Discourse features for author profiling

Ben Verhoeven & Walter Daelemans
CLiPS Research Center, University of Antwerp

University of Groningen, 1 April 2016
Author profiling

Predicting author’s sociological or psychological characteristics

- Age
- Gender
- Personality
- Education level
- Region of origin
- ...
Author profiling

Profile of this person

- 25
- Male
- Extraverted
- Highly educated
- Antwerp, Flanders
State of the art

**Supervised machine learning**
Binary or multi-class classification using SVMs

**Methodology**
Tenfold cross-validation

**Focus**
Short social media texts
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)
And deeper?

Discourse

Semantics
Discourse

What

- relations between sentences
- coherent structure
- situating text in the world

How

- discourse relational devices (DRD)
Currently, document representations for author profiling experiments are mostly limited to word-based features, sometimes utilising syntactic information. We are investigating whether discourse characteristics as features might improve the document representation. We hypothesize that groups of people with a common sociological or psychological factor (e.g. gender) might organise discourse in a similar way, e.g. by using similar discourse structures, similar connectives and similar ways of structuring text in space and time.
Discourse

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

Source
Extracted word lists from Dutch Wiktionary
- 1,300 adverbs
- 82 conjunctions

Annotation
Separate annotation for adverbs and conjunctions
- Intuitive annotation by one annotator
- Items can belong to multiple categories
Categorie: Voegwoord in het Nederlands

Pagina's in categorie "Voegwoord in het Nederlands"

Deze categorie bevat de volgende 82 pagina's, van in totaal 82.

A
- aangenomen
- aangezien
- al
- aleer
- alhoewel
- als
- alsmede
- alsof
- allang
- alvorens
- anders
- annex

B
- blijkens

C
- contravenen
- contras
- contraterende
- contraterend
- contratan
- contratera
- cohort

D
- daar
- daarentegen
- dan
- dan wel
- dat

E
- echter

F
- evengoed

G
- genoodzaakt

H
- hetzij
- hoewel

I
- indien
- ingeval

J
- juist

K
- kwaad

L
- later

M
- maar
- mits
- mitsdien

N
- naardiens
- naargelang
- naarmate
- nadat
- noch
- nochten
- nu

O
- opligtend

P
- passend

Q
- quoad

R
- reeds

S
- sinds
- stel

T
- teneinde
- tenzij
- tenwoord
- toen
- toer
- tot
- totdat

U
- uitgezonderd

V
- vermits
- voor
- vooraleer
- vooroord

W
- waar
Discourse

Categories for adverbs

Based on ANS Dutch grammar (Haeseryn et al., 1997)

- Place/direction: *opzij, nergens*
- Time: *pas, wanneer*
- Frequency: *soms, doorgaans*
- Grade/intensity: *erg, nogal*
- Quantification: *bijna, ook*
- Manner: *graag, anders*
- Modality: *misschien, wellicht*
- Negation: *nergens, niet*
- Conjunction: *immers, trouwens*
- Preposition: *buiten, onderin*
- Question: *hoe, wanneer*
Categories for conjunctions

Based on Penn Discourse Treebank tagset (PDTB Research Group, 2007)

- **TEMPORAL**
  - Synchronous: *terwijl*
  - Asynchronous: *alvorens, nadat*

- **CONTINGENCY**
  - Cause: *dankzij, want*
  - Condition: *aangezien, als*
Discourse

Categories for conjunctions

Based on 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*
Experiment

Gender prediction
Given a text, predict the gender of the author

Corpora

- Blogger corpus: 301,080 instances
- CLiPS Stylometry Investigation (CSI) corpus
  (Verhoeven & Daelemans, 2014)
  - Reviews: 1,298 instances
  - Essays: 517 instances (appeared not to be enough)
Experiment

Association analysis
Correlation analysis relating numerical to binary variable
- Predictors (numerical): relative counts of discourse categories
- Outcome (binary): gender

Logistic regression
- Relative features are normalized
- Fit binomial glm
- Coefficients converted to probabilities
- Computed 95% and 99% two-tailed confidence intervals for statistical significance
Results

Probabilities (of class 1)

- Reviews: between 0.070 and 0.407
- Blogs: between 0.229 and 0.337

So all are more related to female (since male = 1)

Interpretability

- Which gender uses which category more?
- How strong is the association?
Results

Corpus frequencies
Counts of each category per gender (per 10,000 words)

Conjunctions

<table>
<thead>
<tr>
<th></th>
<th>Reviews</th>
<th>Blogs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>Concession</td>
<td>152.04</td>
<td>160.92</td>
</tr>
<tr>
<td>Alternative</td>
<td>30.55</td>
<td>31.56</td>
</tr>
<tr>
<td>Exception</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Comparison</td>
<td>154.91</td>
<td>162.93</td>
</tr>
<tr>
<td>Condition</td>
<td>84.73</td>
<td>72.71</td>
</tr>
<tr>
<td>Expansion</td>
<td>360.65</td>
<td>373.41</td>
</tr>
<tr>
<td>Instantiation</td>
<td>11.46</td>
<td>10.04</td>
</tr>
<tr>
<td>Restatement</td>
<td>12.65</td>
<td>13.49</td>
</tr>
</tbody>
</table>
## Results

### Corpus frequencies

Counts of each category per gender (per 10,000 words)

### Adverbs

<table>
<thead>
<tr>
<th>Category</th>
<th>Reviews M</th>
<th>Reviews F</th>
<th>Blogs M</th>
<th>Blogs F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>745.18</td>
<td>764.88</td>
<td>296.49</td>
<td>294.60</td>
</tr>
<tr>
<td>Preposition</td>
<td>802.70</td>
<td>755.41</td>
<td>281.24</td>
<td>276.30</td>
</tr>
<tr>
<td>Question</td>
<td>48.93</td>
<td>56.84</td>
<td>15.19</td>
<td>14.81</td>
</tr>
<tr>
<td>Manner</td>
<td>576.90</td>
<td>567.70</td>
<td>210.53</td>
<td>211.55</td>
</tr>
<tr>
<td>Frequency</td>
<td>25.54</td>
<td>32.87</td>
<td>8.75</td>
<td>8.97</td>
</tr>
<tr>
<td>Negation</td>
<td>103.35</td>
<td>117.07</td>
<td>31.47</td>
<td>31.05</td>
</tr>
</tbody>
</table>
Conclusion

- Some significant associations, yet several are weak
- More significance in blogs (bigger corpus)
- Two categories relevant in both corpora: Condition and Preposition are both used more by men
- Frequency differences between corpora related to different genres
Evaluation of word lists

- Word lists from Wiktionary aren’t perfect
  - Many old, outdated words
  - Some obvious words missing
- Evaluate lists by comparing with corpora
  - Representativeness of Wiktionary?
Evaluation of word lists

Extraction of word lists from corpora

- Part-of-speech-tagged corpora (TwNC, CSI, Personae, PAN, Blogger, Netlog, ...)
- Extract adverbs and conjunctions with frequencies
- Cleaned version: deleted very infrequent words and obvious mistakes

How to compare them?

- What percentage of Wiktionary words actually occur in corpora?
- How many (and which) words in top of frequent words are not in Wiktionary?
### Evaluation of word lists

<table>
<thead>
<tr>
<th></th>
<th>Adverbs</th>
<th>Conjunctions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiktionary total</td>
<td>1,300</td>
<td>81</td>
</tr>
<tr>
<td># words in uncleaned corpora list</td>
<td>12,895</td>
<td>1,281</td>
</tr>
<tr>
<td># words in cleaned corpora list</td>
<td>1,800</td>
<td>214</td>
</tr>
<tr>
<td>% of Wiktionary in uncleaned corpus</td>
<td>65.5</td>
<td>55.0</td>
</tr>
<tr>
<td>% of Wiktionary in cleaned corpus</td>
<td>57.9</td>
<td>39.5</td>
</tr>
</tbody>
</table>

#### Adverbs

6 words from top 100 not in Wiktionary:

*meer, allemaal, waarom, echter, zelf, meest*

#### Conjunctions

5 words from top 50 not in Wiktionary:

*behalve, zonder, door, hoezeer, gelijk*
Thanks for your attention.

Ben Verhoeven  
CLiPS, University of Antwerp  
ben.verhoeven@uantwerpen.be