Nationality Classification using Name Embeddings

Junting Ye, Stony Brook University

Joint work with Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, Steven Skiena
Outline

• Background
  – Problem Definition
  – Challenges
  – Related Work
• Name Embedding
• NamePrism
  – Nationality Classifier
  – Ethnicity Classifier
• Experiment
• Application
Background

• Problem Definition
  – **Input**: Name (first and last names)
  – **Output**: Nationality *(See demo)*

• Why important?
  – Ads/News/posts **recommendations**;
  – **Sociology** applications;
  – **Biomedical** applications;
Challenges

- Common last names among nationalities
  - **Lee**: British & Chinese
  - **Roy**: British & Indian
  - More when considering ~40 nationalities

- Complexity in naming
  - **Marriage**: Angelina Jolie --> Angelina Pitt
  - **Immigration**: Neta-Lee Hershlag --> Natalie Portman
  - **Colonization**: Jacky Chan, born in HK
  - **Slavery**: change both
Related Work

• Binary classifiers:
  – Hispanic [Buechley 1976]
  – Chinese [Coldman, Braun and Gal]
  – Indian [Harding, Dews and Sir]

• Multi-class/Hierarchical:
  – Decision & HMM, subs [Treetapituk and Giles 2012] http://www.textmap.com
  – Logistic Regression, subs [Treeatpituk and Giles 2012] http://singularity.ist.psu.edu/ethnicity
  – kNN & Instance based [Torvik and Agarwal 2016] http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py

Name Ethnicity Classifier

About
The ethnicity classifier was developed at the Computer Science Department, Stony Brook University, under the guidance of Prof. Steven Skiena. In this project, we tried to aggregate entities into different ethnic groups. We formed a hierarchical structure of ethnicities and then used our classifier to predict the ethnicity of a given entity at different levels of this decision tree. The training data was obtained from Wikipedia.

This Web-UI allows the user to enter a name or a set of names which can then be classified into appropriate ethnic groups by our classifier model.

Ethnie -- the ethnicity predictor (for bibliographic records)

First= Bo Last= Svensson Submit

Prediction (cutoff = 90%) = NORDIC
Prediction (cutoff = 60%) = NORDIC

| Ethnicity  | Prob | First | Last | probF | probL |
|------------|------|-------|------|-------|-------|
| NORDIC     | 99.24| Bo    | Svensson | 24.807| 99.765|
| CHINESE    | 0.72 | Bo    | Svensson | 73.27 | 0.0   |
| KOREAN     | 0.02 | Bo    | Svensson | 1.923 | 0.0   |
| ITALIAN    | 0.01 | Bo    | Svensson | 0.0   | 0.235 |

How did we do it? A combination of temporal predictions of author names in PubMed for: Bo and Svensson

Compare it to another ethnicity predictor: EthnicSeer

Compare it to an approximate lastname distribution worldwide: Forebears

While you're at it you might as well check out Genni -- the temporal gender predictor says Bo Svensson is Male

If you want json instead of html, just append &format=json to the url but if you plan to harvest our data, please be a good netizen -- let me know (vtorvik@illinois.edu) and pause your program at least a second between requests.
Related Work

• Limits:
  – Small training set from Wikipedia
  – Coarse ethnicity/nationality taxonomy
  – Substring features, not working on logograms
  – Existing systems only for latinized names
**Name Embedding**

- **Word Embedding**
  - Vector representation for *words*
  - *Co-occur* within context/moving window in *articles*
  - Similar words have similar embedding

Do not go gentle into that good night, Old age should burn and rave at close of day; Rage, rage against the dying of the light.

- **good**: 0.83, -0.14, -0.12, 0.33 ...
- **gentle**: 0.35, -0.09, -0.04, 0.59 ...
- **rage**: -0.21, 0.01, 0.02, -0.48 ...
Name Embedding

- **Name Embedding**
  - Vector representation for name parts
  - Co-occur within context/moving window in sorted contact lists (*homophily principle*)
  - Similar name parts have similar embedding

Gerda_Zavada@ Roxana Carmen, Adina Margine, Radoi Seicaru, Drînd Ramona, Cristian Usurelu, ...
Chilap_ja@ leung Ja, Chow Iris, Ken Ja, Betty Cheung, Chan Stone, Donna Tang, Ricky Lap, ...
balbirsingh@ Krishan Singh, Neeraj Kumar, Pankaj Bawa, Vijay Kumar, Jaspinder Kumar, ...

roxana@0 -0.089013, 0.382334, 0.068602, -0.670069, 0.289275, 0.721922, -0.042753, ...
carmen@1 0.256830, 0.212766, -0.282459, -0.615443, -0.107637, 0.348170, 0.082246, ...
adina@0 0.324114, 0.261639, -0.144281, 0.143037, 0.101616, 0.316141, 0.426601, ...
margine@1 0.215786, -0.109732, -0.790512, 0.440265, -0.247357, -0.341169, 0.653681, ...
Name Embeddings

- Gender (First names)

2D projection (left) of 5K popular first names’ embeddings. Orange are male names, salmon for females and gray for unlabeled. Same-gender names cluster together, indicating similar embeddings.
Name Embeddings

- Ethnicity (Last names)

2D projection (left) of 5K popular last names’ embeddings. Same-ethnicity names stand close indicating similar embeddings. Insets (left to right) highlight *White*, *Black* and *Hispanic* names.
Name Embeddings

- Nationality (Last Names)

2D projection (left) of 5K popular last names’ embeddings. Two distinct Asian clusters. Left: Chinese/ Vietnamese names. Right: South Asian names.
Methodology

- Overview
  - Hierarchical Classification
  - Naïve Bayes Model
  - Example:

![Diagram showing a hierarchical classification for Barack Obama's name]

**Algorithm 1: NamePrism**, a hierarchical nationality classifier

```
Input: first/last name \(v_f, v_l\); nationality taxonomy; estimated parameter sets \(P_{tr}, P_{em}, P_{p/s}, P_{ch}\).
Output: nationality prediction \(T\).

1. Init \(T = \) root class;
2. while \(T\) is not a leaf class do
   3. for each child class \(N_i\) of \(T\) do
      4. for each name part \(v \in \{v_f, v_l\}\) do
         5. if \(v \in V_{tr}\) then
            6. \(P(v|N_i) = P_{tr}(v|N_i)\);
         7. else if \(v \in V_{em}\) then
            8. \(P(v|N_i) = P_{em}(v|N_i)\);
         9. else
            10. \(P(v|N_i) = \sigma\); # \(\sigma\) is a small constant
      11. if neither of \(v_f, v_l\) in \(V_{tr}\) or \(V_{em}\) then
         12. for each child class \(N_i\) of \(T\) do
            13. for each name part \(v \in \{v_f, v_l\}\) do
               14. if \(v \in V_{p/s}\) then
                  15. \(P(v|N_i) = P_{p/s}(v|N_i)\);
               16. else if \(v \in V_{ch}\) then
                  17. \(P(v|N_i) = P_{ch}(v|N_i)\);
            18. \(P(N_i|v_f, v_l) \propto P(v_f|N_i) \cdot P(v_l|N_i) \cdot P(N_i)\);
      19. \(T = \arg\max_{N_i} P(N_i|v_f, v_l)\);
   20. return \(T\);
```
Methodology

- **Parameter estimation**

  - **Training Data:**
    \[
P_{tr}(v_i|N) = \frac{C(v_i, N)}{C(N)}, v_i \in V_{tr}
    \]

  - **Name Embedding:**
    \[
P_{em}(v_i|N) = \frac{P_{em}(N|v_i)P(v_i)}{P(N)}, v_i \in V_{em}
    \]
    \[
P_{em}(N|v_i) = \frac{1}{k}\sum_{v_j \in kNN(v_i)} P_{tr}(N|v_j)
    \]

  - **Prefix/Suffix:**
    \[
P_{p/s}(v_i|N) = \frac{1}{|PS(v_i)|}\sum_{v_j \in PS(v_i)} P_{tr}(v_j|N), v_i \in V_{p/s}
    \]

  - **Name Characters:**
    \[
P_{ch}(v_i|N) = \frac{1}{|CH(v_i)|}\sum_{v_j \in CH(v_i)} P_{tr}(v_j|N), v_i \in V_{ch}
    \]

  E.g. “Barack” in *African* 5 times, and **5000** African names in total: \(P_{tr}(“Barack” | African) = 0.001\)

  Assuming \(k = 1\), “Barack” nearest neighbor in embedding space “Yemi”, \(P_{em}(African | “Barack”) = P_{tr}(African | “Yemi”),\) then apply Bayes Rule.

  Assuming there are two “co-prefix” first names of “Barack”, Barclay and Baratheon. \(P_{p/s}(“Barack” | African) = \frac{1}{2} (P_{tr}(“Barclay” | African) + P_{tr}(“Baratheon” | African))\)

  For a name part in Chinese (e.g. “中山”), use the average distribution of all name parts in Chinese.
Dataset

• Name Embedding:
  – Contact lists: 57M, account holders’ names removed

• Name Labels:
  – 68M from email users (pairs of <name, registration country>), covering 59 countries;
  – 6M Twitter users, cleaned from 43M Twitter profiles, covering 118 countries;
Taxonomy

Treemap of nationality taxonomy. Nested blocks within a larger block are its child nodes. **118 countries/regions**, covering over **90% world population**, are assigned to **39 leaf nationalities**. The taxonomy is constructed based on Cultural, Ethnic and Linguist similarities.
Sanity Check

- Similarity between countries
  - Country Representation:
    \<..., 0.02, ..., 0.003, ...\>
    
    
    ... Barack@0, ..., Obama@1, ...
  - Cosine similarity
### Performance Comparison

| Nationality            | Name# | HMM | Ethnea | Embd | Prism | Prism* | Wikipedia Data | Name# | HMM | Ethnea | Embd | Prism | Prism* | Email/Twitter Data |
|------------------------|-------|-----|--------|------|-------|--------|----------------|-------|-----|--------|------|-------|--------|-------------------|
| GreaterAfrican         | 11K   | 0.428 | 0.532 | 0.480 | 0.543 | 0.486 | 31K | 0.269 | 0.389 | 0.554 | **0.645** | 0.622 |
| GreaterEuropean        | 113K  | 0.863 | 0.903 | 0.927 | **0.932** | 0.899 | 225K | 0.725 | 0.815 | 0.861 | **0.920** | 0.902 |
| Asian                  | 24K   | 0.654 | 0.670 | 0.711 | 0.745 | **0.748** | 123K | 0.674 | 0.709 | 0.763 | **0.910** | 0.904 |
| Muslim*                | 7K    | 0.380 | 0.563 | 0.538 | **0.615** | 0.611 | 13K | 0.204 | 0.374 | 0.602 | **0.612** | 0.533 |
| Africans*              | 4K    | 0.285 | 0.268 | 0.282 | **0.314** | 0.259 | 18K | 0.174 | 0.288 | 0.458 | 0.636 | **0.659** |
| WestEuropean           | 49K   | 0.631 | 0.724 | 0.709 | 0.747 | **0.756** | 143K | 0.553 | 0.735 | 0.780 | 0.873 | **0.878** |
| EastEuropean*          | 9K    | 0.488 | 0.517 | 0.466 | 0.575 | **0.629** | 38K | 0.301 | 0.582 | 0.726 | 0.794 | **0.812** |
| British*               | 44K   | 0.611 | 0.760 | 0.789 | **0.794** | 0.768 | 35K | 0.361 | 0.578 | 0.627 | 0.648 | **0.689** |
| Jewish*                | 11K   | **0.313** | 0.111 | 0.095 | 0.129 | 0.183 | 9K | 0.097 | 0.361 | 0.301 | **0.405** | 0.387 |
| GreaterEastAsian       | 15K   | 0.637 | 0.626 | 0.642 | 0.690 | **0.706** | 97K | 0.625 | 0.656 | 0.713 | **0.907** | 0.895 |
| IndianSubContinent*    | 9K    | 0.523 | 0.660 | 0.768 | **0.769** | 0.746 | 26K | 0.438 | 0.721 | 0.855 | **0.912** | 0.903 |
| Italian*               | 14K   | 0.521 | 0.543 | 0.595 | **0.634** | 0.613 | 11K | 0.233 | 0.453 | 0.665 | 0.713 | **0.763** |
| Hispanic*              | 11K   | 0.403 | **0.600** | 0.397 | 0.521 | 0.538 | 69K | 0.432 | 0.724 | 0.676 | 0.850 | **0.864** |
| Nordic*                | 5K    | 0.400 | 0.587 | **0.713** | 0.709 | 0.709 | 23K | 0.303 | 0.653 | 0.767 | **0.783** | **0.783** |
| French*                | 14K   | 0.428 | 0.523 | 0.602 | 0.600 | **0.624** | 27K | 0.203 | 0.426 | 0.738 | **0.769** | 0.750 |
| Germanic*              | 5K    | 0.254 | 0.410 | 0.401 | 0.403 | **0.412** | 13K | 0.140 | 0.431 | 0.582 | 0.629 | **0.653** |
| Japanese*              | 8K    | 0.646 | **0.724** | 0.456 | 0.547 | 0.695 | 57K | 0.674 | 0.788 | 0.434 | 0.928 | **0.939** |
| EastAsian*             | 7K    | 0.499 | 0.455 | 0.609 | **0.621** | 0.549 | 40K | 0.270 | 0.340 | 0.723 | **0.834** | 0.811 |
| Weighted Avg.          | —     | 0.492 | 0.607 | 0.619 | 0.648 | **0.651** | — | 0.364 | 0.580 | 0.642 | 0.790 | **0.795** |

F1 scores on a 13-leaf taxonomy. Existing methods: **HMM** and **Ethnea**; **Embd** only uses parameters from name embeddings; **Prism** is **NamePrism** with world population as priors.
### Performance Comparison

| Nationality     | Wikipedia | Email/Twitter |
|-----------------|-----------|---------------|
|                 | Name#     | Seer | Prism | Name#     | Seer | Prism |
| Muslim          | 7K        | 0.560 | **0.646** | 13K        | 0.422 | **0.688** |
| EastEuropean    | 9K        | **0.739** | 0.596 | 38K        | 0.343 | **0.804** |
| British         | 44K       | **0.852** | 0.843 | 35K        | 0.577 | **0.726** |
| Indian          | 9K        | 0.768 | **0.779** | 26K        | 0.639 | **0.880** |
| Hispanic        | 11K       | **0.605** | 0.558 | 69K        | 0.610 | **0.871** |
| Germanic        | 5K        | 0.464 | **0.487** | 13K        | 0.433 | **0.694** |
| French          | 14K       | **0.676** | 0.650 | 27K        | 0.482 | **0.802** |
| Italian         | 14K       | **0.707** | 0.641 | 11K        | 0.329 | **0.728** |
| EastAsian       | 7K        | **0.824** | 0.635 | 40K        | 0.418 | **0.848** |
| Japanese        | 8K        | **0.875** | 0.550 | 57K        | 0.902 | **0.929** |
| **Weighted Avg.** | — | **0.751** | 0.700 | — | 0.571 | **0.831** |

F1 scores on 10 nationalities. *EthnicSeer* performs slightly better on Wikipedia data but it was an unfair comparison because it was trained on it. *NamePrism* performs significantly better on a larger test set from Email/Twitter.
Performance Comparison

- Performance on taxonomy with 39 leaf nationalities

| Nationality         | Name#  | Prism | Nationality   | Name#  | Prism |
|---------------------|--------|-------|---------------|--------|-------|
| CelticEnglish*      | 3505K  | 0.725 | SouthAsian*   | 2623K  | 0.890 |
| Jewish*             | 11K    | 0.396 | African       | 606K   | 0.589 |
| Muslim              | 1475K  | 0.741 | EastAsian     | 6157K  | 0.920 |
| Greek*              | 259K   | 0.887 | Hispanic      | 6892K  | 0.907 |
| Nordic              | 195K   | 0.731 | European      | 5371K  | 0.836 |
| Nubian*             | 577K   | 0.650 | Japan*        | 65K    | 0.836 |
| Maghreb*            | 47K    | 0.148 | Malay         | 2596K  | 0.863 |
| ArabPeninsula*      | 172K   | 0.510 | Chinese*      | 2901K  | 0.928 |
| Turkic              | 78K    | 0.676 | Portuguese*   | 2683K  | 0.886 |
| Pakistanis          | 179K   | 0.511 | Philippines*  | 1137K  | 0.724 |
| Persian*            | 423K   | 0.656 | Spanish*      | 3072K  | 0.851 |
| Finland*            | 30K    | 0.739 | German*       | 1278K  | 0.739 |
| Scandinavian         | 165K   | 0.704 | Baltics*      | 12K    | 0.408 |
| WestAfrican*        | 315K   | 0.563 | French*       | 2674K  | 0.825 |
| SouthAfrican*       | 66K    | 0.370 | Russian*      | 121K   | 0.716 |
| EastAfrican*        | 225K   | 0.574 | EastEurope*   | 65K    | 0.492 |
| SouthKorea*         | 68K    | 0.861 | SouthSlavs*   | 68K    | 0.570 |
| Indochina           | 528K   | 0.901 | Italian       | 1153K  | 0.745 |
| CentralAsian*       | 3K     | 0.196 | Cambodia*     | 1K     | 0.162 |
| Turkey*             | 75K    | 0.687 | Vietnam*      | 502K   | 0.913 |
| Bangladesh*         | 78K    | 0.578 | Thailand*     | 18K    | 0.592 |
| Pakistan*           | 101K   | 0.449 | Malaysia*     | 242K   | 0.480 |
| Denmark*            | 49K    | 0.662 | Indonesia*    | 2354K  | 0.870 |
| Sweden*             | 74K    | 0.607 | Romania*      | 329K   | 0.663 |
| Norway*             | 42K    | 0.620 | Italy*        | 825K   | 0.710 |
| Myanmar*            | 7K     | 0.607 |              |        |       |

Weighted Avg. — 0.806
Application

• Anomaly Detection
  – Indonesian politician named Jeffrie Geovanie
  – +50% of the followers of British, Russian, or Indian nationality
  – Indonesian is the primary language
Takeaway

• Introduce *name embedding*, which captures gender, ethnicity and nationality signals.

• Propose *NamePrism*, which achieves state-of-the-art performance.

• Open API to support related research. (http://www.name-prism.com)
Questions