A Comparative Study of Extremely Low-Resource Transliteration of the World's Languages

Winston Wu and David Yarowsky
Center for Language and Speech Processing, Johns Hopkins University
{wswu,yarowsky}@jhu.edu

Overview

For low-resource languages, there is often very little training data with which to train transliteration models. We compare several transliteration methods on the task of transliterating names into English. To handle unknown characters, we experiment with a pre- and post-transliteration step. We find that the phrase-based system performed best overall, but a simple method of combining inputs from multiple languages yielded much larger gains.

Methods

- Unidecode (baseline), a naïve method of Unicode to ASCII string mapping
- Moses (Koehn et al., 2007), a phrase-based machine translation system
- Sequitur (Bisani and Ney, 2008), a grapheme-to-phoneme system
- OpenNMT (Klein et al., 2017), a neural MT system

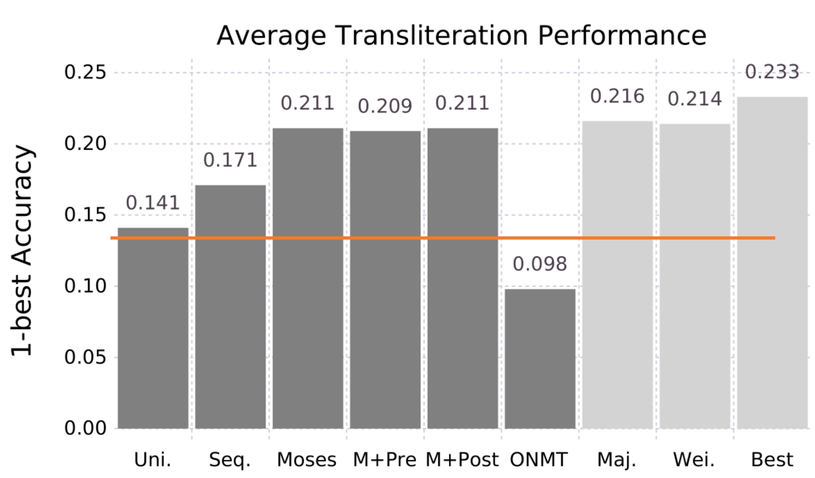
Data

We use the Bible Names Translation Matrix (Wu, Vyas, and Yarowsky, 2018). This data contains 1129 named entities across 529 languages. Due to the tiny number of examples for each language pair, this is considered *extremely low-resource*.

English	Yipma	Hanga	Koongo	Hanunoo
jesus	jizaazai	yeesu	yesu	hisus
christ	kɨraazɨ	kristu	klisto	kiristu
israel	yɨzɨrelɨ	juusi	isaeli	israil
david	devitɨ	dawuda	davidi	dabid
paul	polɨ	pool	pawulu	pablu
peter	pitai	piita	petelo	pidru
egypt	yɨzɨpɨ	yijipi	ngipiti	ihiptu
jerusalem	jeruzaalemɨ	jiruusilim	yelusalemi	hirusalim

Results

Moses exhibited higher exact match accuracy than the other approaches. Pre- and post-transliteration on Moses input/output had little effect. Standard methods of system combination showed slightly higher accuracy.



Experiment and Analysis

We trained transliteration systems from all languages into English, using the names as aligned bitext. Here, we summarize our observations of each system and compare the output with that of Moses, which achieved the highest accuracy for the single language pair systems.

Unidecode

- Language independent
- Cannot handle multiple letter substitutions (o -> us)
- Sometimes simpler model is better

Apurinã (apu)	English	Unidecode	Moses
épeso	ephesos	*epeso	*ephesus[4]
xório	julius	*xorio	julius
nikorao	nicolas	*nikorao	*nicolaus

Lang	Unidecode	Moses
Siyin (csy)	enoch	*enoc
Guahibo (guh)	jordan	*jordam
Ukranian (ukr)	puteoli	*putheoli
Murrinh-patha (mwf)	moses	*mouseus

OpenNMT

- Not enough data to train a good model
- Prefers shorter words

Lang	Source	Moses	OpenNMT
Qaqet (byx)	aleksandria	alexandria	*alandria
Frafra (gur)	metusela	*methushelah	*metusel
	alekzander	*alechzander	alexander
Hiri Motu (hmo)	eparona	*epharon	ephron
	mikaela	*micael	michael

Multi-source NMT

By adding a language word to the start of the input, we trained a neural MT system to learn a multi-source to single-target transliteration model, where the input is the concatenation of all training pairs for all languages. This greatly increased the size of the training data and resulted in a one-best exact match accuracy of 69%.

Source	Target
<ann> p o l ọ k i s	p o l l u x
<bnp> p o l u k s	p o l l u x
<kwf> p o l a k s	p o l l u x
<msy> p o l l u k s	p o l l u x
<mti> p o r a k u s	p o l l u x
<mto> p ó l u x	p o l l u x
<ncj> p ó l u x	p o l l u x
<rus> п о л л у к с а	p o l l u x

Sequitur

- Similar to Moses
- Makes linguistically plausible mistakes

Lang	Sequitur	Moses
Amele (aey)	elam	*ilam
	abilene	*abylene
Balinese (ban)	cleopas	*clopas
Bukawa (buk)	bartimaeus	*batimeas
Hawaiian Pidgin (hwc)	castor	*casthor
	phrygia	*phirygia
Hote (hot)	miletus	*miretus
	philetus	*piletus
	troas	*troaz

Resolving Unknown Characters

Analogous to resolving OOVs in MT, transliteration has the problem of unknown characters. We experiment with using pre- and post-transliteration to resolve unknown characters. Pre-transliteration converts the input characters into ASCII, potentially reducing the character set. Post-transliteration converts the output characters into ASCII, potentially transliterating characters that did not get handled by the translation system.

Pre-transliteration helps in certain cases but conflates character mappings in other cases.

Lang	Source	Moses	+Pre
Ankave (aak)	segɨria	*cenria	cenchrea
Greek (ell)	εὔα	*eὔa	eva
Ukranian (ukr)	марта	*marta	martha
Armenian (hye)	սողոմոնը	solomon	*solomone
Russian (rus)	косам	cosam	*kosam
Ossetian (oss)	тимейы	timaeus	*timee

Post-transliteration helps in a small number of languages.

English	malchus	felix	crete
Moses	* 말 chus	*pel 릭 s	* 크 re 테
+Post	malchus	*pelrigs[6]	*keurete

марта → Pre → marta → Moses → marta

말고의 → Moses → 말chus → Post → malchus

Bible Names Translation Matrix: *Creating a Translation Matrix of the Bible's Names Across 591 Languages*, Wu, Vyas, and Yarowsky (LREC 2018)

Code and data at `github.com/wswu/trabina`