# The Person Index Challenge: Extraction of Persons from Messy, Short Texts

Markus Schröder, Christian Jilek, Michael Schulze and Andreas Dengel

German Research Center for Artificial Intelligence

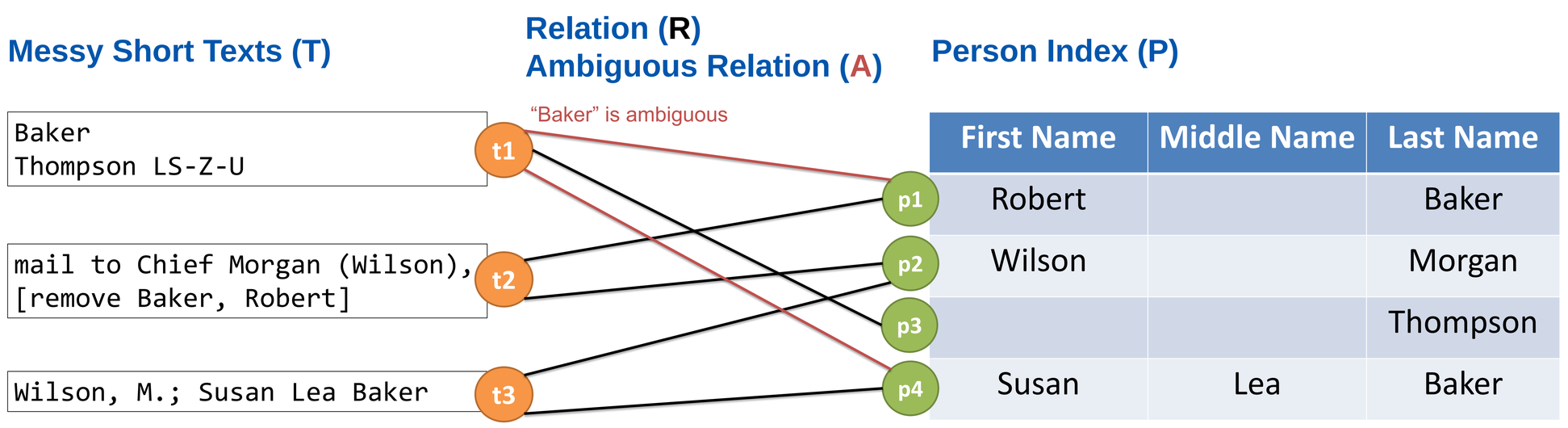
Technische Universität Kaiserslautern

## Abstract

When persons are mentioned in texts with their first name, last name and/or middle names, there can be a high variation which of their names are used, how their names are ordered and if their names are abbreviated. If multiple persons are mentioned consecutively in very different ways, especially short texts can be perceived as "messy". Once ambiguous names occur, associations to persons may not be inferred correctly. Despite these eventualities, in this paper we ask how well an unsupervised algorithm can build a person index from short texts. We define a person index as a structured table that distinctly catalogs individuals by their names. First, we give a formal definition of the problem and describe a procedure to generate ground truth data for future evaluations. To give a first solution to this challenge, a baseline approach is implemented. By using our proposed evaluation strategy, we test the performance of the baseline and suggest further improvements. For future research the source code is publicly available.

## Introduction

- given: messy, short texts with names; may be ambiguous
  - high variation: which names are used, how names are ordered and if they are abbreviated
- expected: person index - a structured table that distinctly catalogs persons
- contributions:
  - formal problem definition
  - procedure that generates ground truth data
  - evaluation strategy to assess the quality of solutions




## Ground Truth Generator

- heavily inspired by concrete data observed in an industrial scenario
  - spreadsheets completed with copy&paste: transfer of names lead to various name variations
- uses patterns to generate mentions of persons in various ways
- several generator parameter settings possible:
  - number of **p**ersons
  - number of generated short **t**exts
  - **max**imum number of mentioned persons per text
  - number of persons with **m**iddle **n**ames
  - degree of **amb**iguity (*n* groups of *n* people share name)

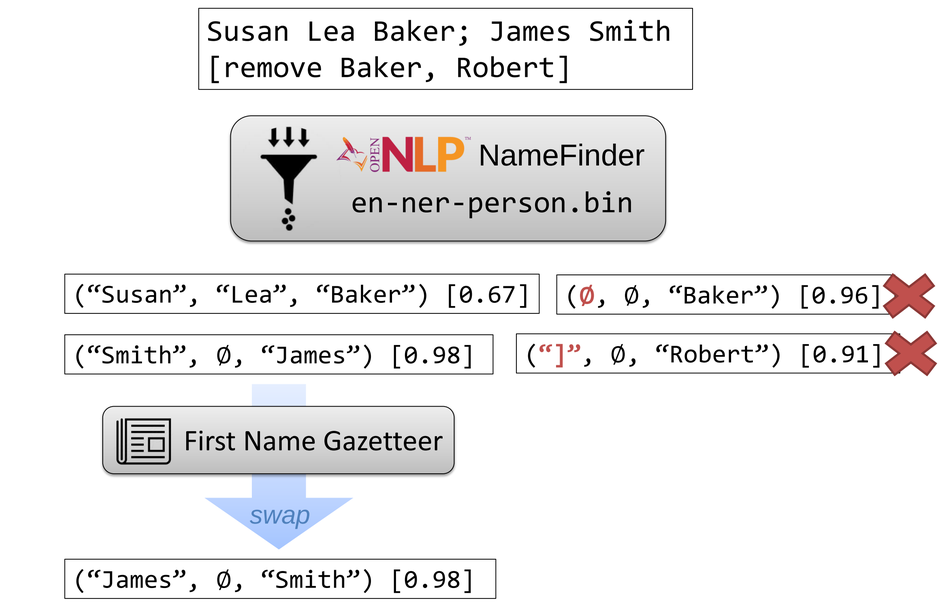
**Name Patterns**

| Nr. | Pattern | Example |
|---|---|---|
| 1 | *fn* | John |
| 2 | *ln* | Kennedy |
| 3 | *fn ln* | John Kennedy |
| 4 | *ln fn* | Kennedy John |
| 5 | *ln, fn* | Kennedy, John |
| 6 | *ln, letter(fn).* | Kennedy, J. |
| 7 | *ln department()* | Kennedy US-Z-G |
| 8 | *department()*↵*ln fn* | US-Z-G↵Kennedy John |
| 9 | *ln fn <lc(ln)@rnd(5).rnd(2)>* | Kennedy John <kennedy@xraok.nc> |
| 10 | *note() role() ln fn* | new Admin Kennedy John |
| 11 | *fn mn ln* | John Fitzgerald Kennedy |
| 12 | *fn letter(mn). ln* | John F. Kennedy |
| 13 | *letter(fn). letter(mn). ln* | J. F. Kennedy |
| 14 | *ln, letter(fn). letter(mn).* | Kennedy, J. F. |

| Nr. | $\|P\|$ | $\|T\|$ | Max | MN | Amb | $prec_P$ | $recall_P$ | $f_P$ | $prec_R$ | $recall_R$ | $f_R$ | $prec_A$ | $recall_A$ | $f_A$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 10 | 0 | 0 | 0 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | - | - | - |
| 2 | 1 | 200 | 0 | 0 | 0 | 0.14 | 1.00 | 0.25 | 0.00 | 0.00 | - | - | - | - |
| 3 | 20 | 200 | 0 | 0 | 0 | 0.63 | 0.85 | 0.72 | 0.82 | 0.72 | 0.77 | - | - | - |
| 4 | 20 | 200 | 10 | 0 | 0 | 0.38 | 0.90 | 0.54 | 0.61 | 0.09 | 0.16 | - | - | - |
| 5 | 20 | 200 | 10 | 4 | 0 | 0.31 | 0.80 | 0.45 | 0.63 | 0.09 | 0.16 | - | - | - |
| 6 | 20 | 200 | 10 | 4 | 2 | 0.39 | 0.75 | 0.52 | 0.47 | 0.41 | 0.44 | 0.03 | 1.00 | 0.06 |
| 7 | 20 | 200 | 10 | 4 | 3 | 0.45 | 0.85 | 0.59 | 0.59 | 0.54 | 0.56 | 0.05 | 0.95 | 0.09 |
| 8 | 40 | 300 | 10 | 4 | 3 | 0.39 | 0.75 | 0.52 | 0.53 | 0.49 | 0.51 | 0.03 | 0.87 | 0.06 |

Performance: building the person index | mapping between short text and person | ambiguity detection

## Baseline Approach




| First Name | Middle Name | Last Name |
|---|---|---|
| Susan | Lea | Baker |
| James | | Smith |

### Suggestion for Improvements

- train detection models that are able to distinguish first name and last name
- consider more context when linking and disambiguating

## Evaluation

- best average performance with first name gazetteer used
- but does not reach f-measure of 0.6
- reasons for performance decline
  - role names or department names are identified as person names
  - falsely extracted persons leads to misconception of ambiguity
  - less text: fewer opportunities to find a correct name pair
  - multiple persons mentioned in one text: names get mixed up
  - persons share first names or last names by accident

**Resources** https://github.com/mschroeder-github/person-index

### Acknowledgement

This work was funded by the BMBF project SensAI (grant no. 01IW20007).

**Contact:**
M.Sc. Markus Schröder
Doctoral Researcher
German Research Center for Artificial Intelligence (DFKI GmbH)
Smart Data & Knowledge Services

Phone: +49 631 20575-2070
Mail: markus.schroeder@dfki.de
Website: http://www.dfki.uni-kl.de/~mschroeder/