Text Analysis and Machine Learning for Stylometrics and Stylogenetics

Walter Daelemans
(in cooperation with Kim Luyckx and Edward Vanhoutte)
CNTS, department of linguistics
walter.daelemans@ua.ac.be

Some recent questions from society to language technology ...
... that we refused to answer
The MA Thesis of Freya Van Den Bossche

- Vice prime minister & finance minister Belgian government
- Media upheaval about rumor that she didn’t write her MA thesis herself
- Journalist: “Can you prove scientifically that she didn’t write it?”

The e-mail of Pascal Vynckje

- Started a website and wrote a book for +50 helping them getting started on the internet
- Suddenly appeared in the shortlist for the “Greatest Belgian” competition after a mobilization by e-mail of the many users of “seniorennet”
- The e-mail (under pseudonym) was traced to his home, but he claimed his grandmother had sent it.
- Journalist: “Can you see from the text that the e-mail was written by himself?”
The most devious one ...

- Colleague: “I got this horrible anonymous review of my paper, can’t you identify which PC member wrote it?”

- Which information can be found out from a text about its author?
- Which methodology would be suitable for extracting this information?

“Stylometrics”

- Combinations of linguistic invariants in text, not under the conscious control of the author, can be used to determine
  - Individual authors (author detection)
  - Characteristics of authors
    - Gender detection
    - Region, age, education level detection
    - Personality detection (?)
    - Period detection (dating)
- As opposed to topic / register / ...
Applications

• Forensic uses (doubtful with current state of the art)
• Customer Relations Management
• Semantic Web automatic meta-information assignment
• Plagiarism detection

Plagiarism Detection

• Current plagiarism detection software
  - based on string matching
  - has severely limited usefulness
    • only works when the plagiarized text is on the WWW or in a user database
• Solution: *Linguistic profiling* (Van Halteren, 2007, ACM TSLP) of an author
  - based on texts known to be written by him/her
  - text that doesn’t match the author’s linguistic profile is suspect
Automatic Text Categorization
(e.g. Sebastiani, 2002, Computing Reviews)

Documents → Bag of words / stemming / Stop list → Term selection (dimensionality reduction) → Document representations → Classifier Building (NB, k-nn, svm, …) → Classifier

Documents → Document topic

General Methodology

Documents → Document meta information

Feature construction (robust text analysis) → Feature selection → Document representations → Classifier Building (Discriminative supervised learning) → Classifier

Documents → Meta info
Milestone: Gender assignment
(Koppel, Argamon, Shimoni, Literary and Linguistic Computing, 2002)

- Documents: British National Corpus (fiction and non-fiction)
- Meta-data: gender of author
- Feature construction:
  - lexical (Function Words)
  - POS (Function Words)
- Supervised learning: linear separator
- Results: gender ~ 80% predictable

Gender Differences

- Use of pronouns (more by women) and some 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 use (by women) and more "informative" (descriptive) language use by men
- Even in formal language use!
- Strong correlation between male language use and non-fiction, and female language use and fiction
**Memory-Based Learning**

- “Lazy” Machine Learning method
  - Based on reuse of stored examples and similarity-based reasoning
- Right bias for natural language processing (?)
  - “Forgetting exceptions can be harmful”
  - Exceptions can be productive (what is “core” and what is “periphery”?)
  - Eager learning methods throw out the exceptions with the noise
Memory-Based Learning

- Basis: k nearest neighbor algorithm:
  - store all examples in memory
  - to classify a new instance $X$, look up the $k$ examples in memory with the smallest distance $D(X,Y)$ to $X$
  - let each nearest neighbor vote with its class
  - classify instance $X$ with the class that has the most votes in the nearest neighbor set
- Choices:
  - similarity metric
  - number of nearest neighbors ($k$)
  - voting weights

TiMBL (http://ilk.uvt.nl)

- Different feature weighting methods (information gain, gain ratio, chi-square, ...)
- Value distance weighting (modified value distance metric)
- Different distance weighting methods (linear, exponential decay, ...)
- Exemplar weighting
- ...
- Different experimental regimes and reporting options (confusion matrix, auc, f-scores, ...)
Memory-Based Shallow Parser: MBSP

- Combination of a number of modules: tokenization, POS tagging, chunking, relation finding, NER.
- Individual modules implemented as TiMBL and Mt servers.
- Versions:
  - Dutch:
    - based on CGN (Spoken Dutch corpus) and d-coi.
  - English:
    - based on Penn treebank (WSJ).
    - adapted from WSJ version for biomedical language using Genia corpus.

<DOCTYPE MBSP SYSTEM 'mbsp.dtd'>
<MBSP>
  <S cnt="s1"/>
  <NP rel="SBJ" of="s1_1">
    <W pos="DT">The</W>
    <W pos="NN" sem="cell_line">mouse</W>
    <W pos="NN" sem="cell_line">lymphoma</W>
    <W pos="NN">assay</W>
  </NP>
  <VP id="s1_1">
    <W pos="VBG">utilizing</W>
  </VP>
  <NP rel="OBJ" of="s1_1">
    <W pos="DT">the</W>
    <W pos="NN" sem="DNA_part">Tk</W>
    <W pos="NN" sem="DNA_part">gene</W>
  </NP>
  <VP id="s1_2">
    <W pos="VBZ">is</W>
    <W pos="RB">widely</W>
    <W pos="VBN">used</W>
  </VP>
  <NP rel="OBJ" of="s1_3">
    <W pos="TO">to</W>
    <W pos="VB">identify</W>
  </NP>
  <NP rel="OBJ" of="s1_3">
    <W pos="JJ">chemical</W>
    <W pos="NNS">mutagens</W>
  </NP>
  <W pos="period">.</W>
</MBSP>
Shallow parsing motivation

- We need efficiency and robustness as well as accuracy in stylometric feature construction
- Full (statistical) parsing
  - Provides too much information
    - Complex syntactic structure (less so for dependency parsing)
  - Provides too little information
    - Concepts
    - Semantic relations: “Who did what to whom, when, where and why?”

Genetic Algorithm Optimization

- Outcome of a ML experiment is based on
  - Algorithm bias (forgetting exceptions ...)
  - Sample size and selection
  - Feature selection and representation
  - Algorithm parameter tuning
  - All combinations of these factors
- Genetic Algorithm
  - Chromosome = experiment description (feature selection and algorithm parameter choice)
  - Fitness = f-score in 10-fold CV
  - Results in the experimental settings “best adapted” to the data.
### Author Detection

- **Long tradition in “humanities computing”**
  (Holmes, Burrows, Baayen, van Halteren, Juola, Hoover, ...)
- **Features:**
  - Word length, sentence length, n-grams, 
distribution of POS tags, frequencies of rewrite 
rules / chunks, word frequency, vocabulary 
richness, ...

### GA results on author detection

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Timbl memory-based learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default</td>
<td>69.4</td>
</tr>
<tr>
<td>Feature selection (fw-bw)</td>
<td>72.6</td>
</tr>
<tr>
<td>Parameter Optimization</td>
<td>70.5</td>
</tr>
<tr>
<td>GA joint FS and PO</td>
<td>80.1</td>
</tr>
</tbody>
</table>
Author Detection

- Text from Flemish national newspaper “De Standaard”
- National politics (similar genre and topic)

<table>
<thead>
<tr>
<th>Class</th>
<th>Training corpus</th>
<th># words</th>
<th>Test corpus</th>
<th># words</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (Anja Otte)</td>
<td>100 articles</td>
<td>57,682</td>
<td>34 articles</td>
<td>20,739</td>
</tr>
<tr>
<td>B (Bart Brinckman)</td>
<td>100 articles</td>
<td>54,479</td>
<td>34 articles</td>
<td>25,684</td>
</tr>
<tr>
<td>O (The Others)</td>
<td>100 articles</td>
<td>62,531</td>
<td>32 articles</td>
<td>21,871</td>
</tr>
</tbody>
</table>

Features

- Lexical (words with highest mutual information)
  - E.g. echter, evenwel (both: however)
- Function words
- Distribution of POS
  - Frequency of POS tags
  - Frequency of verb POS (basic and full)
  - Distribution of patterns of POS
    - E.g. DET NUM ADJ N
- Flesch-Kincaid readability
Results

- Using GA optimization, overall score on this data improved to ~ 80%, which seems to be a limit for this type of data
- Note that register and topics were kept relatively constant
- Accuracy / reliability too low for forensic applications
  - Contra e.g. Carole Chaski (UK), regularly asked in criminal cases, who claims > 99% accuracy when given 200 words of text
Personality Detection

- Are personality traits such as extraversion reflected in writing style?
- Seminal work by Gill & Oberlander on extraversion and neuroticism
  - Not in a prediction (classification) context but in a descriptive statistics context
  - Disregards effect of combinations of features
  - Based on e-mail

Previous hypotheses and observations

- Extraverts
  - Use fewer hedges (confidence)
  - More verbs, adverbs and pronouns (vs. nouns, adjectives, prepositions)
  - Less formal
  - Fewer negative emotion words more positive emotion words
  - Fewer hapaxes
  - More present tense verbs
  - Fewer negation and causation words
  - Fewer numbers and less quantification
  - Less concrete
Meyers-Briggs

- Meyers-Briggs Type Indicator (MBTI)
  - Based on Carl Jung personality typology
  - Classifies person’s preferred types
- Dichotomies:
  - extraversion / introversion [attitude, orientation]
  - sensing / intuition [perception, concrete vs. possible]
  - thinking / feeling [decision making style]
  - judging / perceiving [planned vs. spontaneous]
- Based on 93 forced-choice questions (2 options)

Meyers-Briggs

- Leads to 16 types: ENTJ (1.8%) ... ESFJ (12.3%)
- Mental functions: ST, SF, NT, NF
- Attitudes: TJ, TP, FP, FJ
- Temperaments: SP (artisan), SJ (guardian), NF (idealist), NT (rational)
- MBTI correlates with "Big Five" personality characteristics
  - extraversion and openness, to a lesser extent with
  - agreeableness and conscientiousness, but not with neuroticism
- Validity and reliability have been questioned
Data Collection (November 2006)

- 145 BA students (from a population of ~200) in a course on interdisciplinary linguistics
- Voluntarily watched the same documentary on Artificial Life (but 2 cinema tickets as incentive)
  - Topic held constant
- Wrote a text of ~ 1200 words
  - Factual description + Opinion
- Did an on-line MBTI-test
- Submitted their profile, the text and some user information via a web-site

Participant characteristics

- Too homogeneous for some experiments
  - 77% female
  - 97% native speaker of Flemish-Dutch
  - 77% from Antwerp region
- MBTI dichotomies:
  - E 80 vs. I 65
  - N 78 vs. S 67
  - F 105 (72%) vs. T 40
  - J 117 (81%) vs. P 28
Participant characteristics

28 ESFJ (provider)  6 ESFP
23 ENFJ (teacher !)  4 ISFP
16 ISFJ (protector)  4 INFP
15 INTJ (mastermind !)  4 ESTJ
15 INFJ
9 ENFP
8 ISTJ
8 ENTJ

Our typical student

• Flemish girl from around Antwerp who likes people and is warm, sympathetic, helpful, cooperative, tactful, down-to-earth, practical, thorough, consistent, organized, enthusiastic, and energetic. She enjoys tradition and security, and will seek a stable life that is rich in contact with friends and family
• (but she is not interested in Computational Linguistics) :-)

Universiteit Antwerpen
Features

- Lexical
  - Selection according to information gain, frequency
  - Binary or numeric features
  - Unigrams and bigrams
- Syntactic
  - POS
  - POS patterns (chunks)
  - Unigrams, bigrams, trigrams
- Readability
- Type / token (vocabulary richness)

Preliminary!

- Combination of all features give best results
- There are some tendencies, but most of the statements about extraversion that we could measure are not present significantly in our data
  - Except some POS effects
- I vs. E: 85.5% accuracy (baselines 55.2%, 50.5%); extraversion (86.3% f-score) > Introversion (84.7% f-score)
- N vs. S: 74.2%
- No results above baseline for other two dimensions
- No GA results yet, not all feature (combinations) implemented yet
POS distribution

• E > I
  - PRO, V, ADV
  - ! ?
• I > E
  • It is the (weighted) combination of many features that determines personality

Lexical Predictors for extraversion

• E
  - ! ?
  - you are
  - each other
  - therefore
  - etc
  - in fact
  - enormous
  - also
  - wanted

• I
  - close
  - external
  - decide
  - calculations
Discussion

- Extraversion is definitely predictable above chance from a combination of lexical and syntactic cues
- Unclear whether accuracy levels are high enough to make this useful beyond academic interest

“Stylogenesis”

- Cooperation with Edward Vanhoutte (literary scholar)
  - Postulates existence of “stylistic genomes” for periods / genres / authors
- Robust text analysis + Machine Learning (clustering here) as a methodology
Methodology

Documents
Memory-based shallow parsing
Feature selection
Document representations
Similarity-based clustering
Dendrograms

Author / period / gender

Data

- Representative samples of 100,000 words
- 50 English and American authors
- 12 time periods between 1525 and 1925

<table>
<thead>
<tr>
<th></th>
<th>Literary works</th>
<th>Non-literary work</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 male authors</td>
<td>2,500,000</td>
<td>Wall Street Journal</td>
</tr>
<tr>
<td>25 female authors</td>
<td>2,500,000</td>
<td></td>
</tr>
<tr>
<td>Number of words</td>
<td>100,000</td>
<td></td>
</tr>
</tbody>
</table>
### Feature Construction

- **Memory-Based Shallow Parser**

- **Feature vectors**
  - One feature vector per author
  - 3 labels: author, gender, period
  - Feature sets: type-token ratio, word length, readability, distribution of parts-of-speech, chunks & function words, NP & VP structure
Cluster Analysis & Interpretation

- Successful clusters
  - Cluster of men (21 out of 30, 70%)
  - Cluster of women (16 out of 20, 80%)
    outliers: Kipling, James, Trollope & Hardy

- Stylistic outliers
  - Wall Street Journal: genre & period
  - Hobbes, Behn, More, Mill & Defoe: period & genre
  - Behn: early female playwright
Cluster Analysis & Interpretation
Feature selection (gender clustering)

- 5 groups of gender clusters
- Wall Street Journal clusters with male authors
- George Eliot clusters with other women
- Female gender problem reduces from 9 to 6
- Stylistic outliers (22%)
  - 6 female authors (Stowe, Austin, Shelley, Ferber, Porter, Behn)
  - 5 male authors (Defoe, Collins, Trollope, James, Hardy)

What could this possibly be good for?

- Complementary to intuitive approach towards literary analysis
- Allows testing of specific literary style hypotheses
- New insights in literary style
- Close cooperation between literary scholars & computational linguists needed
Conclusions

- Robust text analysis + Machine learning is a powerful combination for inferring author characteristics from data
- Applications
  - in author detection & profiling "styleme" (plagiarism detection)
  - in author characteristic detection (period, region, gender, personality, ...)
  - in literary studies “stylogenetics”
- Results are encouraging but not really much more than that
- Text characteristics are the result of many factors. How do we focus on only those of interest?