Xing Shi, Kevin Knight ¹ Heng Ji ²

¹Information Sciences Institute
Computer Science Department
University of Southern California
{xingshi, knight}@isi.edu

²Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
jih@rpi.edu

June 24, 2014
Overview

1. Introduction
2. Data
3. Evaluation
4. Model
5. Training
   - Phoneme-based model
   - Phoneme-phrase-based model
   - Word-based model
   - Hybrid training/decoding
6. Experiments
7. Conclusion and Future work
Can people speak a language they don’t know?
Yes, use a phrasebook

Thank you for this wonderful meal.

Diamond 谢谢你这顿美餐。

San ke you fou re si wan de fou mi ou

三可油否热斯弯德否迷欧
Yes, use a phrasebook
Yes, use a phrasebook

He want to know: Where is the restroom?
Yes, use a phrasebook

What if we want to say something **beyond** the phrasebook?
Or, a speech-to-speech translator

from: proto-knowledge.blogspot.com

However, direct Human interactivity is much more fun!
Our solution

- Easily pronounceable
  - Both input $T(S)$ and output $T'(S)$ are in speaker’s language.

- Understandable by listener
  - $T'(S)$ sounds like $T(F)$.
  - $T(F)$ and $T(S)$ has the same meaning.
Our solution

Demo
Our solution

T(S): text in speaker's language S
Foreign language F → S-to-F Machine Translation → T(F): translation of T(S) → Our System → T'(S): phonetic rendering of T(F)

谢谢你  Thank you  三可 有
A collection of 1312 <Chinese, English, Chinglish> phrasebook tuples.  
1182 for training, 65 for development and 65 for test.

<table>
<thead>
<tr>
<th>Chinese</th>
<th>已经八点了吗</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>It’s eight o’clock now</td>
</tr>
<tr>
<td>Chinglish</td>
<td>意思埃特额克劳克闹 (yi si ai te e ke lao ke nao)</td>
</tr>
<tr>
<td>Chinese</td>
<td>这件衬衫又时髦又便宜</td>
</tr>
<tr>
<td>English</td>
<td>this shirt is very stylish and not very expensive</td>
</tr>
<tr>
<td>Chinglish</td>
<td>迪思舍特意思危锐思掉利失安的闹特危锐伊克思班西五</td>
</tr>
</tbody>
</table>

1 Dataset at http://www.isi.edu/natural-language/mt/chinglish-data.txt

Shi, X., Knight, K. and Ji, H. (USC & RPI) How to Speak a Language
<table>
<thead>
<tr>
<th>Frequency Rank</th>
<th>Chinese</th>
<th>Chinglish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>de</td>
<td>si</td>
</tr>
<tr>
<td>2</td>
<td>shi</td>
<td>te</td>
</tr>
<tr>
<td>3</td>
<td>yi</td>
<td>de</td>
</tr>
<tr>
<td>4</td>
<td>ji</td>
<td>yi</td>
</tr>
<tr>
<td>5</td>
<td>zhi</td>
<td>fu</td>
</tr>
</tbody>
</table>

**Table:** Top 5 frequent syllables in Chinese and Chinglish
Evaluation

T(C)
Foreign language (E)nglish

C-to-E Machine Translation → T(E)

Our System → T'(C)

R'(C):
Reference of Phonetic rendering

1

Chinese Speech Synthesizer
Evaluation

T(C) → C-to-E Machine Translation → T(E) → T'(C) → Chinese Speech Synthesizer

R'(C): Reference of Phonetic rendering

1

2

T(E,Human)

T(C)
Foreign language (E)nglish

Our System

Shi, X., Knight, K. and Ji, H. (USC & RPI) How to Speak a Language June 24, 2014 15 / 31
Model: Cascade FSTs

Chinese

MT

Eword

FST A

Epron

Pinyin-split

wFST B

Pinyin

FST C

Chinglish

谢谢

translate.google.com

Thank you

CMU Pron Dict
(Weide, 2007)

TH EY N K Y UW

s an k e y ou

Deterministic Rules

san ke you

Pron Dict

三可由
Model: Cascade FSTs

Chinese

MT

Eword

FST A

Epron

wFST B

Pinyin-split

FST C

Pinyin

wFST E

Chinglish

谢谢

Thank you

translate.google.com

CMU Pron Dict
(Weide, 2007)

TH E Y N K Y UW

Need to learn from data

s an k e y ou

Deterministic Rules

san ke you

Pron Dict

三可由
Construct <Epron, Pinyin-split> training pairs.

Mapping schema: 1-to-1, 1-to-2 and 2-to-1.

EM to learn parameters in wFST B, e.g. $P(\text{g e} | \text{g})$.

Viterbi alignments:

- grand
- 哥 软 的

$g \quad r \quad ae \quad n \quad d$
$g \quad e \quad r \quad uan \quad d \quad e$
### Table: Learned translation tables for the phoneme based model

| labeled Eppron | Pinyin-split | $P(p|e)$ |
|----------------|--------------|---------|
| d              | d            | 0.46    |
|                | d e          | 0.40    |
|                | d i          | 0.06    |
|                | s            | 0.01    |
| ao r           | u            | 0.26    |
|                | o            | 0.13    |
|                | ao           | 0.06    |
|                | ou           | 0.01    |
Alignment using phoneme-based model is fine.
When decoding test data, choices of target phonemes are context sensitive.

Decoding “grandmother”:

```
g r ae n d m ah dh er
g e r an d e m u e d e
```

reference Pinyin-split sequence:

```
g e r uan d e m a d e
```
Intuition: model the substitution of longer sequences \(^2\).

Viterbi alignment using Phoneme-based model:

\[
\begin{array}{cccccccc}
g & r & ae & n & d & m & ah & dh & er \\
g e & r & uan & d e & m & a & d & e \\
\end{array}
\]

Extract phoneme phrase pairs:

\[
g \rightarrow g \ e \\
g \ r \rightarrow g \ e \ r \\
\ldots \\
r \rightarrow r \\
r \ ae \ n \rightarrow r \ uan \\
\ldots
\]

\(^2\text{(Koehn et al., 2003)}\)
• Construct \(\langle\text{Eword, Pinyin}\rangle\) training pairs.
• Mapping schema: 1-to-[1,7].
• EM to learn parameters in wFST E, i.e. \(P(\text{nai te}|\text{night})\).
• Viterbi alignments:
  
<table>
<thead>
<tr>
<th>wake</th>
<th>up</th>
</tr>
</thead>
<tbody>
<tr>
<td>wei ke</td>
<td>a pu</td>
</tr>
</tbody>
</table>

• Error happen due to sparsity: “tips” and “ti pu si” only appear once.
  
<table>
<thead>
<tr>
<th>accept</th>
<th>tips</th>
</tr>
</thead>
<tbody>
<tr>
<td>a ke sha pu</td>
<td>te ti pu si</td>
</tr>
</tbody>
</table>
Hybrid training

- Intuition: Combine two models during **training** phrase.
- Use phoneme-based model to help word-based model:

  ![Diagram]

  - Errors are fixed:
    - accept
    - a ke sha pu te
    - tips
    - ti pu si
Hybrid decoding

- Intuition: Combine two models during decoding phrase.
<table>
<thead>
<tr>
<th>Language Type</th>
<th>Chinese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>等等我</td>
<td>wait for me</td>
</tr>
<tr>
<td>Reference English</td>
<td>唯特佛密(wei te fo mi)</td>
<td>wait for me</td>
</tr>
<tr>
<td>Reference Chinglish</td>
<td>唯特佛密(wei te fo mi)</td>
<td>wait for me</td>
</tr>
<tr>
<td>Hybrid Chinglish</td>
<td>位忒佛密(wei te fo mi)</td>
<td>wait for me</td>
</tr>
<tr>
<td>Human-dictated English</td>
<td>年年夜饭都要吃些什么</td>
<td>what do you have for the Reunion dinner</td>
</tr>
<tr>
<td>ASR English</td>
<td>我忒度优嗨佛佛得瑞优你恩低呢</td>
<td>what do you have for the reunion dinner</td>
</tr>
<tr>
<td>Human-dictated English</td>
<td>我忒度优嗨佛佛得瑞优你恩低呢</td>
<td>what do you have for the Reunion dinner</td>
</tr>
<tr>
<td>ASR English</td>
<td>我忒度优嗨佛佛得瑞优你恩低呢</td>
<td>what do you high for 43 Union Cena</td>
</tr>
</tbody>
</table>
## Experiments: English-to-Pinyin decoding accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>Coverage</th>
<th>Error Rate on covered text</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word based</td>
<td>29/65</td>
<td>0.042</td>
<td>0.664</td>
</tr>
<tr>
<td>Word-based hybrid training</td>
<td>29/65</td>
<td>0.029</td>
<td>0.659</td>
</tr>
<tr>
<td>Phoneme based</td>
<td>63/65</td>
<td>0.583</td>
<td>0.611</td>
</tr>
<tr>
<td>Phoneme-phrase based</td>
<td>63/65</td>
<td>0.136</td>
<td>0.194</td>
</tr>
<tr>
<td>Hybrid training/decoding</td>
<td>63/65</td>
<td>0.115</td>
<td>0.175</td>
</tr>
</tbody>
</table>

![Diagram of the process]

### Diagram Description:
- **T(C)**: Reference of Phonetic rendering
- **T(E)**: Foreign language (English) Translation
- **C-to-E Machine Translation**: Converts English to Chinese
- **Our System**: Processes the translated Chinese
- **T'(C)**: Chinese Speech Synthesizer

---

## Experiments: Human Dictation Accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>Error Rate vs. reference English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictation from Reference Chinglish</td>
<td>0.477</td>
</tr>
<tr>
<td>Phoneme based</td>
<td>0.696</td>
</tr>
<tr>
<td>Hybrid training and decoding</td>
<td>0.496</td>
</tr>
</tbody>
</table>

![Diagram showing the process of human dictation accuracy]

## Experiments: No Human in the Loop

<table>
<thead>
<tr>
<th>Model</th>
<th>Error Rate vs. reference English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word based</td>
<td>0.925</td>
</tr>
<tr>
<td>Word-based hybrid training</td>
<td>0.925</td>
</tr>
<tr>
<td>Phoneme based</td>
<td>0.937</td>
</tr>
<tr>
<td>Phoneme-phrase based</td>
<td>0.896</td>
</tr>
<tr>
<td>Hybrid training and decoding</td>
<td>0.898</td>
</tr>
</tbody>
</table>

![Diagram of the system flow](image-url)
Conclusion

- **Goal:** Help people speak foreign languages
  - Provide native phonetic spellings that approximate the sounds of foreign phrases
  - Use a cascade of FSTs
  - Improve the model by adding phrases and combining models in both training and decoding phase

For future:

- More Language Pairs
Thank you! & QA

Demo: http:/\cage.isi.edu:8080