

NUS PENNSTATE

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

Ergin Elmacioglu<sup>1</sup>, Yee Fan Tan<sup>2</sup>, Su Yan<sup>1</sup>,  
Min-Yen Kan<sup>2</sup> and Dongwon Lee<sup>1</sup>

<sup>1</sup>The Pennsylvania State University, USA  
<sup>2</sup>National University of Singapore, Singapore

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

NUS PENNSTATE

## Web People Search Task

target person name  
↓  
YAHOO!

↓

Title Snippet  
URL  
Web page

↓

↓

- 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

E. Elmacioglu, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features

NUS PENNSTATE

## Our System

- Goal
  - To compare the usefulness of various features for the Web People Search Task
- Architecture

Input web pages → Feature vectors → Clusters

Cosine similarity + Single link hierarchical agglomerative clustering + Minimum similarity threshold

E. Elmacioglu, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features

NUS PENNSTATE

## Features

- Overview
  - Tokens
  - Named entities
  - Links
  - Page URL
- All features weighted by TF-IDF except links

E. Elmacioglu, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features

NUS PENNSTATE

## 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... → Charles, Chelsea, Morrice, Edward, Fox, London, ...

Dr. Edward A. Fox holds a Ph. D. and M.S. in Computer Science from Cornell University, ... → Dr., Edward, A., Fox, Cornell, University, ...

E. Elmacioglu, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features

NUS PENNSTATE

## 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





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., ...

E. Elmacioglu, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features

NUS PENNSTATE SemEval 2007 Workshop, Association of Computational Linguistics

## Features

- Hostname (H), Domain (D)**
  -  → http://www.cs.ualberta.ca/~lindek/
   
 → http://www.cs.ualberta.ca/~pinchak/
   
 } common hostname
  -  → http://www.cs.ualberta.ca/~lindek/
   
 → http://armena.cs.ualberta.ca/lindek/downloads/sim.tgz
   
 } common domain
- 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"

E. Elmasri, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 7

NUS PENNSTATE SemEval 2007 Workshop, Association of Computational Linguistics

## Features

- Page URLs (U)**

http://www.cs.ualberta.ca/~lindek/

  - URL itself tells quite a lot
    - Home page of "lindek"
    - CS department, University of Alberta, Canada
  - MeURLin (Kan and Nguyen Thi, 2005)
    - Tokens (http, www, cs, ualberta, ca, lindek)
    - URI parts (scheme:http, hostname:cs, user:lindek, ...)
    - N-grams (ca ualberta, ualberta cs, cs www, www lindek)
    - Length of tokens
    - ...

E. Elmasri, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 8

NUS PENNSTATE SemEval 2007 Workshop, Association of Computational Linguistics

## 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

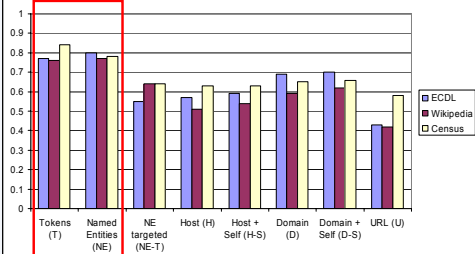
} merged into single data set
- Evaluation measure**
  - Purity and inverse purity (Hotho et al., 2003)
  - F-measure ( $\alpha = 0.5$  and  $\alpha = 0.2$ )

E. Elmasri, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 9

NUS PENNSTATE SemEval 2007 Workshop, Association of Computational Linguistics

## Evaluation

- F ( $\alpha = 0.5$ ) and similarity threshold 0.2**



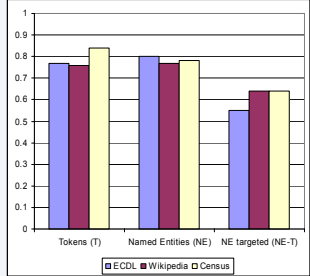
Feature	ECDL	Wikipedia	Census
Tokens (T)	~0.75	~0.85	~0.80
Named Entities (NE)	~0.75	~0.80	~0.75
NE targeted (NE-T)	~0.65	~0.65	~0.65
Host (H)	~0.55	~0.55	~0.55
Host + Self (H-S)	~0.60	~0.60	~0.60
Domain (D)	~0.65	~0.65	~0.65
Domain + Self (D-S)	~0.60	~0.60	~0.60
URL (U)	~0.45	~0.45	~0.45

E. Elmasri, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 10

NUS PENNSTATE SemEval 2007 Workshop, Association of Computational Linguistics

## 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



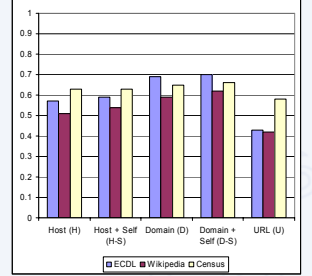
Feature	ECDL	Wikipedia	Census
Tokens (T)	~0.75	~0.80	~0.85
Named Entities (NE)	~0.75	~0.80	~0.80
NE targeted (NE-T)	~0.55	~0.65	~0.65

E. Elmasri, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 11

NUS PENNSTATE SemEval 2007 Workshop, Association of Computational Linguistics

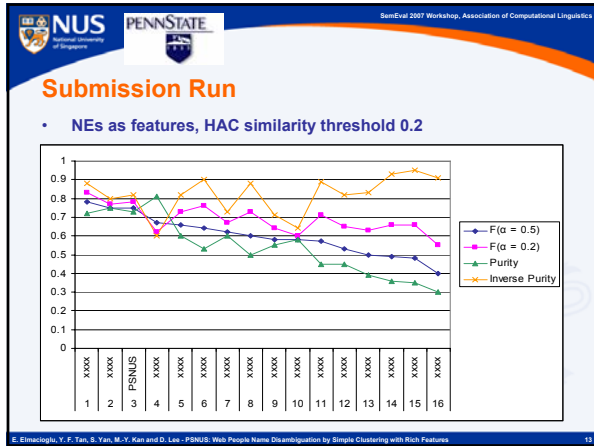
## Analysis

- 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



Feature	ECDL	Wikipedia	Census
Host (H)	~0.55	~0.55	~0.55
Host + Self (H-S)	~0.60	~0.60	~0.60
Domain (D)	~0.65	~0.65	~0.65
Domain + Self (D-S)	~0.60	~0.60	~0.60
URL (U)	~0.45	~0.45	~0.45

E. Elmasri, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 12



NUS PennState SemEval 2007 Workshop, Association of Computational Linguistics

### 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

E. Elmeleggi, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 14

NUS PennState SemEval 2007 Workshop, Association of Computational Linguistics

### Thank you

E. Elmeleggi, Y. F. Tan, S. Yan, M.-Y. Kan and D. Lee - PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features 15