selection from data
* CRF with AV
  * AV can be extracted from large corpora without any manual segmentation
  * AV of un-segmented English names from training, development, and test data might help enhancing E2C transliteration

The Concept of AV Score

* AV Score
  * The representation accommodates both the character position of a string and the string's likelihood ranking by the logarithm

$$f(s) = r, if\ 2^r \leq x \leq 2^{r+1}$$

  * The logarithm ranking mechanism is inspired by Zipf's law to alleviate the potential data sparseness of infrequent strings

Example of AV Score and CRF Labeling Format

* Example of AV Score

$$AV(s) = min\{L_{av}(s), R_{av}(s)\}$$

AV(RA) =32	**AV(RAB) = 32** **AV(FRA) = 40**

* CRF labeling format

$$f(s) = r, if\ 2^r \leq x \leq 2^{r+1}$$

RA	$log_2(32) = 5$	**R B** **A E**	**R $5B$** **A $5E$**

Example of CRF Training Data with AV

Input	AV Feature					Label
	1 Char	2 Char	3 Char	4 Char	5 Char	
R	7S	5B	4B	2B	0B	B 拉
A	7S	5E	$4B_1$	2B	0B	I
N	6S	5E	4E	$2B_2$	$0B_2$	B 纳
A	7S	5E	3E	$2B_1$	0B	I
R	7S	5E	3E	$2B_2$	0I	I
D	7S	2E	3E	2E	0E	B 德

Experimental Data

* NEWS 10
  * Development Set : 5792 name pairs
  * Training Set : 31961 name pairs
  * Test Set : 3000 name pairs
* NEWS 09
  * Development Set : 2896 name pairs
  * Training Set : 31961 name pairs
  * Test Set : 2896 name pairs

Evaluation Metrics (ACC)

* Word accuracy in Top-1 (ACC)
  * Measures correctness of the first transliteration candidate in the candidate list

$$ACC = \frac{1}{N}\sum_{i=1}^{N}\begin{Bmatrix}1\ if\ \exists\ r_{i,j}: r_{i,j} = c_{i,1};\\ 0\ otherwise\end{Bmatrix}$$

Evaluation Metrics (Mean F-score)

* Fuzziness in Top-1 (Mean F-score)
  * Measures how different, on average, the top transliteration candidate is from its closest reference

$$LCS(c,r) = \frac{1}{2}\left(|c| + |r| - ED(c,r)\right)$$

$$r_{i,m} = \arg\min\left(ED(c_{i,1}, r_{i,j})\right)$$

$$R_i = \frac{LCS(c_{i,1}, r_{i,m})}{|r_{i,m}|} \quad P_i = \frac{LCS(c_{i,1}, r_{i,m})}{|c_{i,1}|} \quad F_i = 2\,\frac{R_i \times P_i}{R_i + P_i}$$

Evaluation Metrics (MRR)

* Mean reciprocal rank (MRR)
  * Measures traditional MRR for any right answer produced by the system, from among the candidates

$$RR_i = \begin{Bmatrix} 1 \; if \exists \; r_{i,j}, r_{i,k} : r_{i,j} = c_{i,k}; \\ 0 \;\; otherwise \end{Bmatrix}$$

$$MRR = \frac{1}{N} \sum_{i=1}^{N} RR_i$$

Evaluation Metrics (MAP_{ref})

* MAP_{ref}
  * Measures tightly the precision in the n-best candidates

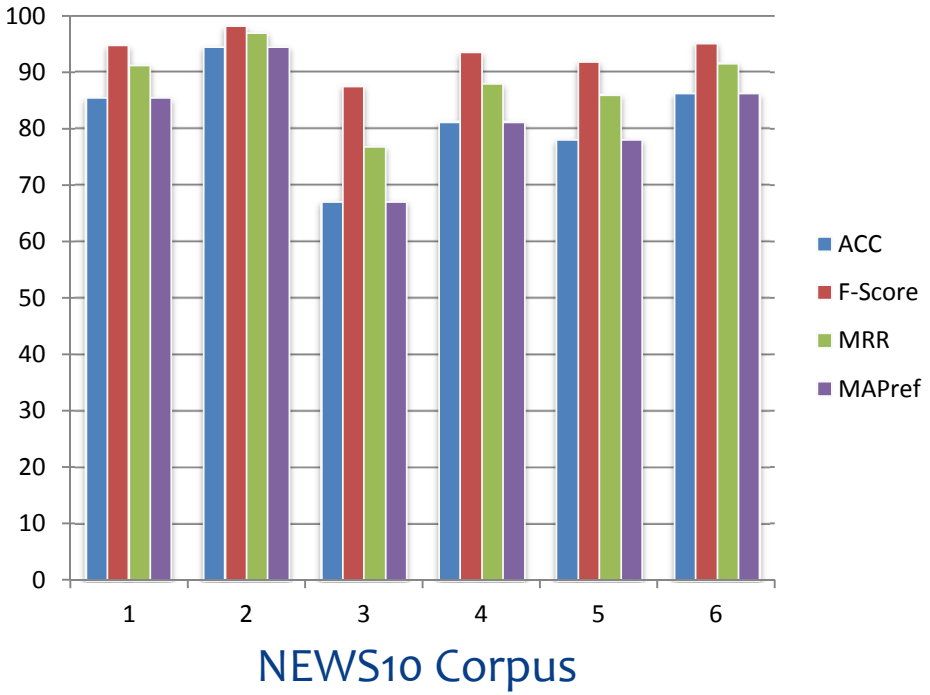
$$MAP_{ref} = \frac{1}{N}\sum_{i}^{N}\frac{1}{n_i}\left(\sum_{k=1}^{n_i} num(i,k)\right)$$

Experiment Design

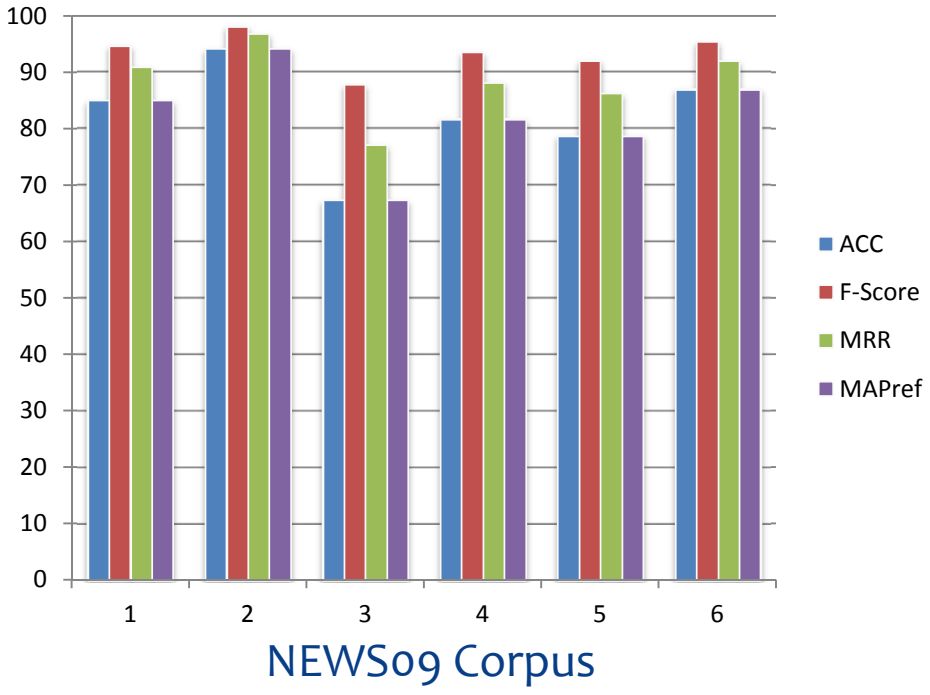
* Pilot tests
  * Both the training set and the development set
  * Optimizing feature combinations and M2M and Wapiti CRF parameters by evaluating of the development set

* The accuracy and F score were compared
  * Between development sets and test sets from NEWS10 and NEWS09

Evaluation Scores of E2C on Development Set

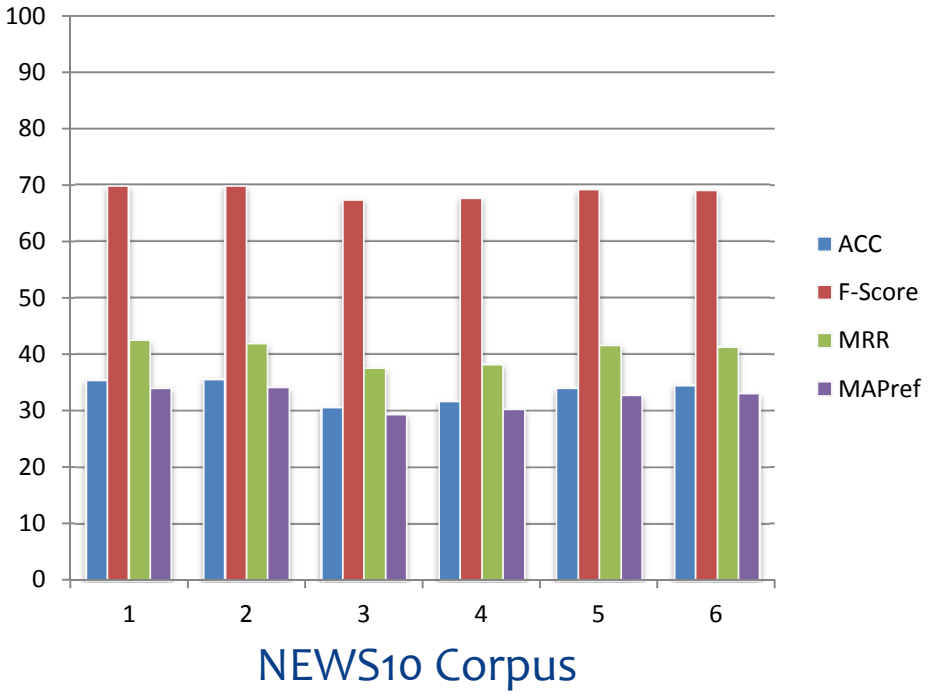


NEWS10 Corpus

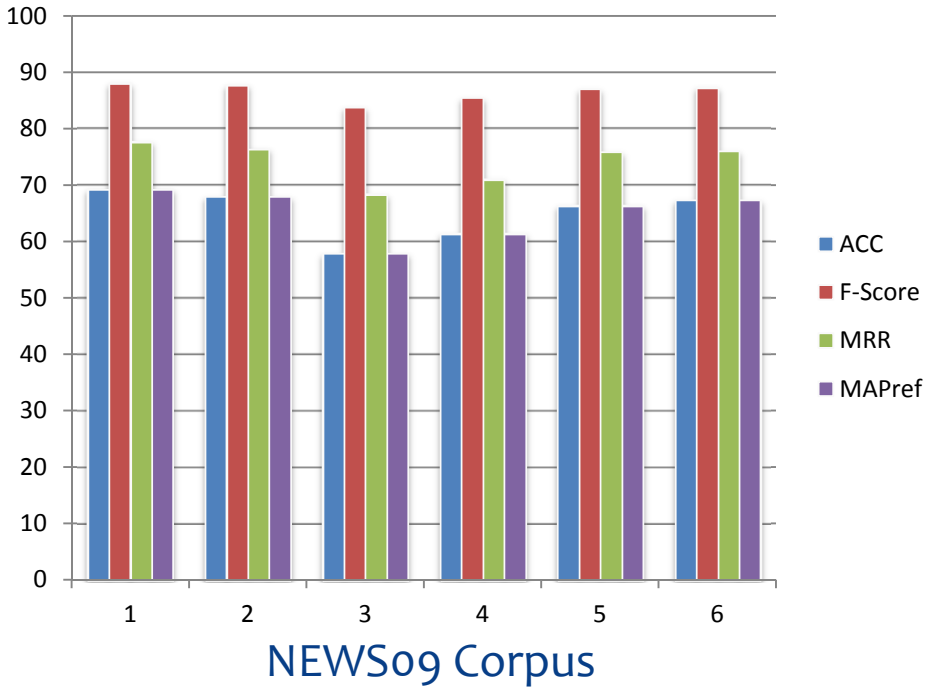


NEWS09 Corpus

Evaluation Scores of E2C on Test Set



NEWS10 Corpus



NEWS09 Corpus

Analyzing of NEWS Data

* Phenomenon of development sets (phrasal named entities)
  * Unseen in training sets
  * Unused in test sets
* Noisy alignments during the training phases

Name pair	Alignment
COMMONWEALTH OF THE BAHAMAS	巴哈马 / 联邦
ARAL SEA	咸 / 海

The C2E Problem

* Problems of Chinese to English (C2E) experiment
  * CRF L-BFGS training requirement (memory)
  * Too many labels and features
  * C2E transliteration is a one-to-many mapping but E2C is a many-to-one mapping

CRF Training Cost

* CRF training cost
  * The time complexity of a single iteration

$$\text{CRF L-BFGS} = O(L^2 NTF)$$

* Contribution rate C
  * realizing which standard runs are better choice

$$C = log_2(L^2 F_{total})$$

Contribution Rate

ID	F_{total}	L	C_{Test}^{Acc}	C_{Test}^{F}	C_{Test}^{MRR}	C_{Test}^{MAP}
1	2,501,328	744	**0.0292**	0.0575	**0.0350**	**0.0280**
2	4,882,872	744	***0.0287***	0.0561	***0.0337***	***0.0275***
3	1,125,744	376	0.0273	**0.0601**	0.0335	0.0261
4	2,322,176	376	0.0275	***0.0588***	0.0332	0.0263
5	2,680,512	1,104	0.0272	0.0552	0.0333	0.0262
6	2,975,280	1,104	0.0275	0.0549	0.0329	0.0263

ID	F_{total}	L	C_{Test}^{Acc}	C_{Test}^{F}	C_{Test}^{MRR}	C_{Test}^{MAP}
1	2,472,300	738	**0.0571**	0.0725	**0.0640**	**0.0571**
2	4,824,306	738	***0.0547***	0.0710	0.0610	***0.0547***
3	1,113,405	373	0.0517	**0.0748**	0.0610	0.0517
4	2,302,156	373	0.0533	***0.0742***	***0.0617***	0.0533
5	2,651,449	1097	0.0530	0.0695	0.0606	0.0530
6	2,946,542	1097	0.0536	0.0695	0.0605	0.0536

Conclusion

* E2C transliteration with AV as additional graphemic features
* Appropriate parameters
  * M2M-aligner
  * Context depth and CRF labeling scheme
* Future research
  * Applying different approaches to recognize C2E transliteration with efficient memory usages

Thanks For Your Listening !

Performance of Other Transliteration System

ACC	F-Score	MRR	MAPref
0.731	0.895	0.812	0.731
0.717	0.890	0.785	0.717
0.713	0.883	0.794	0.713
0.666	0.864	0.765	0.666
0.652	0.858	0.755	0.652
0.646	0.867	0.747	0.646
0.643	0.854	0.745	0.643
0.621	0.852	0.718	0.621
0.619	0.847	0.711	0.619
0.607	0.840	0.695	0.607

Reference: Li *et al.* 2009. Report of NEWS 2009 Machine Transliteration Shared Task.

Six Configurations of CRF Labeling

ID	Feature Template	AV	Label	
			Tag	Chinese Char
1	C_0, C_{-1}, C_1 C_{-2}, C_2 $C_0C_1, C_{-1}C_0$ $C_{-2}C_1, C_1C_2$	No	*B*, *I*	*B* and *I*
2	C_0, C_{-1}, C_1 C_{-2}, C_2 $C_0C_1, C_{-1}C_0$ $C_{-2}C_1, C_1C_2$	Yes	*B*, *I*	*B* and *I*
3	C_0, C_{-1}, C_1 C_{-2}, C_2 $C_0C_1, C_{-1}C_0$ $C_{-2}C_1, C_1C_2$	No	*B*, *I*, *E*	*B*
4	C_0, C_{-1}, C_1 C_{-2}, C_2 $C_0C_1, C_{-1}C_0$ $C_{-2}C_1, C_1C_2$	Yes	*B*, *I*, *E*	*B*
5	C_0, C_{-1}, C_1 $C_0C_1, C_{-1}C_0$	No	*B*, *I*, *E*	*B*, *I* and *E*
6	C_0, C_{-1}, C_1 $C_0C_1, C_{-1}C_0$	Yes	*B*, *I*, *E*	*B*, *I* and *E*