Finding Chinese Personal Names Automatically in Unrestricted Texts

Maosong SUN (Tsinghua University, Beijing)
Lawrence CHEUNG (CityU, HK)
1. What Chinese Personal Names Look Like
## Structure of Chinese Personal Names

<table>
<thead>
<tr>
<th>Full Name</th>
<th>Husband's Surname</th>
<th>Surname</th>
<th>Given Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H1 ( (H2) )</td>
<td>S1 ( (S2) ) + G1 ( (G2) )</td>
<td></td>
</tr>
<tr>
<td>1 李鹏</td>
<td></td>
<td>李</td>
<td>鹏</td>
</tr>
<tr>
<td>2 邓小平</td>
<td></td>
<td>邓</td>
<td>小 平</td>
</tr>
<tr>
<td>3 诸葛亮</td>
<td></td>
<td>诸 葛</td>
<td>亮</td>
</tr>
<tr>
<td>4 东方闻樱</td>
<td></td>
<td>东 方</td>
<td>闻 樱</td>
</tr>
<tr>
<td>5 陈方安生</td>
<td>陈</td>
<td>方</td>
<td>安 生</td>
</tr>
<tr>
<td>6 诸葛东方闻樱</td>
<td>诸 葛</td>
<td>东 方</td>
<td>闻 樱</td>
</tr>
</tbody>
</table>

- Minimal length: 2 Chinese characters;
- Maximal length: 6 Chinese characters;
- Frequent length: 2-3 Chinese characters.
2. The Role of Automatic Identification of Chinese Personal Names in Information Technology
Response of Sina( http://sina.com.cn), the most popular Chinese Internet search engine, to the query “中将” (lieutenant general):
The response of Sina (regarding Chinese personal names) to the query “中将” (lieutenant general):

快讯：陈中将与俄罗斯选手争夺跆拳道女子67公斤以上级金牌

黄健中将与奥利弗-斯通合作 中美合拍《赛金花》

Irrelevant results are retrieved due to the pretty poor performance of Sina search engine in processing Chinese personal names.
Text-to-Speech Conversion

Bell Labs Mandarin
Text-to-Speech Synthesis

http://www.bell-labs.com/project/tts/mandarin-gb.html

我的老板查金泰不同意他弟弟查建国先生的看法。

My boss Zha Jin-Tai did not agree to the opinion of his younger brother, Mr Zha Jian-Guo.

华国锋曾任中华人民共和国国务院总理。

Hua Guo-Feng is the former premier of the People’s Republic of China.
I see that Deng Xiao-Ping greets with Jiang Ze-Min.

I see week star Chi open together study friend greet.
3. The Characteristics of Chinese Personal Names
No Explicit Marks as Capital Initial in English and with too Many Potential Ambiguities!

- Character sets for surname and given name are strictly a subset of Chinese character set, and are decentralized to some extent.
- Some characters in the above-said sets may be mono-syllabic words.
  钱玉钦睡觉。
  *Qian Yu-Qin sleeps.*
  钱玉爱睡觉。
  *Qian Yu-Ai sleeps.*  *Qian Yu likes to sleep.*
- The mono-syllabic words in the above-said sets could be either content words or function words.
- Some multi-syllabic words can be involved in Chinese personal names starting at every possible position.
  王朝闻(dynasty)  马胜利(victory)  严肃(serious)
Why the Task Difficult? -- An Example at Extreme

A couplet (对联) in ancient China:
鱼游石孔秋江冷
柏成林丛夏岳高
• Every character can be mono-syllabic common word.
  fish swim stone aperture Autumn river cold
cypress constitute forest crowd Summer mountain high
Fish swim within stone apertures of the cold river in Autumn;
and cypresses constitute the crowsed forest, the mountain in
Summer is so high.
• Every character can be used as both surname and given name.
  Totally 22 possible candidates of Chinese personal names
  (without considering the case of the husband’s surname)!
  鱼游, 鱼游石, 游石, 游石孔, 石孔, 石孔秋,
  孔秋, 孔秋江, ..., 丛夏, 丛夏岳, 夏岳, 夏岳高, 岳高
4. Algorithm for Automatic Identification of Chinese Personal Names
Knowledge Exploited in the Algorithm

Type 1: Internal information:
- Probability of being a Chinese personal name
  \[ \log(\text{Prob}(吕欽)) = -5.019326 \quad >> \]
  \[ \log(\text{Prob}(和广)) = -6.200005 \]
- Nature of characters
  李逵 郑筱云 刘景藜
- Construction
duplication: “媛媛”“强强”“白毛”“潇洒”

Type 2: External information:
- Local context
title: 博士(Dr.) 教授(professor)
- Special pattern
  “以 < CN> {title1} 为 <title2> ”
  以潘杜泉为团长的香港工会代表团, ...
Chinese Personal Name Identification as a Part of Chinese Word Segmentation System

CSeg&Tag1.1 DEMO

Sample input:
陈中昨天去天津了。
王美丽真生气了。

Sample output:
陈中/nr 昨天/t 去/v 天津/ns 了/u 。/w
王美丽/nr 真/d 生气/a 了/y 。/w
5. Continue ...

The next part of this talk:
by Lawrence CHEUNG (CityU, HK)