Translating Words You’ve Never Seen

Nguyen Bach

Language Technologies Institute
Student Research Symposium
September 2006

Joint work with
Bing Zhao, Ian Lane, and Stephan Vogel
Translating Words You’ve Never Seen

Vietnamese news:
“Rô-ma-nô Prô-di đang đối mặt với thử thách thực sự đầu tiên trong cuộc vị của mình”

VN-EN Machine Translation:
“is facing the first real battle of the premiership”

Translating Words You’ve Never Seen
“Romano Prodi is facing the first real battle of the premiership”
Problem

- Machine Translation systems have a limited vocabulary, yet human language has an unlimited vocabulary.
- Out-of-vocabulary (OOV) words are guaranteed to produce errors.
- OOV words contain names.
- **Challenge:** Can we translate unseen names?
Examples

Name pairs: source and target languages
- GE: Konstantinopolis – En: Constantinople
- AR: KwmArAtwnjA – En: Kumaratunga
- SP: Adelaida – En: Adelaide
- FN: Tariffi – En: Tariff

Conjecture: Names can be transformed from spelling in one language to spelling in other languages

Let’s try with Arabic-English.
Severe Issues in Arabic Names

- Arabic is romanized by BAMA toolkit
  - خفاجي xfAjy
- Romanized Arabic names:
  - Missing vowels in writing:
    - خفاجی xfAjy
  - Possible correct translations:
    - mHSn Mahasin / Muhasan / Mahsan
Approaches

- Rule-based Transliteration
- Transliteration as Translation
Rule-based Transliteration - Examples

Examples
- AR: KwmArAtwijnA – En: Kumaratunga

Rules
- If source Begins with “Kwm” then Begin of target is “Kum”
- If source Ends with “ja” then End of target is “ga”
- If Middle of source is “A” then Middle of target is “a”
- If Middle of source is “twn” then Middle of target is “tun”

Templates
- Source | Target | Position | Score
- Kwm | Kum | Begin | 1.4324
Rule-based Architecture Overview

Bilingual NE corpus → Generator → Learner → Translation Hypothesis → Applicator → Picker → Top N Candidates

Training

Decoding
Rule-based Architecture Overview

- **Training - Generator:**
  - Given “lybyry” & “liberian” how many possible rules?
  - A: Alignment by calculating edit distance
    
    | lybyry | lybyry |
    |--------|--------|
    | |||| | |||| |
    | liberian | liberian |

  - Use all optimal paths to extract rules according to alignment paths
  - Distinguish rules for begin, middle, and end
  - Use consonants to anchor rule
Rule-based Architecture Overview

From 5820 pairs
Total: 19957 different rules
Max freq: 379
Min freq: 1

Head list
379 An an Begin
345 q ca Begin
303 X sh Begin
286 nd nd Middle
283 ry ri End
273 ny ni End
252 kt ph Begin
252 qr car Begin
219 x kha Begin
217 x kh Begin
Rule-based Architecture Overview

- Training - Learner:
  - How to know which rule is good or bad?
  - For each rule, apply it to the held-out data & use reduction of character errors as figure of merit

- Decoding - Applicator:
  - Application order: Begin -> End -> Middle
  - Confidence threshold: filter out unreliable rules
  - Application strategy: for each source word, find all possible rules, and apply them in order
Rule-based Architecture Overview

- Decoding - Picker:
  - Many possible transliteration candidates
  - Phonetic similarity to select top N candidates

- Decoding - Spell checker:
  - Google Suggest
  - Web frequency
Evaluation Setup

- **Training data:**
  - CMU Ar-En SMT: Vocabulary: 470K Arabic, 179K English; LM: 6-gram, 800 million words
  - 80K Arabic-English name pairs (GALE FOUO)
- **Blind test set:** Arabic-English Tides 2003
  - 286 unique tokens were left un-translated
  - Among them: 97 un-translated unique person, location names

<table>
<thead>
<tr>
<th>Arabic</th>
<th>BAMA</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>قشطة</td>
<td>qXTp</td>
<td>Qishta</td>
</tr>
<tr>
<td>ايتساخرف</td>
<td>fAytAxr</td>
<td>Weizsacker</td>
</tr>
<tr>
<td>والدكاني</td>
<td>sWldHmAny</td>
<td>al-Dahmani</td>
</tr>
<tr>
<td>يلويغيرز</td>
<td>zylwygyr</td>
<td>Zellweger</td>
</tr>
<tr>
<td>نايسين</td>
<td>vAksyn</td>
<td>Thaksin</td>
</tr>
</tbody>
</table>

- **Evaluation method:** Edit distance between hyp against possibly multiple references
  - Src = “mHmd”  Ref = Muhammad / Mohammed
  - Accept translation if edit distance = 1
Rule-based Performance

Rule-based Transliteration performance with top N candidates in Arabic Tides 2003 Eval set

- Baseline is 0%

16% 48%
Transliteration-as-Translation (T.a.T)

- Borrow Statistical Machine Translation (SMT) framework

Ideas

- Name pairs as sentence pairs in letter level.
- Train n-gram letter Language Model
- Train IBM models 4
- Decode with InterACT decoder
- Web spell checker
T.a.T Architecture

- Name pairs
- Preprocessing
- Letter-alignment Model (IBM Model 4)
- Language Model (5-gram)
- Phrase pairs
- Spell Checker
- Internet
- N-best Translation Hypothesis

Translating Words You've Never Seen
T.a.T: Key Points

- **Alignment:**
  - GIZA +
  - Fertility: restricted to 4
  - Letter classes: vowel, consonant, unknown
  - Monotone alignment
  - Log-linear model for phrase alignment with 13 features are computed for each n-gram

- **Decoding**
  - Monotone

- **Web spell checker (same as Rule-based)**
  - Web statistics and Google suggestions
Evaluation (Rule-based vs. T.a.T)

Rule-based Transliteration vs. Transliteration-as-Translation by percentage of different top N candidates in Arabic Tides 2003 Eval set

Significantly outperform rule-base

10/20/2006

Translating Words You've Never Seen
Evaluation (Google vs. T.a.T)

- The Arabic-English Google Web Translation (Google)
- Accuracy **45%** (as in June 20) for the 1-best hypothesis while T.a.T archives **52%**

<table>
<thead>
<tr>
<th>Source</th>
<th>Reference</th>
<th>T.a.T</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>سوماي</td>
<td>Sumaye</td>
<td>Sumaye</td>
<td>Somai</td>
</tr>
<tr>
<td>هازوميتسو</td>
<td>Hazumitsu</td>
<td>Hazumitsu</td>
<td>Hazoumitso</td>
</tr>
<tr>
<td>بلاد</td>
<td>Yalahow</td>
<td>Ylahu</td>
<td>Elaho</td>
</tr>
<tr>
<td>نكاختت</td>
<td>Nikbakht</td>
<td>Nkbakht</td>
<td>Nkbacht</td>
</tr>
<tr>
<td>ميكولاس</td>
<td>Mikulas</td>
<td>Mikulas</td>
<td>Mikoias</td>
</tr>
<tr>
<td>كومارانتوهج</td>
<td>Kumaratunga</td>
<td>Kumaratunga</td>
<td>Kumaratung</td>
</tr>
<tr>
<td>هندان</td>
<td>Handan</td>
<td>Handan</td>
<td>Hamedan</td>
</tr>
<tr>
<td>لماراندران</td>
<td>Mazandaran</td>
<td>Mazandaran</td>
<td>Mazandaran</td>
</tr>
<tr>
<td>ويكراكسينغه</td>
<td>Wickremasinghe</td>
<td>Wickramsin</td>
<td>The Ekermsingh</td>
</tr>
</tbody>
</table>
Incorporating T.a.T to SMT (GALE)

Arabic text source sentence

كولombo 4 يناير / شينخوا/ حذر رئيس الوزراء السريلانكي رانيل ويكرمسينغه الرئيسة تشاندريكا كوماراتونجا من مغبة تدمير عملية السلام التي ترعاها النرويج

SMT hypothesis

- in colombo 4 January 1997, the xinhua / warned by the prime minister {UNK رانيل ويكرمسينغه السريلانكي} chairperson {UNK تشاندريكا كوماراتونجا} cautioned the destruction of the peace process sponsored by norway

SMT with T.a.T

- in colombo 4 january 1997 , the xinhua / warned by the prime minister Sri Lankan Ranil Wikramasinghe charperson Chandrika Kumaratunga cautioned the destruction of the peace process sponsored by norway

Reference translation

- Colombo 04/01 (Xinhua) Sri Lankan Prime Minister Ranil Wickremasinghe warned the country's President Chandrika Kumaratunga of the consequences of destroying the peace process sponsored by the Norwegians
Conclusion & Future Work

- Unseen names convey key information.
- We can **translate unseen names** for Arabic-English MT system.
- Our system performance is comparable with the state-of-the-art system.

Future

- Configure to other languages pairs (Italian – English, Chinese – English, ...)
- Incorporate to cross-lingual IR systems
- Translating OOV words from Speech Recognition
Acknowledgements

- Stephan Vogel, Alex Waibel
- Bing Zhao, Ian Lane, Sanjika Hewavitharana - SMT - InterACT
- GALE, STEEM projects, InterACT fellowship.

Thank you