Genealogical Record Linkage on International Data

Randy Wilson
FamilySearch.org
wilsonr@ldschurch.org

March 13, 2008

Record Linkage

- The process of identifying pairs of records that refer to the same thing.
- Used in medicine (Dunn, 1946), advertising, business, government, and genealogy.
- Most popular approach has been **Probabilistic Record Linkage** (Newcombe et al., 1959; Fellegi & Sunter, 1969).

Traditional Record Linkage

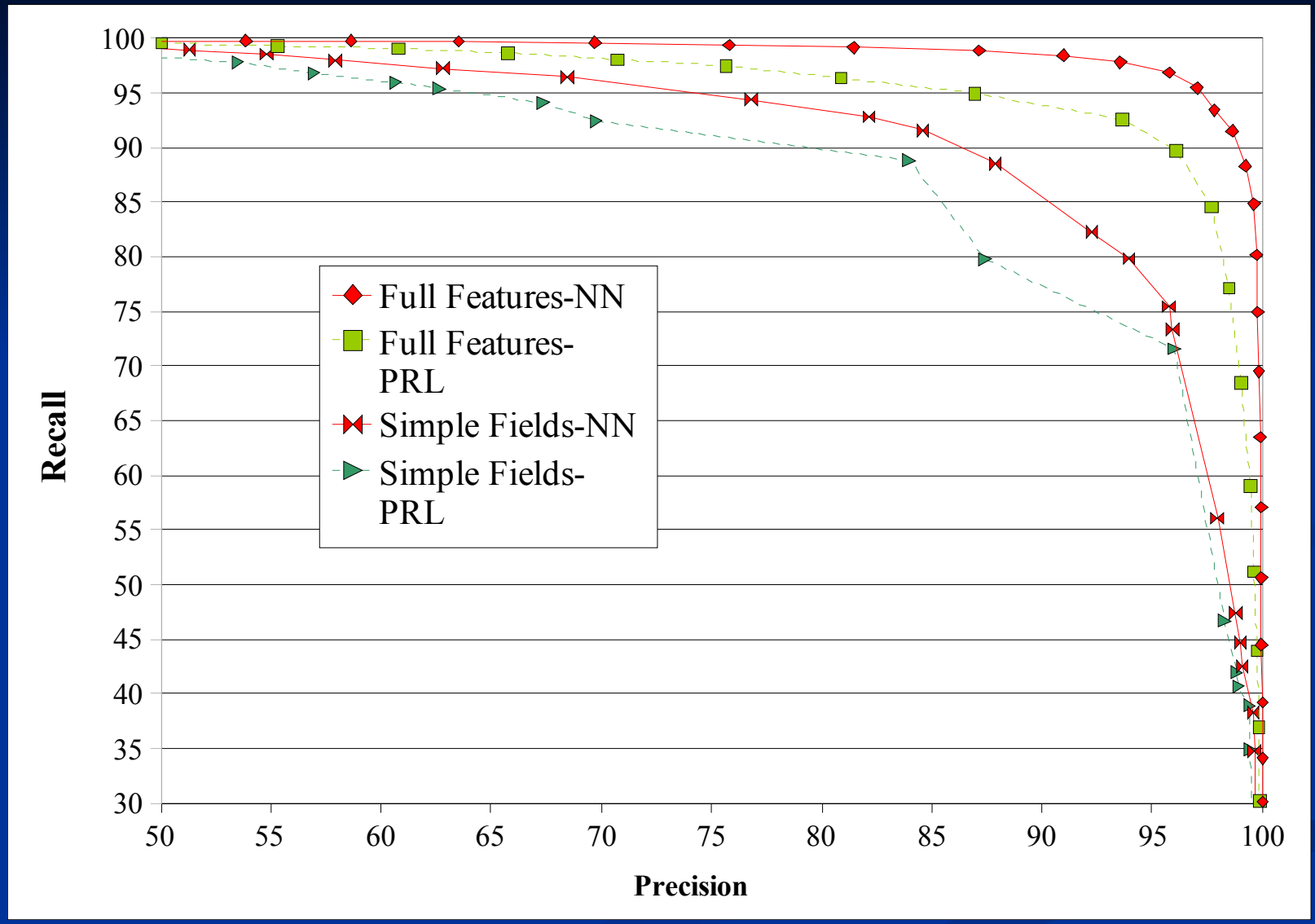
- Probabilistic Record Linkage
  - Simple field agreement/disagreement weights
    - Given name agrees: +3.7
    - Given name disagrees: -2.5
    - Mother's surname agrees: +1.7
    - Etc.
  - Probabilistic formulas for calculating weights
    - Equivalent to "Naïve Bayes" classifier.

Better Record Linkage

- Complex features
  - GivenName=-1: -2.1132 (Conflicting name)
  - GivenName=1: 0.5125 (Initials match, or names are somewhat similar)
  - GivenName=2: 1.083
  - GivenName=3: 0.2176
  - GivenName=4: 1.5075
  - GivenName=5: 0.6978
  - GivenName=6: 1.1639
  - GivenName=7: 1.084 (Multiple agreeing name pieces)
  - DiedBeforeBorn: -12.03 (One person died before the other was born)
- Neural network training

Empirical Results

- 80,000 labeled pairs of genealogical records.
- Measuring accuracy:
  - Threshold=score above which a pair is considered to be a “match” by the algorithm.
  - Recall = percent of known matching records that are above a threshold.
  - Precision = percent of records above a threshold that are really a match.


Recall
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
50
55
60
65
70
75
80
85
90
95
100
Precision
Full Features-NN
Full Features-PRL
Simple Fields-NN
Simple Fields-PRL

Data Variation

- **Names**: Same real person can have different names in records, due to:
  - Nicknames (“Bob” vs. “Robert”)
  - Maiden vs. married names (“Elizabeth Turner” vs. “Elizabeth Smith”)
  - Spelling variations (“Elizabeth” vs. “Elisabeth”)
  - Initials (“John Henry” vs. “John H.”)
  - Typographical errors (“John” vs. “Jhon”)
  - Illegible handwriting
  - Noise words (“Mrs.”, “infant”, “or”)

Data Variation

- **Dates**:
  - Formatting differences (“12 Jun 1850”, “6/12/1850”, “1850.12.6”)
  - Estimates (“1850”, “about 1848”)
  - Typographical errors (“1701” vs. “1710”)
  - Calendar changes (1782/1783)
- **Places**:
  - **Abbreviation: (“SLC”, “Salt Lake, UT”, “Salt Lake City, Salt Lake, Utah, USA”)**
  - **Different levels (“Utah” vs. “Provo, UT”)**
  - **Place name changes: (“Istanbul”, “Constantinople”)**
  - **Boundary changes over time**

Reducing variation

- Normalization
  - Lower case
  - Remove punctuation
  - Tokenize
- Standardization
  - Name group ID
  - Standard dates in common calendar
  - Place ID

Challenges of International Data

- **Scripts** (Unicode); **Name order**; **Delimeters**
  - **Chinese**
    - Chinese: 黃 德纘
    - Romanized: Huang Te-Tsuan
  - **Japanese**
    - Kanji [Chinese]: 鈴木 栄吉
    - Katakana: スズキ　エイキチ
    - Hiragana: すずき　えいきち
    - Romanized: Suzuki Eikichi
  - **Korean**
    - Hanja [Chinese]: 金 聲繡
    - Hangul: 김 성수
    - Romanized: Kim Seong-su
  - **Cyrillic** (Russian)
    - Cyrillic: Иван Овсянников
    - Romanized: Ivan Ovsyannikov

Handling “Mrs.”

- “Mrs. John Smith” vs. “Elizabeth Turner”
- **Chinese**:
  - <Husband's name>“fu-ren” (夫人)
- **Japanese**:
  - <Husband's surname> 夫人 ; or
  - <Husband's surname> 夫人 <given in kana>
- **Korean**:
  - <Husband's name>“ui buin” (“ 의 부인”)
- **Cyrillic**:
  - <Husband's surname> Г-жа (“Госпожа”=*gospozha*)

Handling “Mrs.” in CJKC

Language	Husband s name	Wife s name
1. English Mrs. [<given>] <surname>	John Smith	Mrs. Smith; or Mrs. John Smith
2. Chinese	黃 德纘 (Huang Te-Tsuan)	黃 德纘夫人 (Huang Te-Tsuan; Mrs.)
3. Japanese <surname>夫人	鈴木 栄吉 (Suzuki Eikichi)	鈴木夫人 (Mrs. Suzuki)
4. Japanese <surname>夫人<given>	鈴木 栄吉 (Suzuki; Eikichi)	鈴木夫人 かの (Suzuki; Mrs.; Kano)
5. Korean <surname given> 의 부인	김 성수 (Kim Seong-Su)	김 성수의 부인 (Kim Seong-Su s wife)
6. Cyrillic	Иван Овсянников Ivan Ovsyannikov	Госпожа Ивана Овсянникова or: Г-жа Ивана Овсянникова (Mrs. Ivana Ovsyannikova)

Handling Mr./Miss

- Chinese
  - <surname> “ssi” (氏)
- Korean
  - <surname> “ssi” (씨)
    - Hanja: 金 氏
    - Hangul: 김 씨
    - Romanized: Kim Ssi
- Japanese
  - <surname> daughter (娘):
    - 鈴木娘 (“Miss Suzuki”)
  - <surname> son (息子)

Patronymic Names

- Scandinavian countries (Sweden, Norway, Denmark, etc.)
- Children of Olaf Svensen:
  - Johan Olafsen
  - Inga Olafsdotter
- Variations
  - Inga Olafsdtr/Olafssen/Olafson/...
  - Inga Svensen (from father's “surname”)
- So map “son/daughter” stems when surname doesn't exactly match.

Cyrillic Patronymic Names

- Varies by country
- Middle name = father's given name + suffix
  - e.g., “-vich” for males
  - “-yevna”, “-ovna”, “ichna”, etc., for females
- Example:
  - Son: Sergey (Сергей)
  - Father: Ivan Popov (Иван Попов)
  - => Sergey Ivanovich Popov
  - (Сергей Иванович Попов)

Parsing Asian Places

- Often no spaces
- General on left; specific on right.
- Guangdong, China: 中國廣東省 .
  - (中國 =China; 廣東省 =Guangdong)
- Do catalog search to find places:
  - Parse left-to-right, looking up substrings in catalog.
  - Recurse with remaining string.
  - Require later strings to be “within” earlier string in catalog. (e.g., ignore “Guangdong, Taiwan” if “China” was already parsed).

Observations

- 877 pairs from 1.6M Chinese records
  - 22% of pairs affected by “Mrs.” or “Miss” issue
  - 16% of records named “Mrs....”; 24% “Miss”
- 5,000 pairs from 1.5M Korean records
  - almost half affected by “Mrs.” or “Miss” issue
  - 40% named “...ssi”; 10% named “wife of...”
- 4,700 pairs from 4M Japanese records
  - 12% affected by “Mrs.” issue
- Asian place names:
  - On small sample, none parsed originally
  - All parsed (as far as catalog went) with new method

Remaining Issues

- Refinement
  - CJKC: Re-evaluate with more reasonable matches
- Additional cultures
  - Africa: Patriarchal naming:
    - Mohamed Takal Yogol [...], son of
    - Takal Yogol [...], son of
    - Yogol [...]
  - Middle East
  - Latin multiple surnames
  - Tribe / Clan (Africa, Korea, ...)

감사합니다 [*kamsamnida*]

спасибо [*spasibo*]

Gracias

ありがとう [*arigato*]

Obrigado

謝謝 [*xièxie*]

Grazie

Thank you