Integrating an Unsupervised Transliteration Model into Statistical Machine Translation

Nadir Durrani, Hassan Sajjad, Hieu Hoang and Philipp Koehn

Transliteration

- Languages are written in different Scripts
  - Russian, Bulgarian and Serbian – Cyrillic Script
  - Urdu, Farsi and Pashto – Arabic Script
  - Hindi, Marathi and Nepalese – Devanagri Script
- Transliteration is conversion from one script to other
  - सीमा (sima) → Seema
  - مورغان (morghan) → Morgan
  - Талботу (tælbət) → Talbot

- Utility
  - Cross Lingual IR
  - Terminology Extraction
  - Machine Translation
    - OOVs, Disambiguation, Closely Related Language
- Transliteration System
  - Rule-based approach
  - Data-driven approach
    - Requires transliteration corpus

Transliteration Mining

Магазин	Shop
власть	Power
Аналог	Analog
Пакистан	Pakistan
Нужда	Need
…	…
…	…
Энтони	Anthony

➡

Аналог	Analog
Пакистан	Pakistan
Энтони	Anthony

Supervised and Semi-Supervised (Sherif and Kondrak, 2007; Kahki et. al., 2011; Noeman and Madkour, 2010)

Fully Unsupervised (Sajjad et. al., 2012)

Unsupervised Transliteration Mining

✓	аналог analog	0.83
✗	система analog	0.05
✓	энтони anthony	0.71
✗	языково linguist	0.001

а | a	0.78	⬆
э| a	0.45	
a | e	0.07	⬇
г| g	0.75	
и| y	0.88	⬇
л| l	0.82	
…	…	⬇
…	…	

аналог analog	энтони anthony	⬆
Нужда Need	власть Power	⬇
Магазин Shop	Пакистан Pakistan	⬆
языково linguist	Америка America	⬆

Mining Model

$$p_1(e,f) = \sum_{a \in Align(e,f)} \prod_{j=1}^{|a|} p(q_j)$$

Transliteration Model

$$p_2(e,f) = \prod_{i=1}^{|e|} p_E(e_i) \prod_{i=1}^{|f|} p_F(f_i)$$

Non-Transliteration Model

- EM-based Algorithm
- with transliteration model, we can score training and extract transliteration corpus
- if we knew which pairs in the training are transliteration we can build transliteration model from these

$$p(e,f) = (1-\lambda)p_1(e,f) + \lambda p_2(e,f)$$

Integration into SMT

- Method 1: Replace OOV words with 1-best
- Method 2: Select best transliteration from n-best in post decoding
- Method 3: Pass transliteration phrase-table into decoder
- Features
  - Transliteration Phrase Table
  - LM-OOV Feature

Lang	Test	B_0	M_1	M_2	M_3	OOV
AR	$iwslt_{11}$	26.75	+0.12	+0.36	+0.25	587
	$iwslt_{12}$	29.03	+0.10	+0.30	+0.27	682
BN	jhu	16.29	+0.12	+0.42	+0.46	1239
FA	$iwslt_{11}$	20.85	+0.10	+0.40	+0.31	559
	$iwslt_{12}$	16.26	+0.04	+0.20	+0.26	400
HI	jhu	15.64	+0.21	+0.35	+0.47	1629
RU	wmt_{12}	33.95	+0.24	+0.55	+0.49	434
	wmt_{13}	25.98	+0.25	+0.40	+0.23	799
TE	jhu	11.04	-0.09	+0.40	+0.75	2343
UR	jhu	23.25	+0.24	+0.54	+0.60	827
Avg		**21.9**	**+0.13**	**+0.39**	**+0.41**	**950**

Lang	Data	$Train_{tm}$	$Train_{tr}$
		Sent	Types
Arabic	IWSLT-13	152K	6795
Bengali	JHU	24K	1916
Farsi	IWSLT-13	79K	4039
Hindi	JHU	39K	4719
Russian	WMT-13	2M	302K
Telugu	JHU	45K	4924
Urdu	JHU	87K	9131

Can we improve results by improving Mining?

–Mined system (MTS) vs. gold-standard system (GST)

–MTS has better rule coverage

)p(پ b) →(ب / al) → ε(ال–

–Gigapixel vs **Al**gega**b**ixel

WMT-14 (HI-EN an RU-EN)

– Best systems in ¾ cases
– Gains from 0.24-1.07
– Integrated in Moses
– used in Syntax based systems

	AR		HI	RU	
Test	$iwslt_{11}$	$iwslt_{12}$	**jhu**	wmt_{12}	wmt_{13}
B_0	26.75	29.03	15.64	33.95	25.98
MTS	27.11	29.33	16.11	34.50	26.38
GST	26.99	29.20	16.11	34.33	26.22
Δ	**-0.12**	**-0.13**	**0.0**	**-0.17**	**-0.16**
	Transliteration Pairs Used				
MTS	6795		4719	302K	
GST	1799		2394	1859	

Acknowledgement: The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287658.