CHIME: An Efficient Error-Tolerant Chinese Pinyin Input Method

Yabin Zheng¹, Chen Li², and Maosong Sun¹

¹Tsinghua University
²University of California, Irvine
Outline

• Introduction
• Related Work
• Correcting a Single Pinyin
  – Finding Similar Pinyins
  – Ranking Similar Pinyins
• Converting Pinyin Sequences to Chinese Words
  – Pinyin-to-Chinese Conversion without Typos
  – Pinyin-to-Chinese Conversion with typos
• Experiments
• Conclusions
Introduction

- What is Chinese Pinyin input method
  - Users cannot type in Chinese characters directly
  - Pinyin input methods are proposed

- Users mentally generate a Chinese word “上海”
- Type in corresponding Pinyin “shanghai”
- Input methods display words with this pronunciation

shanghai
1. 上海  2. 伤害  3. 上还  4. 商海  5. 上
Introduction (cont.)

- A beginner of Chinese language
  - 篮球 (basketball), lanqiu or lanchiu
- Users in southern China
  - 开花 (bloom), kaihua or kaifa
Introduction (cont.)

- Users may make typos when typing Pinyins

- Users have to identify and correct typos

- We need error-tolerant Pinyin input method
Introduction (cont.)

• Two challenges in developing “CHIME” (CHinese Input Method with Errors)
  • Accuracy

和  hé  hè  hú  huo  huó  huò

人  rén

• Efficiency
Correcting a Single Pinyin

- Pinyin dictionary $D$, an input Pinyin $p$ that is not in $D$
- Find a set of similar candidate Pinyins $w \in D$
  - Similarity measure: edit distance
  - Empirically keep top-3 candidate Pinyins

\[ p = \text{sanghaai} \]
Finding Similar Pinyins

- Efficient similarity search
  - State-of-the-art Index structure and search algorithm (Ji et al., 2009)
  - woemng gounai le sanghaai shengchang de niulai

<table>
<thead>
<tr>
<th>Input Pinyin</th>
<th>Similar Candidate Pinyins</th>
</tr>
</thead>
<tbody>
<tr>
<td>woemng</td>
<td>women, weng, wodang</td>
</tr>
<tr>
<td>gounai</td>
<td>goumai, dounai, guonei</td>
</tr>
<tr>
<td>le</td>
<td>le</td>
</tr>
<tr>
<td>sanghaai</td>
<td>shanghai, canghai, wanghuai</td>
</tr>
<tr>
<td>shengchang</td>
<td>shengchan, zhengchang, shangchang</td>
</tr>
<tr>
<td>de</td>
<td>de</td>
</tr>
<tr>
<td>niulai</td>
<td>niunai, niupai, niuli</td>
</tr>
</tbody>
</table>
Ranking Similar Pinyinns

- Given a mistyped Pinyin $p$, rank candidate $p'$ using $Pr(p'|p)$
  - Noisy channel error model
    \[
    Pr(p'|p) = \frac{Pr(p|p')Pr(p')}{Pr(p)} \propto Pr(p|p')Pr(p')
    \]

- Estimate conditional probability
  \[
  Pr(sanghaai|shanghai) = Pr('h'->'~')Pr('~'->'a')
  \]
Pinyin-to-Chinese Conversion without Typos

- Convert a Pinyin sequence $P = p_1 p_2 \ldots p_k$ to the most probable sequence of Chinese word $W = w_1 w_2 \ldots w_k$

$$
\hat{W} = \arg \max_w Pr(W | P) \\
= \arg \max_w \frac{Pr(W)Pr(P | W)}{Pr(P)} \\
= \arg \max_w Pr(W)Pr(P | W) \\
= \arg \max_w Pr(W) \prod_i Pr(p_i | w_i)
$$

- $Pr(W)$ is estimated using a bigram language model

$$Pr(W) = Pr(w_1) Pr(w_2 | w_1) Pr(w_3 | w_2) \ldots Pr(w_n | w_{n-1})$$
Pinyin-to-Chinese Conversion with Typos

- $P = p_1 p_2 \ldots p_k$ ($P$ have typos), $P'$ denotes the correct Pinyin sequence
- Given $P'$, Pinyin sequence $P$ and word sequence $W$ are conditionally independent

$$
\hat{W} = \arg \max_w Pr(W \mid P) \\
= \arg \max_w \sum_{P'} Pr(P' \mid P) Pr(W \mid P') \\
= \arg \max_w \sum_{P'} \frac{Pr(W) Pr(P' \mid P) Pr(P' \mid W)}{Pr(P')} \\
= \arg \max_w \sum_{P'} Pr(W) Pr(P \mid P') Pr(P' \mid W) \\
= \arg \max_w \sum_{P'} Pr(W) \prod_i Pr(p_i \mid p_i') Pr(p_i', w_i).
$$

$Pr(W) = Pr(w_1) Pr(w_2 \mid w_1) Pr(w_3 \mid w_2) \ldots Pr(w_n \mid w_{n-1})$

$Pr(p \mid p') = \prod_{e \in T} Pr(e)$
Framework of CHIME

- Correct mistyped Pinyin in the Pinyin sequence
- Convert corrected Pinyin sequence to Chinese words
Experimental Settings

- Sun-Pinyin software
  - Pinyin dictionary and language model
  - 104,833 Chinese words and 66,797 Pinyins
- Lancaster corpus (McEnery and Xiao, 2004)
  - Five native-speakers type in 2,000 sentences for evaluation
  - 679 sentences (34%) contain one or more typos
  - 885 typos are collected in total
- Computer with AMD Core2 2.20GHz CPU and 4GB memory, C++ compiled with a GNU compiler
Probabilities of Edit Operations

- $Pr(e)$ is not uniformly distributed
  
  $Pr(p | p') = \prod_{e \in T} Pr(e)$

- $Pr(‘z’->‘s’) > Pr(‘z’->‘p’)$
  - ‘z’ and ‘s’ are adjacent on the keyboard
  - ‘z’ and ‘s’ pronounce similarly

- Heuristic rules based on Chinese-specific features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Example pairs of similar Pinyin letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front and back nasal sound</td>
<td>‘ang’ - ‘an’, ‘ing’ - ‘in’, ‘eng’ - ‘en’</td>
</tr>
<tr>
<td>Retroflex and blade-alveolar</td>
<td>‘zh’ - ‘z’, ‘sh’ - ‘s’, ‘ch’ - ‘c’</td>
</tr>
<tr>
<td>Letters with similar pronunciations</td>
<td>‘z’ (茲) - ‘c’ (词) - ‘s’ (丝), ‘n’ (呢) - ‘l’ (勒), ‘b’ (播) – ‘p’ (泼)</td>
</tr>
</tbody>
</table>
Evaluation Metrics

- \( E_1 \): A mistyped Pinyin is not detected, Detection error rate \( \text{DER} = \frac{E_1}{T} \)
- \( E_2 \): A mistyped Pinyin is not suggested to the correct Pinyin, Correction error rate \( \text{CorrER} = \frac{E_2}{T} \)
- \( E_3 \): A mistyped Pinyin is not converted to the correct Chinese word, Conversion error rate \( \text{ConvER} = \frac{E_3}{T} \)

Commercial software Sogou-Pinyin for comparison

<table>
<thead>
<tr>
<th>Metric</th>
<th>DER</th>
<th>CorrER</th>
<th>ConvER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHIME</td>
<td>37.40%</td>
<td>52.43%</td>
<td>53.56%</td>
</tr>
<tr>
<td>Sogou</td>
<td>70.62%</td>
<td>91.19%</td>
<td>91.75%</td>
</tr>
</tbody>
</table>
Efficiency Evaluation

- Average processing time: 12.9ms/sentence
- Processing time decreases with more letters typed in
- Additional processing time of 4.97ms for CHIME
**Saved Typing Efforts**

- CHIME can return Chinese words before users type in a complete Pinyin sequence

<table>
<thead>
<tr>
<th>Original Pinyin Sequence</th>
<th>Actual Pinyin Sequence</th>
<th>Converted Chinese Words</th>
<th>Saved Typing Efforts</th>
</tr>
</thead>
<tbody>
<tr>
<td>woemng gounai le sanghaai shengchang de niulai</td>
<td>woem gouna l sangh shengc d niulai</td>
<td>我们 购买了 上海 生产 的 牛奶</td>
<td>26.1%</td>
</tr>
<tr>
<td>zaichang oizhe fenfen juqi shexiangj</td>
<td>zaicha oiz fenfe juqi shexiangj</td>
<td>在场 记者 纷纷 举起 摄像机</td>
<td>16.2%</td>
</tr>
<tr>
<td>zhuajin richang shenghou de uifu gongzuo</td>
<td>zhuaji rich shengh d uifu gongz</td>
<td>抓紧 日常 生活 的 恢复 工作</td>
<td>22.5%</td>
</tr>
<tr>
<td>yixie caidtan changjina gfennu le</td>
<td>yix caidt changj gfennu l</td>
<td>一些 彩电 厂家 愤怒 了</td>
<td>24.2%</td>
</tr>
<tr>
<td>quanguo xiyhiji shohoufuwu youxiu changj</td>
<td>quang xiyhiji shohouf youxiu changj</td>
<td>全国 洗衣机 售后服务 优秀 厂家</td>
<td>16.7%</td>
</tr>
<tr>
<td>shenchanguosheng de duo suhy laodongmimixing chanpin</td>
<td>shenchanguos d duo suhy laodongmimi chap</td>
<td>生产过剩 的 多属于 劳动密集型 产品</td>
<td>22.6%</td>
</tr>
<tr>
<td>shangpingjingji qianglie fuhuan kexuejishu de zhichi</td>
<td>shangpingji qiangl fuhuan kexuejish d zhic</td>
<td>商品经济 强烈 呼唤 科学技术 的 支持</td>
<td>19.2%</td>
</tr>
<tr>
<td>tigao canping zhiliang yu guanni suiping</td>
<td>tiga canp zhilia yu guanni suipi</td>
<td>提高 产品质量 与 管理 水平</td>
<td>20.0%</td>
</tr>
<tr>
<td>jinnialai sulian gounei de gezhong maodun zhujian tuchu</td>
<td>jinniala sulian goune d gezh maod zhujian tuc</td>
<td>近年来 苏联 国内 的 各种 矛盾 逐渐 突出</td>
<td>18.2%</td>
</tr>
<tr>
<td>changchtunshi chengshi guihua guanli tiaoil</td>
<td>changchtunsh chengs guih guanl tiaoil</td>
<td>长春市 城市 规划 管理 条理</td>
<td>14.0%</td>
</tr>
</tbody>
</table>
Related Work

- Pinyin-to-Chinese conversion
  - Statistical segmentation and language model based approach [Chen and Lee, 2000]
    - They only correct single-character errors
  - Extract Chinese Pinyin names from English text and suggest corresponding Chinese characters [Kwok and Deng, 2002]
    - They only convert Pinyin names to Chinese characters
  - Commercial Pinyin input methods use rule-based approaches to handle typos
Related Work (cont.)

- English spelling corrections
  - Noisy channel models based on generic string-to-string edit operations (Brill and Moore, 2000)
  - Pronunciation information is useful for English spelling correction (Toutanova and Moore, 2002)
  - Query log and click-through data in English spelling correction (Cucerzan and Brill, 2004; Sun et al., 2010; Whitelaw et al., 2009)
  - These methods are not directly applicable to the Chinese language
Conclusion and Future Work

**Conclusion**
- Error-tolerant features are important for Chinese Pinyin input method
- CHIME finds similar Pinyins for a mistyped Pinyin and ranks candidate Pinyins using language-specific features
- CHIME detects and corrects Pinyin sequence, and finds most likely sequence of Chinese words
- CHIME achieves both a high accuracy and efficiency

**Future Work**
- Correct a mistyped Pinyin that included in the Pinyin dictionary
- Support acronym Pinyin input (e.g. “zg” for “中国”)
Reference

Reference (cont.)

Thanks & QA