Author profiling: more linguistics and explanation

Ben Verhoeven
In close collaboration with Walter Daelemans

Presented as a JOTA Lecture
Ljubljana, Slovenia
22 November 2016

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

Text mining

Three layers of information in text

- Objective
  - Facts, concepts, characteristics of concepts, relations between concepts, . . .
  - Who does what, where, how and why?
- Subjective
  - Opinion, sentiment, emotion, . . .
  - Who believes what about what?
- Metadata - Profile
  - Age, gender, region, . . .
  - What do we know about the author?

Example

- Objective
  - Who? Damjan Popič
  - Did what? Presented at TEDxKranj on the Janes project

- Subjective
  - What? The presentation was succesful
  - Who believes this? Darja Fišer

Example

- Metadata – Profile
  - Who? Darja Fišer
  - Age? Mid-thirties
  - Gender? Female
  - Personality? Extravert
  - Education? Highly educated

Stylometry

The quantitative study of stylistic characteristics of a text

Writing style

A combination of invariant and unconscious decisions in language production on all linguistic levels, uniquely associated with specific authors or groups of authors

→ Human Stylome Hypothesis (Van Halteren et al. 2005)

Computational stylometry

- Authorship identification
  - Attribution - attribute text to one of limited set of authors
  - Verification - is unknown text written by given author?
- Author profiling
  - Prediction of sociological or psychological characteristics of an author

Text categorization

- Class representation
- Document representation (features)
- Supervised machine learning method

Class representation

Author profiling

- Age (e.g. 10s, 20s, 30s, 40s, …)
- Gender (e.g. male vs. female)
- Location
- Personality
- Education
- Ideology
- Mental health

Brief catalogue of features

Numeric

- Complexity, readability
- Vocabulary richness
  - Type-token ratio
  - Hapax legomena
- Averages or distributions of
  - Syllable length
  - Word length
  - Sentence length

Character-level

- Letter frequency
- Punctuation
- Spelling errors
- Character n-grams

Brief catalogue of features

Word-level

- Word n-grams
- Special dictionaries
- Morphology: prefixes and suffixes

Syntax

- Part-of-speech distributions
- Frequencies of syntactic chunks (e.g. NP = Det + Adj + N)

…

Which documents?

Data with associated classes needed to train a classifier.

Not that many existing resources
(especially for Dutch)

Issues

- Authorial profile can be hard to get
- Not all freely available
  - Non-disclosure agreements
  - Anonymization problems
- None have more than 2 kinds of meta-data

Why do we want all meta-data?

- All aspects have an influence on the author's writing style
- More importantly: these aspects are reflected in the same kind of features
  - E.g. pronouns (Pennebaker, 2011)
- Solutions:
  - control for some aspects
  - balance the data
  - take all aspects into account

Some resources for personality

- Essays dataset (Pennebaker, later Mairesse)
  - English stream-of-consciousness texts by students
- myPersonality (Stillwell & Kosinski)
  - Large-scale data collection through Facebook app, many languages
- Personae (Luyckx & Daelemans)
  - Dutch essays, written by students
- **CSI Corpus** (Verhoeven & Daelemans)
  - Dutch papers, essays and reviews written by students
- **TwiSty Corpus** (Verhoeven, Daelemans & Plank)
  - Multilingual Twitter stylometry corpus

CLiPS Stylometry Investigation (CSI)

- Corpus in two genres: essays and reviews
- Large amount of meta-data
- Multitude of purposes
  - Mostly in computational stylometry
- Freely available
- Yearly expansion

CSI Corpus

Author meta-data

- Age
- Gender: male/female
- Sexual orientation*: straight or LGBT
- Region of origin: Belgian provinces or The Netherlands
- Personality profile: Big Five and MBTI*

* Provided optionally

Personality typologies

Big Five

- **O**penness to experience
- **C**onscientiousness
- **E**xtraversion
- **A**greeableness
- **N**euroticity

Score 0-100 per trait

MBTI (Myers-Briggs Type Indicator)

- Extravert – Introvert
- Thinking – Feeling
- Sensing – iNtuition
- Judging – Perceiving

Dichotomy with score 0-100

CSI Corpus

Document meta-data

- Genre
  - Essays, papers: written for Dutch proficiency course (university level), formal text
  - Reviews: special assignment
- Document info
  - Topic, sentiment, veracity (true/false) of reviews
  - Grades of papers and essays

CSI Corpus

Corpus size

Genres	# docs	# tokens	Avg. length	Std. dev.
Reviews	1298	202,827	156	65
Essays	517	565,885	1095	734
Total	1815	768,712		

CSI Corpus

Advantages

- Multiple purposes
- Yearly expansion
- Text from similar sources (within each genre)
- Enables cross-genre experiments

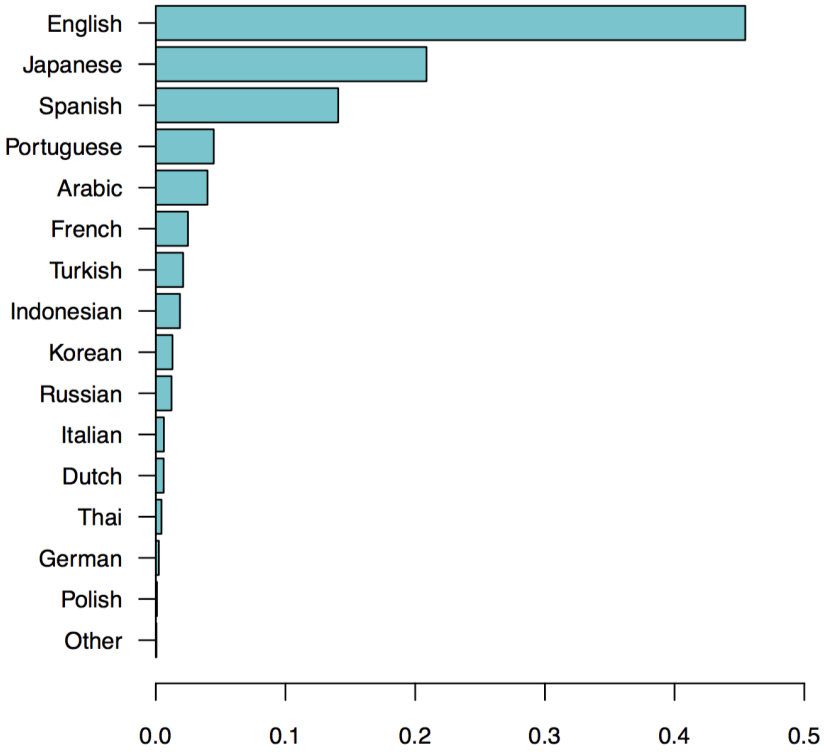
Disadvantages

- Opportunistic nature (restricted to authors at hand) influences balance of meta-data

Twitter Stylometry (TwiSty)

TwiSty Corpus

- Large-scale multilingual Twitter corpus for personality and gender
- All Western European languages in top 20 of Twitter frequencies, apart from English
  - IT, NL, DE, ES, PT, FR



TwiSty Corpus

- Developed on idea of Plank & Hovy (2015)
  - Twitter mining for only one week
  - Search for MBTI types via API
  - Only English
  - Annotating gender
  - Result
    - 1500 authors
    - 1.2M tweets

Refresher: MBTI

- Myers-Briggs Type Indicator
  - Extraversion vs. Introversion
  - iNtuitive vs. Sensing
  - Thinking vs. Feeling
  - Judging vs. Perceiving
- 16 Types
  - E.g. ESTJ, ISFP, ENTP, …

TwiSty Corpus

Data collection

- Twitter search instead of mining through API
- Search for combination of each MBTI type with language-specific words
- Download HTML

Italian	*che, sono, fatto*
Dutch	*ik, jij, het, persoonlijkheid*
German	*ich, bist, Persönlickheit, dass*
French	*suis, c'est, personnalité*
Spanish	*soy, tengo, personalidad*
Portuguese	*sou, personalidade*

TwiSty Corpus

Data clean-up

- Filter out tweets that were not relevant:
  - Not about author
    - *@schrooten ok, ik heb deze test destijds met een uitgebreide vragenlijst op mijn werk gedaan. Meerdere van mijn collega PM-ers zijn ESTJ...*
  - Ambiguity of type
    - *Volgens mij ben ik zowel INTJ als ESTJ -- het eerste als ik me rot voel, het tweede als het goed gaat. #beetjevreemd*
  - In different language
    - ***Estj** seregas muzon4**ik**? **Het**. O, nu tad davaj daj timati, etoj dj dljee.;D*
- Label for gender

TwiSty Corpus

- Corpus size in profiles

DE	IT	NL	FR	PT	ES
411	490	1,000	1,405	4,090	10,772

- Corpus size in tweets

	Total	Mean	SD	Median
German	952,549	2,318	819	2,628
Italian	932,785	1,904	912	2,146
Dutch	2,083,484	2,083	963	2,426
French	2,786,589	1,983	932	2,254
Portuguese	8,833,132	2,160	878	2,456
Spanish	18,547,622	1,722	952	1,930

TwiSty Corpus

Language Identification

- Many bilingual/polyglot Twitter users
- Tweet-level identification
- Majority voting approach with three language identifiers

Tool	Authors	# Langs
langid.py	Lui & Baldwin (2012)	97
langdetect	Nakatani (2010)	53
ldig	Nakatani (2012)	17

TwiSty Corpus

- Corpus size in tweets

	Total	Confirmed	% Confirmed
Italian	932,785	658,332	70.6
Dutch	2,083,484	1,541,259	74.0
German	952,549	713,744	74.9
Spanish	18,547,622	13,493,445	72.8
French	2,786,589	1,995,865	71.6
Portuguese	8,833,132	6,353,763	71.9

Experiment

- Instances: 200 tweets per user
- Preprocessing: normalize urls, hashtags, mentions and tokenize
- Features: character and word n-grams
- Model: LinearSVC
- Evaluation: 10-fold cross-validation

Gender prediction

Language	WRB	MAJ	F-score
DE	50.28	53.75	**77.62**
IT	54.78	65.46	**73.29**
NL	50.04	51.41	**82.61**
FR	51.84	59.60	**83.80**
PT	52.15	60.36	**87.55**
ES	51.00	57.06	**87.62**

Personality prediction

Lang	Trait	WRB	MAJ	F-score
DE	I-E	60.22	72.61	72.27
	S-N	71.03	82.43	74.49
	T-F	51.16	57.62	**59.03**
	J-P	53.68	63.57	61.99
IT	I-E	65.54	77.88	77.78
	S-N	75.60	85.78	79.21
	T-F	50.31	53.95	52.13
	J-P	50.19	53.05	*47.01*
NL	I-E	53.02	62.28	**62.90**
	S-N	57.66	69.57	**70.49**
	T-F	51.47	58.59	**59.95**
	J-P	52.00	60.00	57.99

Personality prediction

Lang	Trait	WRB	MAJ	F-score
FR	I-E	54.77	65.44	**66.49**
	S-N	68.00	80.00	78.90
	T-F	50.65	55.68	**58.22**
	J-P	52.13	60.32	56.79
PT	I-E	53.36	62.97	**66.69**
	S-N	65.60	76.08	73.42
	T-F	51.27	57.98	**61.62**
	J-P	50.87	56.61	56.53
ES	I-E	50.00	50.49	**61.09**
	S-N	55.42	66.47	61.54
	T-F	51.63	59.04	**59.73**
	J-P	51.53	58.75	56.08

Conclusion

- Large-scale, “opportunistic”, multilingual social media corpus

- Gender prediction works very well
- Personality prediction is more difficult, yet possible

Slovene Twitter

- Currently, working on gender prediction on Slovene tweets
  - Twitter corpus from Janes project
    - 6,500 authors
    - 2/3 male, 1/3 female

Explanation

- Lack of effort in stylometry to explain results, despite some great early examples

- Argamon & Koppel (2003)
  - Use of pronouns (more by women) and certain types of noun modification (more by men)
    - ‘Male’ words: a, the, that, these, one, two, more, some
    - ‘Female’ words: I, you, she, her, their, myself, yourself, herself
  - More ‘relational’ language by women, more ‘informative/rational’ language by men
    - Even in formal language (non-fiction)

More linguistics

- Discourse

- Semantics

Discourse

- What
  - relations between sentences
  - coherent structure
  - situating text in the world
- How
  - discourse relational devices (DRD)

Discourse

Features

- Dictionary with categories for different kinds of discourse structure
- Frequencies of categories are an approximation of their use

Discourse

- Penn Discourse Treebank tagset

  (PDTB Research Group, 2007)

  - TEMPORAL
    - Synchronous: terwijl
    - Asynchronous: alvorens, nadat
  - CONTINGENCY
    - Cause: dankzij, want
    - Condition: aangezien, als

Discourse

- Penn Discourse Treebank tagset
  (PDTB Research Group, 2007)
  - COMPARISON
    - Contrast: oftewel
    - Concession: ofschoon, wanneer
  - EXPANSION
    - Conjunction: alsook, eveneens
    - Instantiation: zoals
    - Restatement: alsof
    - Alternative: noch, hetzij
    - Exception: uitgezonderd
    - List: en

Ambiguity

- Nothing much changed **while/TIME** I was away.
- **While/CONCESSION** I wouldn't recommend a night-time visit, by day the area is lovely.
- One person wants out, **while/CONTRAST** the other wants the relationship to continue.

$\Rightarrow$ Weighting

Dictionary Creation

- Penn Discourse Treebank (PDTB): text with annotated discourse connectives
  - Make dictionary of connectives with weighted classes
- Extrapolated this dictionary to other languages
  - Using multilingual lexica of discourse markers created from aligned Europarl corpora

Dictionary Creation

- 86 English seed words from PDTB
- Number of translations found
  - Dutch: 335
  - German: 341
  - Slovene: 299
- On average 2 categories per connective
- Mean strength of strongest category: 90%

Ongoing research

- Evaluate this dictionary on German annotated lexicon: DimLex
- Experiments using discourse dictionaries for Dutch & English gender classification on news corpora

Hvala za pozornost!

Ali imate vprašanja?

ben.verhoeven@uantwerpen.be
@verhoevenben