Authorship attribution of e-mail as a multi-class task

Notebook for PAN at CLEF 2011
Amsterdam, The Netherlands
21 September 2011

Kim Luyckx
CLiPS Computational Linguistics Group
University of Antwerp, Belgium

Authorship attribution

- Delicate balance between
  - Discriminative features & approach
  - Scalability: sensitivity to differences in author set size, data size, text length
- Text categorization approach
  - 1. features 2. discriminative learning
  - Common in the field
- Often binary SVM classifiers: one-vs.-all or one-vs.-one

Writing style

- Assumptions
  - identity interacts with writing style
  - aspects you are unconscious of
  - analysis of writing style allows us to identify the author
- Identity = mix of age, gender, personality, education level, ideology,…

Data set specifics

- SMALL and LARGE authorship identification scenarios
- Challenging materials (Enron E-mail Corpus)
  - Quite a large group of suspects (26 and 72, resp.)
  - Short texts (+/- 60 words/e-mail)
  - Skewed class distributions (10,000 words in 200 e-mails vs. 500 words in 10 e-mails)
  - Small-world data set but a lot of internal variation (meetings, financial information *etc*.)

Approach

- Pre-processing
  - Tokenization
  - Removed everything between <omni> </omni> tags
    - Lost training data for 2 authors in both scenarios
- Text categorization approach
  - Extract features & determine the most relevant ones
  - SVMs to build a model & test it on test data

Features

- CHR *n*-grams
- *n*-grams of LEX items
- DISC: *however, nevertheless, on the contrary*
- MOD: *can, could, would, shall*
- Ranking & selection
  - Chi-square for feature relevance ranking
  - Restricted to top-1000

SVM *multiclass*

- Joachims (1999,2002)
- Open-source
- Model all classes simultaneously, instead of one by one
- *C* 'soft margin parameter'
  - High C ~ hard-margin classification
  - Low C introduces a lot of training errors

Development results

- Without parameter tuning *C=5,000*
- Tuning of *C* yielded no significant difference in results

SMALL

	Macro F_1	Micro F_1
CHR3	**37.1**	**59.4**
LEX1	33.1	54.9
DISC	4.5	8.6
MOD	2.0	6.5
CHR-var	26.9	49.7
LEX-var	**34.0**	**57.3**
CHR+LEX	31.4	54.1

LARGE

	Macro F_1	Micro F_1
CHR3	27.3	40.6
LEX1	**28.8**	**42.2**
DISC	1.7	3.4
MOD	1.7	4.4
CHR-var	22.0	35.6
LEX-var	**31.2**	**46.1**
CHR+LEX	24.5	38.2

Test results

- Expectations
  - CHR3 > LEX-var in SMALL
  - LEX-var > LEX1 in Large

SMALL

	Macro F_1	Micro F_1
CHR3 (9/17)	34.3	62.0
LEX-var (6/17)	**37.1**	**64.2**
WINNERS	**47.5**	**71.7**

LARGE

	Macro F_1	Micro F_1
LEX1 (7/18)	34.0	50.0
LEX-var (9/18)	**34.2**	**52.2**
WINNERS	**52.0**	**65.8**

Which features are in LEX-var?

- Dates, locations
- Expressions of politeness (*thanks, regards, you soon*)
- E-mail specifics (*attached is*)
- Pronouns
- Argumentation (*for he*)
- Company names (*Reliant, Dominion, Enpower*)
- Domain-specific words (*pipeline*)

Conclusions

- *What is our ceiling?*
  - *What is humanly possible?*
  - *What is reasonably possible?*
- *Is it realistic to think we will get an answer?*
- *Severe lack of theory in the field*
  - *What is authorial style?*
  - *What do character* n-grams bring us?

Measuring writing style

- In *reality*, no one knows what writing style is
  - independent of the genre, register, topic?
    - can you recognize the author of a letter in a newspaper article?
  - independent of
    - the author's maturity in writing?
    - familiarity with the topic?
    - his/her mood?
- … consequences for validity of approaches suggested!

Contact

- kim.luyckx@ua.ac.be
- http://www.clips.ua.ac.be/~kim