

Name Translation in SMT: Learning When to Transliterate

Authors: U. Hermjakob, K. Knight and
H. Daume III

Presenter: Waleed Ammar

Contributions

1. Method for transliterating words
2. Identification of words to be transliterated
3. Integration in SMT system
4. Metric to evaluate NE translation

Contributions

1. Method for transliterating words
2. Identification of words to be transliterated
3. Integration in SMT system
4. Metric to evaluate NE translation

How to Transliterate a word?

The Problem

1. Phonetic ambiguity in source language

- سلم transliterates to “silm”
- ... or salam, sallam, sallim or sollam?
- سَلَّمَ VS. سلم

2. Proper spelling in target language

- جون كلارك transliterates to “Jon Clark”
- ... or John Clark?

How to Transliterate a word?

Search-Oriented Approach

1. Define $\text{translit-cost}(\text{ar-word}, \text{en-word})$
2. Define $\text{score}(\text{en-word} | \text{ar-word})$
 $= \log\text{-freq}(\text{en-word}) / 20 - \text{translit-cost}(e, a)$
3. Transliteration
 $= \text{argmax}_{\text{en-word}} \text{score}(\text{en-word} | \text{ar-word})$

How to Transliterate a word?

translit-cost(ar-word, en-word)

- Think Levenshtein edit distance
- 732 rules:
 - Substituting “q” for “ق” costs 0
 - Substituting “k” for “ق” costs 0.2
 - Substituting “k\$” for “\$ق” costs 0
 - Substituting “sh” for “ش” costs 0.1
 - ...
- No rule applies => match fails
- Accumulated cost > threshold => match fails

How to Transliterate a word?

translit-cost caveats

- 700+ rules per LP
- Longest applicable rule
- Context
- Multi-token transliterations
- Style flags

How to Transliterate a word?

$$\operatorname{argmax}_{\text{en-word}}$$

- Transliteration
= $\operatorname{argmax}_{\text{en-word}} \text{score}(\text{en-word} | \text{ar-word})$
- Too many candidate en-words (3.5M)
- Solution:
 - Map English n-grams to consonant skeletons
Rachmaninoff $\rightarrow$ rkmnnf, rmnnf, rsmnnf, rtsmnnf
 - Build a reverse index(skeleton) $\rightarrow$ English n-grams
 - Map an Arabic word to a consonant skeleton
 - Use the reverse-index to find candidate English n-grams
 - Compute score() for each candidate

How to Transliterate a word?

Evaluation

Contributions

1. Method for transliterating words
2. Identification of words to be transliterated
3. Integration in SMT system
4. Metric to evaluate NE translation

Which words to transliterate?

The Problem

وبينما حمل الجيش السوداني عبر الناطق الرسمي باسمه
الصوارمي خالد سعد الحركة الشعبية مسؤولية أمن وسلامة
الرهائن الصينيين، طالب السفير الصيني بالخرطوم ليو شياو
فوانغ الحكومة السودانية ببذل قصارى جهدها لإنقاذ رعايا
بلاده.

Which words to transliterate?

The Problem

1. Named Entities

Translit: **Al-Gaysh Sudanese**

Translit: **Al-Haraka Al-Shabeyya** **armi Khaled**

Translit: **Liu Xiao Fuang** **e's Movement** **armi Khaled**

Correct: **Liu Xiao Fuang**

الصوارمي خالد سعد الحركة

الرهائن الصينيين، طالب السفير الصيني بالخرطوم **ليو**

شياو فوانغ الحكومة السودانية ببذل قصارى جهدها لإنقاذ

رعايا بلاده.

Which words to transliterate?

The Problem

1. Named Entities
2. Out of Vocab

وبينما حمل الجيش السوداني عبر الناطق الرسمي باسمه
الصوارمي خالد سعد الحركة الشعبية مسؤولية أمن وسلامة
الرهائن الصينيين، طالب السفير الصيني بالخرطوم ليو شياو
فوانغ الحكومة السودانية ببذل قصارى جهدها لإنقاذ رعياء
بلاده.

Which words to transliterate?

The Problem

1. Named Entities
2. Out of Vocab

وبينما حمل الجيش السوداني عبر الناطق الرسمي باسمه
الصوارمي خالد سعد الحركة الشعبية مسؤولية أمن وسلامة
الرهائن الصينيين، طالب السفير الصيني بالخرطوم ليو
شياو فوانغ الحكومة السودانية ببذل قصارى جهدها لإنقاذ
رعايا بلاده.

Which words to transliterate?

The Problem

1. Named Entities
2. Out of Vocab
3. Transliterate-me tagger

وبي؟ما حمل الجيش السود؟اني غير الناطق الرسمي باسمه
الصوارمي خالد سعد الحركة الشعبية مسؤولية أمن وسلامة
الرهائن الصينيين، طالب السفير الصيني بالخرطوم ليو شياو
فوانغ الحكومة السودانية ببذل قصارى جهدها لإنقاذ رعايا
بلاده.

Which words to transliterate?

Transliterate-me Tagger

1. Take a bitext

صِقْلِيَّةٌ هِيَ أَكْبَرُ جَزِيرَةٍ فِي الْبَحْرِ الْأَبْيَضِ الْمَتَوَسِّطِ

Sicily is the largest island in the Mediterranean Sea

Which words to transliterate?

Transliterate-me Tagger

1. Take a bitext
2. Find transliteration pairs (P:99.5%, R:95%)

صِقْلِيَّةٌ هِيَ أَكْبَرُ جَزِيرَةٍ فِي الْبَحْرِ الْأَبْيَضِ الْمَتَوَسِّطِ

Sicily is the largest island in the Mediterranean Sea

Which words to transliterate?

Transliterate-me Tagger

1. Take a bitext
2. Find transliteration pairs
3. Tag Arabic words that can be transliterated

صِقْلِيَّةٌ هي أكبر جزيرة في البحر الأبيض المتوسط

Sicily is the largest island in the Mediterranean Sea

Which words to transliterate?

Transliterate-me Tagger

1. Take a bitext
2. Find transliteration pairs
3. Tag Arabic words that can be transliterated
4. Get rid of English

صِقْلِيَّة هي أكبر جزيرة في البحر الأبيض المتوسط

Which words to transliterate?

Transliterate-me Tagger

1. Take a bitext
2. Find transliteration pairs
3. Tag Arabic words that can be transliterated
4. Get rid of English
5. Split tagged Arabic sentences into train/test

Which words to transliterate?

Transliterate-me Tagger

1. Take a bitext
2. Find transliteration pairs
3. Tag Arabic words that can be transliterated
4. Get rid of English
5. Split tagged Arabic sentences into train/test
6. Train the “transliterate-me” tagger

Which words to transliterate?

Transliterate-me Features

صِقْلِيَّة هي أكبر جزيرة في البحر الأبيض المتوسط

- Surface form, $f(\text{surface_form})$
- Previous two words, $g(\text{prev_two_words})$
- Next two words, $h(\text{prev_two_words})$
- Prefixes
- Suffixes
- Total = 250 features

Which words to transliterate?

Transliterate-me Caveats

- Transliteration matching:
 - Multi-token names
 - Many-to-one links
 - Capital English words
 - Stopwords removed
 - Country/nationality name equivalence
 - Split prefixes
- Tagger:
 - “Stat section”
 - Averaged perceptron

Which words to transliterate?

Transliterate-me Evaluation

Reference	Precision	Recall	F-meas.
Raw test corpus	87.4%	95.7%	91.4%

- Test set = 10K sents

Contributions

1. Identification of words to be transliterated
2. Method for transliterating words
3. Integration in SMT system
4. Metric to evaluate NE translation

Integration in End-to-End SMT Approach

1. Tag Arabic sentence to be translated using “transliterate-me” tagger
2. Transliterate tagged items with unreliable counts in bitext
3. Add transliterations to the phrase table with the “transliterated” feature set to 1
4. At runtime, transliteration entries in the phrase table compete with original entries

Contributions

1. Identification of words to be transliterated
2. Method for transliterating words
3. Integration in SMT system
4. Metric to evaluate NE translation

How to Know We Did a Good Job?

NEWA: Named Entity Weak Accuracy

- Traditional metrics make no distinction btw dropping a comma and dropping a NE
- $$\text{NEWA} = \frac{\text{\# of NEs correctly translated}}{\text{Total \# of NEs}}$$
- A named entity is correctly translated if one of its proper translations appears in the translated sentence

End-to-End Evaluation

- Metric: NEWA
- NE type: all
- NE count: 1730
- Sentence count: 637

Gold Standard	BBN GS
Human 1	87.0%
Human 2	85.3%
Human 3	90.4%
Human 4	86.5%
SMT System	80.4%

End-to-End Evaluation

- NE types: PERson, GeoPoliticalEntity, ORGanization, FACility, PER.Nominal, LOCation
- BLEU: 50.70 → 50.96

NE Type	Count	Baseline SMT	SMT with Transliteration
---------	-------	-----------------	-----------------------------

All types	1730	1519 (87.8%)	1552 (89.7%)
-----------	------	--------------	--------------

Criticism

- Transliteration rules and their costs are hand-crafted
- Can't generate new words
- Unsupported claim about applicability to other language pairs
- Re-annotation partially simulate the transliteration algorithm by using transliterator output and ngram frequencies