# *Personae*, a corpus for author and personality prediction from text

Kim Luyckx & Walter Daelemans

CNTS Language Technology Group
University of Antwerp, Belgium

LREC 2008
May 28th, 2008

# Computational Stylometry

Writing style reflects

- Topic, register, genre (text)
- Identity, gender, region, age, personality (author)

Two-stage approach

- Selection of predictive features $\Longrightarrow$ Language Technology
- Machine Learning for categorization

*Personae corpus*: innovative for

- Author prediction
- Personality prediction

# Author Prediction

| | Traditional approach | Our approach |
|---|---|---|
| AUTHORS | small, closed set of authors | large number of authors |
| DATA | lots of data per author | limited data |
| RESULTS | upper-90% | exploratory experiments |
| APPS. | disputed authorship<br>unrealistic forensic linguistics<br>PROBLEM ALMOST SOLVED | plagiarism detection<br>realistic forensic linguistics<br>LOT OF WORK TO DO! |

forensic linguist C.Chaski "99.9% sure that *x* did not write the text (and *y* did)"

$\Longleftrightarrow$

short e-mail, limited training data, lots of candidate authors

CNTS Language Technology Group

# Personality Prediction

Relatively new in computational stylometry

- LangPsy: direct correlation between personality & language (Gill, 2003; Campbell & Pennebaker, 2003)
- CompLing: little research (Argamon *et al.*, 2005; Nowson & Oberlander, 2007; Mairesse *et al.*, 2007)
- Five-Factor Model of Personality: OCEAN
- Stream-of-consciousness essays by (psychology) students

We take it further

1. Texts on non-personality related topic
2. Corpus of Dutch written language
3. Lots of authors + limited data

CNTS Language Technology Group

# *Personae* Corpus

- 200K words of Dutch
- 145 student essays about a documentary on Artificial Life (factual description & opinion)
  $\Longrightarrow$ genre, register, topic & age are kept relatively constant
- avg. 1,400 words/student (1 text/student)
- Released copyright to University of Antwerp
- Online MBTI test

CNTS Language Technology Group

# Myers-Briggs Type Indicator (MBTI)

- Forced-choice test
- Carl Jung's personality typology
- Categorization according to 4 preferences:
  - Introversion & Extraversion (*attitudes*) $\sim$ Extraversion
  - iNtuition & Sensing (*information-gathering*) $\sim$ Openness
  - Feeling & Thinking (*decision-making*) $\sim$ Agreeableness
  - Judging & Perceiving (*lifestyle*) $\sim$ Conscientiousness
- Parallels with Five-Factor Model of Personality
- Controversial domain
- Consensus over at least first 2 dimensions
- Apps.: career & marriage counseling, group dynamics, HR,...

CNTS Language Technology Group

# Structure

- Too homogeneous for experiments on gender, mother tongue, or region
- Interesting distributions in at least 2 MBTI preferences

| I | E |
|---|---|
| .45 | .55 |

| N | S |
|---|---|
| .54 | .46 |

| F | T |
|---|---|
| .72 | .28 |

| J | P |
|---|---|
| .81 | .19 |

CNTS Language Technology Group

# Exploratory experiments with *Personae* corpus

- author -

- personality -

# 1. Selection of predictive features

- Memory-Based Shallow Parsing (MBSP): tagging, chunking & id. of syntactic relations (Daelemans & van den Bosch, 2005)
- AA instances
  - Texts are split in 10 fragment, 9 in training and 1 in test
  - Feature vector per fragment + author label
- PP instances
  - 90% of authors in training, 10% in test
  - Feature vector per author + personality type
- Type-token ratio, readability components, function word distributions, *n*-grams of fine- & coarse-grained POS, *n*-grams of words
- $\chi^2$ metric
- Single feature sets + combinations

CNTS Language Technology Group

# 2. Machine Learning

- TiMBL (Daelemans *et al.*, 2003)
  - Implementation of kNN
  - Extensions for nominal features & relevance weighting
  - Does not abstract away from exceptions
  - Compares test instance to all training instances in memory
  - Better fit for limited data than eager learners?
- 10-fold cross-validation & micro-average of confusion matrix
- AA: 145 author classes
- PP
  - 8 binary classification tasks (e.g. I or not-I)
  - 4 tasks: distinguish between two poles (e.g. I or E)

| Features | Accuracy |
|---|---|
| tok | 29.17% |
| fwd | **32.83%** |
| lex 1 | **34.00%** |
| lex 2 | 22.90% |
| lex 3 | 12.00% |
| cgp 1 | 30.00% |
| cgp 2 | 31.17% |
| cgp 3 | 28.21% |
| pos 1 | **34.48%** |
| pos 2 | 30.55% |
| pos 3 | 17.10% |
| lex1 + pos1 | 40.69% |
| lex1 + pos1 + tok | 48.28% |
| lex1 + pos1 + tok + fwd1 | 48.28% |
| lex1 + tok | **49.21%** |

Table: TiMBL results in authorship attribution on 145 authors

CNTS Language Technology Group

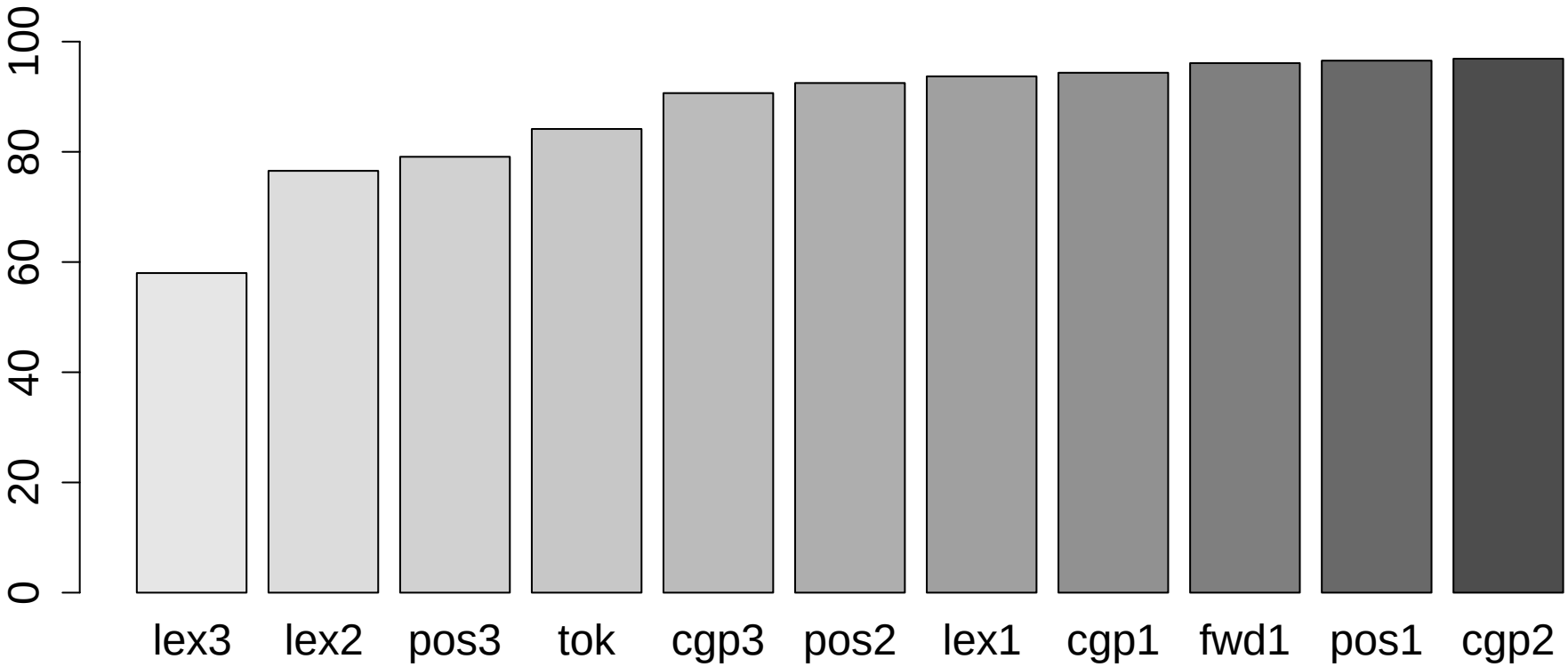


Table: TiMBL results in 100 AA experiments with random samples of 2 authors

# Related Work in Authorship Attribution

- Van Halteren, 2005
  - 8 authors, 9 texts per author
  - 97% correctly classified
- Argamon *et al.*, 2003
  - 20 authors
  - 25% - 45% accuracy (depending on topic)
- Madigan *et al.*, 2005
  - 114 authors, 200 texts/author
  - Error rates between 97% and 20%
- Koppel *et al.*, 2006
  - Thousands of candidate authors, lots of blogs/author
  - Answer in 31.3% of cases, correct almost 90% of the time

CNTS Language Technology Group

| Task | Feature set | Precision | Recall | F-score | Accuracy |
|---|---|---|---|---|---|
| **Introverted** | lex 3-grams | 56.70% | 84.62% | 67.90% | 64.14% |
| | *random* | *44.1%* | *46.2%* | | |
| **Extraverted** | cgPOS 3-grams | 58.09% | 98.75% | 73.15% | 60.00% |
| | *random* | *54.6%* | *52.5%* | | |
| **iNtuitive** | cgPOS 3-grams | 56.92% | 94.87% | 71.15% | 58.62% |
| | *random* | *48.7%* | *48.7%* | | |
| **Sensing** | POS 3-grams | 50.81% | 94.03% | 65.97% | 55.17% |
| | *random* | *40.3%* | *40.3%* | | |
| **Feeling** | lex 3-grams | 73.76% | 99.05% | 84.55% | 73.79% |
| | *random* | *72.6%* | *73.3%* | | |
| **Thinking** | lex 1-grams | 40.00% | 50.00% | 44.44% | 65.52% |
| | *random* | *28.2%* | *27.5%* | | |
| **Judging** | lex 3-grams | 81.82% | 100.00% | 90.00% | 82.07% |
| | *random* | *77.6%* | *76.9%* | | |
| **Perceiving** | lex 2-grams | 26.76% | 67.86% | 38.38% | 57.93% |
| | *random* | *6.9%* | *7.1%* | | |

Table: TiMBL results for eight binary classification tasks

CNTS Language Technology Group

| Task | Feature set | F-score [INFJ] | F-score [ESTP] | Average F-score | Accuracy |
|---|---|---|---|---|---|
| **I vs. E** | lex 3-grams | 67.53% | 63.24% | 65.38% | 65.52% |
| | *random* | | | | *49.7%* |
| | *majority* | | | | *55.2%* |
| **N vs. S** | pos 3-grams | 58.65% | 64.97% | 61.81% | 62.07% |
| | *random* | | | | *44.8%* |
| | *majority* | | | | *53.8%* |
| **F vs. T** | lex 3-grams | 84.55% | 13.64% | 49.09% | 73.79% |
| | *random* | | | | *60.7%* |
| | *majority* | | | | *72.4%* |
| **J vs. P** | lex 3-grams | 90.00% | 13.33% | 51.67% | 82.07% |
| | *random* | | | | *63.5%* |
| | *majority* | | | | *80.7%* |

Table: TiMBL results for four discrimination tasks

CNTS Language Technology Group

# Related Work in Personality Prediction

- Argamon *et al.*, 2005
  - Stream-of-consciousness + deep self analysis essays
  - Lexical stylistic features
  - **E** 57%, **N** 58% acc.
- Nowson & Oberlander, 2007
  - Feature selection & training on small, clean blog corpus
  - Testing on large, automatically selected corpus
  - **O** skewed, **C** 56.6%, **E** 50.6%, **A** 52.9%, **N** 55.8% acc.
- Mairesse *et al.*, 2007
  - Stream-of-consciousness essays by psych. students
  - Top-down approach
  - Pos/neg emotion words, self-references,...
  - **O** 62.1%, **C** 55.3%, **E** 55.0%, **A** 55.8%, **N** 57.4% acc.

CNTS Language Technology Group

# Conclusions

*Personae* corpus is innovative for AA & PP

- Large number of authors
- Limited data
- Closer to natural situation

Authorship Attribution

- 145 authors: almost 50% accuracy
- Using combinations leads to significant improvements

Personality Prediction

- First 2 personality dims are predicted fairly accurately
- Good results in 6 out of 8 binary classification tasks
- Skewed class distributions: around 51% and 46% F-score

# Further Research

- Test features suggested in PP literature
- Effect* of number of authors
- Effect* of limited data
- Lazy vs. eager ML algorithms
- Authorship Verification: *one-vs.-all*
- Genetic Algorithm optimization

* features, performance, Machine Learners

CNTS Language Technology Group

# Contact

Kim Luyckx
kim.luyckx@ua.ac.be
http://www.cnts.ua.ac.be/∼kim

Walter Daelemans
walter.daelemans@ua.ac.be
http://www.cnts.ua.ac.be/∼walter

Stylometry Project

Sponsored by FWO (National Research Fund - Flanders)

http://www.cnts.ua.ac.be/∼kim/Stylometry.html

# Questions?

"Piled Higher and Deeper" by Jorge Cham - www.phdcomics.com