Pacific Neighborhood Consortium (PNC)
City University of Hong Kong
January 15 - 20, 2001

香港城市大學
City University
of Hong Kong

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

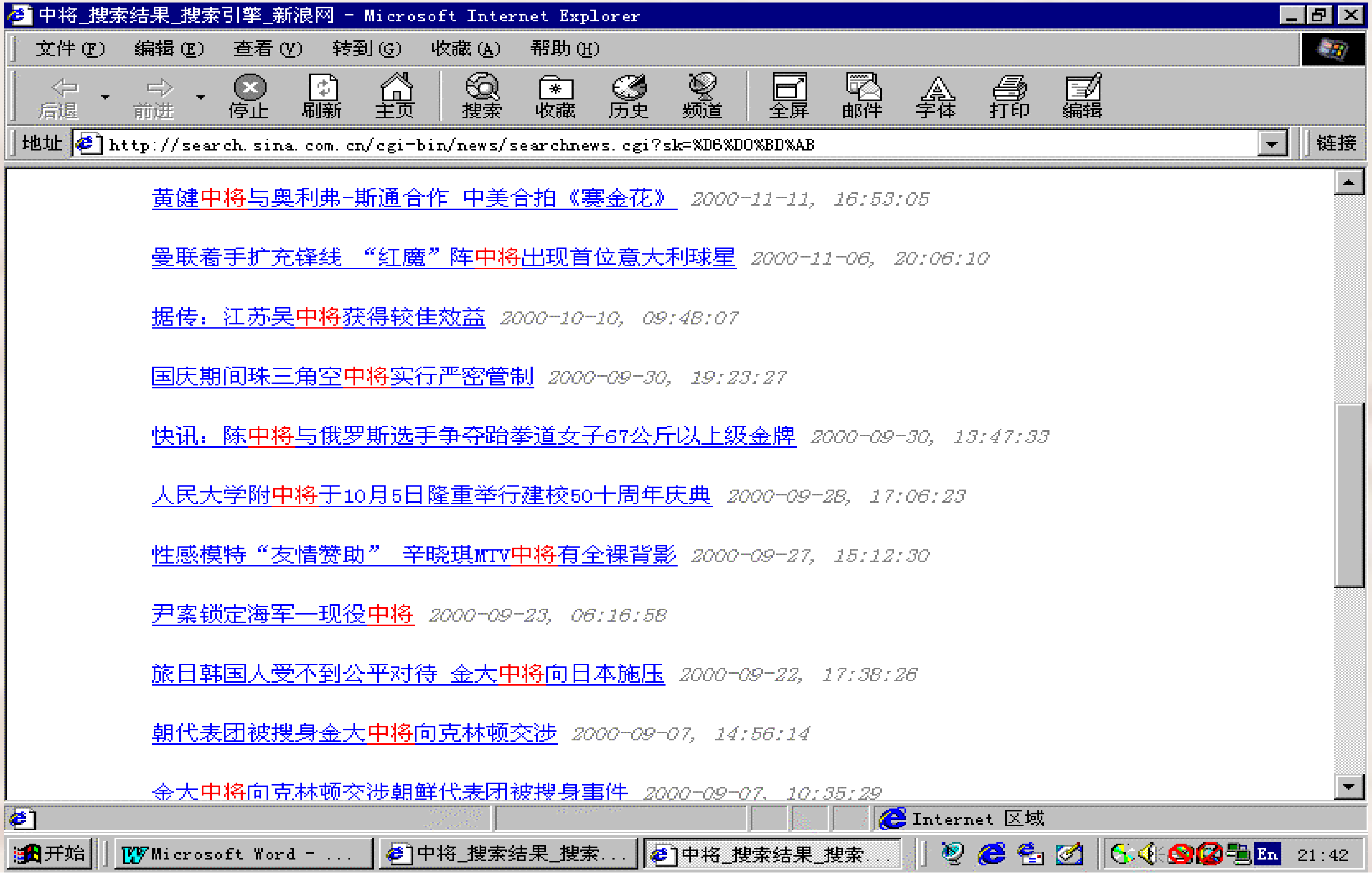
	Full Name	Husband's Surname			Surname			Given Name	
		H1	(H2)	+	S1	(S2)	+	G1	(G2)
1	李鹏				李			鹏	
2	邓小平				邓			小	平
3	诸葛亮				诸	葛		亮	
4	东方闻樱				东	方		闻	樱
5	陈方安生	陈			方			安	生
6	诸葛东方闻樱	诸	葛		东	方		闻	樱

- 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

Information Retrieval

Response of Sina(http://sina.com.cn), the most popular Chinese Internet search engine, to the query “中将”(lieutenant general):



中将_搜索结果_搜索引擎_新浪网 - Microsoft Internet Explorer
文件(F) 编辑(E) 查看(V) 转到(G) 收藏(A) 帮助(H)
后退 前进 停止 刷新 主页 搜索 收藏 历史 频道 全屏 邮件 字体 打印 编辑
地址
链接

http://search.sina.com.cn/cgi-bin/news/searchnews.cgi?sk=%D6%D0%BD%AB

黄健中将与奥利弗-斯通合作 中美合拍《寨金花》 2000-11-11, 16:53:05

曼联着手扩充锋线 “红魔”阵中将出现首位意大利球星 2000-11-06, 20:06:10

据传：江苏吴中将获得较佳效益 2000-10-10, 09:48:07

国庆期间珠三角空中将实行严密管制 2000-09-30, 19:23:27

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 2000-09-30, 13:47:33

人民大学附中将于10月5日隆重举行建校50十周年庆典 2000-09-28, 17:06:23

性感模特“友情赞助” 辛晓琪MTV中将有全裸背影 2000-09-27, 15:12:30

尹案锁定海军一现役中将 2000-09-23, 06:16:58

旅日韩国人受不到公平对待 金大中将向日本施压 2000-09-22, 17:38:26

朝代表团被搜身金大中将向克林顿交涉 2000-09-07, 14:56:14

金大中将向克林顿交涉朝鲜代表团被搜身事件 2000-09-07, 10:35:29

Information Retrieval(Continue)

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

Lucent Technologies
Bell Labs Innovations

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。

Zha Zha(Cha)

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。

Hua4 Ceng(Zeng) Hua1

Hua Guo-Feng is the former premier of the People's Republic of China.

Chinese to English Machine Translation

transtar 泽星网

泽星科技 沟通无限

http://www.transtar.com.cn/transtar/chinese/netbar/onlinetrans.asp

我看见邓小平同江泽民打招呼。

Transtar: *I see that Deng Xiao-Ping greets with Jiang Ze-Min.*

我看见周星驰同张学友打招呼。

Transtar: *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 crowed 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

Type1: Internal information:

- Probability of being a Chinese personal name

  lg(*Prob*(吕钦))= -5.019326 >>

  lg(*Prob*(和广))= -6.200005

- Nature of characters

  李逵 郑筱云 刘景藜

- Construction

  duplication: “媛媛”“强强”“毛毛”“潇潇”

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

Input

Output

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)