PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features

Ergin Elmacioglu¹, Yee Fan Tan², Su Yan¹, Min-Yen Kan¹ and Dongwon Lee¹

¹The Pennsylvania State University, USA
²National University of Singapore, Singapore

{ergin, syan, dongwon}@psu.edu
{tanyeefa, kanmy}@comp.nus.edu.sg

Our System

• Goal
  – To compare the usefulness of various features for the Web People Search Task

• Architecture
  ![Architecture Diagram]

Features

• Tokens (T)
  – Stemmed words from web pages

• Named entities (NE)
  – We consider people, organizations, locations
  – Each NE token a feature

| Born Edward Charles Morrice Fox in Chelsea, London… | Dr. Edward A. Fox holds a Ph. D. and M.S. in Computer Science from Cornell University,… |
| Charles, Chelsea, Morrice, Edward, Fox, London,… | Dr. Edward, A, Fox, Cornell, University,… |

Web People Search Task

- Number of different entities unknown
- Number of clusters unknown
- Different entities with same name may appear on a page
  – Clusters can overlap
- Web pages are free form, no standard structure

Features

• Overview
  – Tokens
  – Named entities
  – Links
  – Page URL

• All features weighted by TF-IDF except links

| Born Edward Charles Morrice Fox in Chelsea, London… | Charles, Morrice,… |
| Dr. Edward A. Fox holds a Ph. D. and M.S. in Computer Science from Cornell University,… | Dr. A,… |

Features

• NE-targeted (NE-T)
  – Motivation: middle names and titles
  – For NEs having a token of target name
  – Extract tokens that are not in target name as features
Features

- Hostname (H), Domain (D)
  - http://www.cs.ualberta.ca/~lindek/
  - http://www.cs.ualberta.ca/~pinchak/
- Two pages link to common rare hostnames/domains?
- Each hostname/domain a feature, weighted by IDF
- Works well for mixed citation problem (Tan et al., 2006)
- Hostname with Self (H-S), Domain with Self (D-S)
- URL of web page is also counted as one of its “links”

Evaluation

- Training data
  - 7 Wikipedia names
  - 10 ECDL names
  - 32 US Census names (Mann, 2003, 2006)
- Test data
  - 10 ACL names
  - 10 Wikipedia names
  - 10 US Census names
- Evaluation measure
  - Purity and inverse purity (Hotho et al., 2003)
  - F-measure (α = 0.5 and α = 0.2)

Analysis

- NE performs better than Tokens for ECDL and Wikipedia
  - Useful to identify related locations and organizations
  - Irrelevant tokens in menus, headers, etc.
- NE targeted (NE-T)
  - Discards too much information, so low recall

- Hostname (H), Domain (D)
  - Domain better than Hostname due to better recall
  - +Self gives slight increase in recall
- Page URLs (U)
  - Highly precise but issue with recall

Evaluation (F(α = 0.5) and similarity threshold 0.2)
Submission Run

- NEs as features, HAC similarity threshold 0.2

Conclusion

- System
  - Feature generation + Clustering

- Comparison between various features
  - Named entities in web pages make good features

- Submission run
  - Achieved 3rd place among 16 teams

Thank you