

Discourse features for author profiling

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

University of Groningen, 1 April 2016

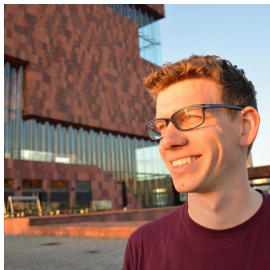
Author profiling

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)

Discourse

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

← → ↻ https://nl.wiktionary.org/wiki/Categorie:Voegwoord_in_het_Nederlands   

 Apps  Watchlist  Facebook  Argenta  Twitter  CLIPS  Ben  Scholar  deredactie.be  Paribas Fortis  Andere bladwijzers

 **WikiWoordenboek**
Het vrije woordenboek

Hoofdpagina
Recente wijzigingen
Nieuwe pagina's
Willekeurige woord
Willekeurige NL-woord
Woord begint met ...
Categorieën

Informatie
De kroeg
Hulp
Helpdesk
Financieel bijdragen

Zusterprojecten
Wikibooks
Wikipedia
Wikiquote
Wikisource
Commons

Afdrukken/exporteren
Boek maken
Downloaden als PDF
Printervriendelijke versie

Hulpmiddelen
Links naar deze pagina
Verwante wijzigingen

Niet aangemeld [Overleg](#) [Bijdragen](#) [Registreren](#) [Aanmelden](#)

Categorie: [Overleg](#) [Lezen](#) [Bewerken](#) [Geschiedenis weergeven](#) 

Categorie:Voegwoord in het Nederlands

 [Hulp](#)

Pagina's in categorie "Voegwoord in het Nederlands"

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

A	H	<ul style="list-style-type: none">sindsstel
<ul style="list-style-type: none">aangenomenaangezienalaleeralhoewelalsalsmedealsofalsookalthansalvorensandersannex	<ul style="list-style-type: none">hetzijhoewel I <ul style="list-style-type: none">indieningeval M <ul style="list-style-type: none">maarmitsmitsdien N <ul style="list-style-type: none">naardiennaargelangnaarmatenadatnochnochtansnu	T <ul style="list-style-type: none">teneindetenzijterwijltoentottotdat U <ul style="list-style-type: none">uitgezonderd V <ul style="list-style-type: none">vermitsvoorvooraleervoordat W <ul style="list-style-type: none">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*

Discourse

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

	Reviews			Blogs		
	M	F		M	F	
Concession	152.04	160.92		61.17	61.98	*
Alternative	30.55	31.56		14.39	14.31	**
Exception	0.00	0.00		0.0050	0.0035	*
Comparison	154.91	162.93		61.41	62.21	*
Condition	84.73	72.71	*	24.06	23.54	**
Expansion	360.65	373.41		149.22	149.33	**
Instantiation	11.46	10.04	*	2.776	2.712	
Restatement	12.65	13.49	*	3.293	3.218	

Results

Corpus frequencies

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

Adverbs

	Reviews			Blogs		
	M	F		M	F	
Place	745.18	764.88	*	296.49	294.60	
Preposition	802.70	755.41	**	281.24	276.30	**
Question	48.93	56.84		15.19	14.81	*
Manner	576.90	567.70		210.53	211.55	**
Frequency	25.54	32.87		8.75	8.97	*
Negation	103.35	117.07		31.47	31.05	**

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

	Adverbs	Conjunctions
Wiktionary total	1,300	81
# words in uncleaned corpora list	12,895	1,281
# words in cleaned corpora list	1,800	214
% of Wiktionary in uncleaned corpus	65.5	55.0
% of Wiktionary in cleaned corpus	57.9	39.5

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`