Vote/Veto Meta-Classifier for Authorship Identification

Roman Kern Christin Seifert Mario Zechner
Michael Granitzer

Institute of Knowledge Management
Graz University of Technology
{*rkern, christin.seifert*}*@tugraz.at*
-
Know-Center GmbH
{*mzechner, mgrani*}*@know-center.at*

CLEF 2011 / PAN / 2011-09-22

Overview

Authorship Attribution System

- Preprocessing
  - Apply NLP techniques
  - Annotate the plain text
- Feature Spaces
  - Multiple feature spaces
  - Each should encode specific aspects
  - Integrate feature weighting
- Meta-Classifier
  - Base classifiers
  - Record performance while training
  - Selectively use the output for combined result

Preprocessing 1/4

Preprocessing Pipeline

- Preprocessing
  - Text lines - characters terminated by a newline
  - Text blocks - consecutive lines separated by empty lines
- Annotations
  - All consecutive annotations operate on blocks only
  - Natural language annotations
  - Slang-word annotations
  - Grammar annotations

Each document is treated separately from each other

Preprocessing 2/4

Natural Language Annotations

- OpenNLP
  - Split sentences
  - Tokenize
  - Part-of-speech tags
- Normalize to lower-case
- Stemming
- Stop-words
  - Predefined list
  - Heuristics (numbers, non-letter characters)

Preprocessing 3/4

Slang Word Annotations

- Smilies
  - `:-) :)  ;-) :-( :-> >:-> >;->`
- Internet Slang
  - `imho imm imma imnerho imnl imnshmfo imnsho imo`
- Swear Words
  - 

Very sparse, only a few documents contain such terminology

Preprocessing 4/4

Grammatical Annotations

- Apply parser component
  - Stanford parser

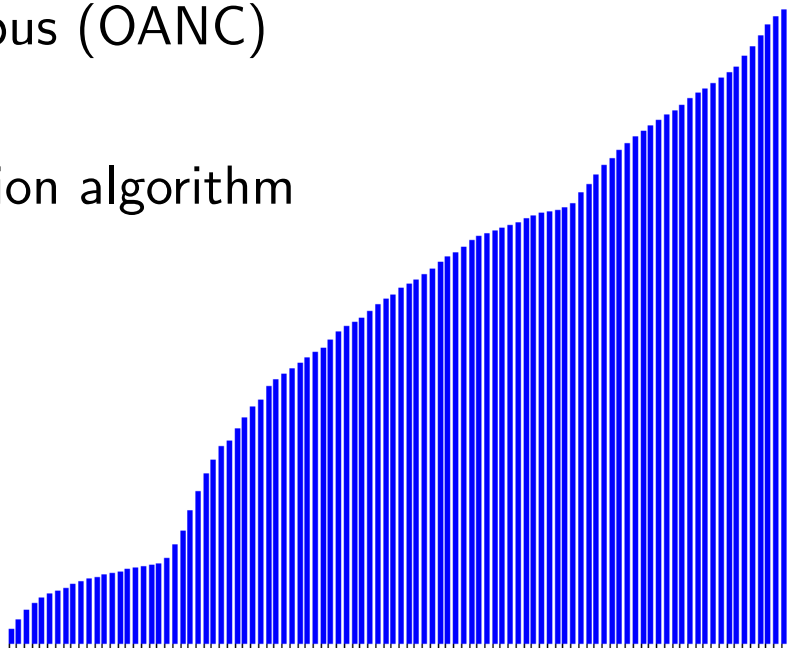
    Klein and Manning [2003]
- Sentence parse tree
  - Structure and complexity of sentences
- Grammatical dependencies
  - Richness of grammatical constructs

    de Marneffe et al. [2006]

Feature Weighting 1/2

Integrate External Resources

- External resources should give more robust estimations
- Word statistics
  - Open American National Corpus (OANC)
- Document splitting
  - Apply a linear text segmentation algorithm
    Kern and Granitzer [2009]
  - About 70,000 documents
    (instead of less than 10,000)
  - About 200,000 terms



Feature Weighting 2/2

Weighting Strategies

- Binary feature value
  - $w_{binary} = \operatorname{sgn} tf_x$
- Locally weighted feature value
  - $w_{local} = \sqrt{tf_x}$
- Externally weighted feature value
  - External corpus, modified BM25 Kern and Granitzer [2010]
  - $w_{ext} = \sqrt{tf_x} * \frac{\log(N - df_x + 0.5)}{df_x + 0.5} * \frac{1}{\sqrt{length}} * DP(x)^{-0.3}$
- Globally weighted feature value
  - Training set as corpus
  - $w_{global} = \sqrt{tf_x} * \frac{\log(N - df_x + 0.5)}{df_x + 0.5} * \frac{1}{\sqrt{length}}$
- Purity weighted feature value
  - Combine all document of an author into one big document
  - $w_{purity} = \sqrt{tf_x} * \frac{\log(|A| - af_x + 0.5)}{af_x + 0.5} * \frac{1}{\sqrt{length}}$

Feature Spaces 1/4

Feature Spaces Overview

- Statistical properties
  - Basic statistics
  - Token statistics
  - Grammar statistics
- Vector space model
  - Slang words $\mapsto$ linear
  - Pronouns $\mapsto$ linear
  - Stop words $\mapsto$ binary
  - Pure unigrams $\mapsto$ purity
  - Bigrams $\mapsto$ local
  - Intro-outro $\mapsto$ external
  - Unigrams $\mapsto$ external

Separate base classifier for each feature space, to be able to individually tune for each feature space

Feature Spaces 2/4

Basic Statistics Feature Space

IG	Feature Name	IG	Feature Name
0.699	text-blocks-to-lines-ratio	0.258	mean-text-block-token-length
0.593	text-lines-ratio	0.243	mean-tokens-in-sentence
0.591	number-of-lines	0.235	max-text-block-line-length
0.587	empty-lines-ratio	0.225	number-of-words
0.429	number-of-text-blocks	0.225	number-of-tokens
0.415	number-of-text-lines	0.207	max-text-block-char-length
0.366	max-words-in-sentence	0.191	number-of-sentences
0.337	mean-text-block-sentence-length	0.189	max-text-block-token-length
0.311	mean-line-length	0.176	number-of-stopwords
0.306	mean-text-block-char-length	0.174	mean-punctuations-in-sentence
0.298	mean-text-block-line-length	0.174	mean-words-in-sentence
0.294	capitalletterwords-words-ratio	0.145	max-tokens-in-sentence
0.292	capitalletter-character-ratio	0.133	number-of-punctuations
0.288	mean-nonempty-line-length	0.122	max-text-block-sentence-length
0.284	max-punctuations-in-sentence	0	number-of-shout-lines
0.278	number-of-characters	0	rare-terms-ratio
0.259	max-line-length		

Feature Spaces 3/4

Token Statistics Feature Space

IG	Feature Name	IG	Feature Name
0.25	token-PROPER_NOUN	0	token-PREPOSITION
0.2248	tokens	0	token-PARTICLE
0.1039	token-length	0	token-PRONOUN
0.0972	token-OTHER	0	token-length-18
0.0765	token-length-09	0	token-length-19
0.0728	token-length-08	0	token-NUMBER
0.0691	token-ADJECTIVE	0	token-CONJUNCTION
0.0691	token-length-ADJECTIVE	0	token-DETERMINER
0.0647	token-length-ADVERB	0	token-length-13
0.0646	token-length-07	0	token-length-14
0.0644	token-length-03	0	token-length-10
0.064	token-length-NOUN	0	token-length-12
0.0636	token-ADVERB	0	token-length-11
0.0614	token-length-VERB	0	token-UNKNOWN
0.0612	token-length-04	0	token-length-16
0.0583	token-length-05	0	token-PUNCTUATION
0.0581	token-length-06	0	token-length-02
0.0524	token-VERB	0	token-length-15
0.0465	token-NOUN	0	token-length-01
0	token-length-17		

Feature Spaces 4/4

Grammar Statistics Feature Space

IG	Feature Name	IG	Feature Name
0.1767	phrase-count	0.0654	relation-advmod-ratio
0.1659	sentence-tree-depth	0.0613	relation-dobj-ratio
0.1569	phrase-FRAG-ratio	0.0612	relation-complm-ratio
0.1538	relation-appos-ratio	0.0605	relation-advcl-ratio
0.15	phrase-S-ratio	0.059	phrase-ADVP-ratio
0.1477	phrase-NP-ratio	0.0585	phrase-INTJ-ratio
0.1165	phrase-VP-ratio	0.0545	relation-cop-ratio
0.1141	relation-nsubj-ratio	0.0525	relation-dep-ratio
0.087	phrase-PP-ratio	0.0523	relation-xcomp-ratio
0.086	phrase-SBAR-ratio	0.04	phrase-LST-ratio
0.0839	relation-prep-ratio	0	phrase-SBARQ-ratio
0.0838	relation-pobj-ratio	0	phrase-SINratio
0.0789	relation-cc-ratio	0	phrase-SQ-ratio
0.0779	relation-conj-ratio	0	phrase-WHADVP-ratio
0.0777	relation-nn-ratio	0	phrase-WHPP-ratio
0.0754	relation-det-ratio	0	phrase-WHNP-ratio
0.0745	relation-aux-ratio	0	relation-rcmod-ratio
0.0694	relation-amod-ratio	0	phrase-UCP-ratio
0.0672	relation-ccomp-ratio	0	phrase-X-ratio
0.0667	relation-mark-ratio		

Classification 1/2

Base Classifiers

- Open-source WEKA library
- Base classifier
  - Statistical feature spaces
    - Bagging with random forests
      Breiman [1996, 2001]
  - Vector space models
    - L2-regularized logistic regression, LibLINEAR
      Fan et al. [2008]

System would allow different classifiers and settings for each feature space

Classification 2/2

Meta Classifiers

- Training phase
  - Records the performance of all base classifiers during training
  - 10-fold cross-validation
  - If precision $> t_p$, the base classifier may vote for a class
  - If recall $> t_r$, the base classifier may veto against a class
- Classification phase
  - Apply all base classifiers, record posterior probabilities
  - If (may vote AND probability $> p_p$) $\rightarrow$ vote for this class
    - $W_c = W_c + (w_c^i \cdot p_c^i)$
  - If (may veto AND probability $< p_r$) $\rightarrow$ veto against this class
    - $W_c = W_c - (w_c^i \cdot p_c^i)$
  - The final base classifier is treated differently, the probabilities are directly added to the weights
  - Class with the highest W_c wins

Evaluation 1/5

Behavior of Base Classifiers (LargeTrain)

Classifier	#Authors Vote	#Authors Veto
basic-stats	4	14
token-stats	5	7
grammar-stats	5	5
slang-words	3	2
pronoun	6	1
stop-words	4	10
intro-outro	25	11
pure-unigrams	6	15
bigrams	20	23

There is an overlap between the classes the classifiers' vote/veto

Evaluation 2/5

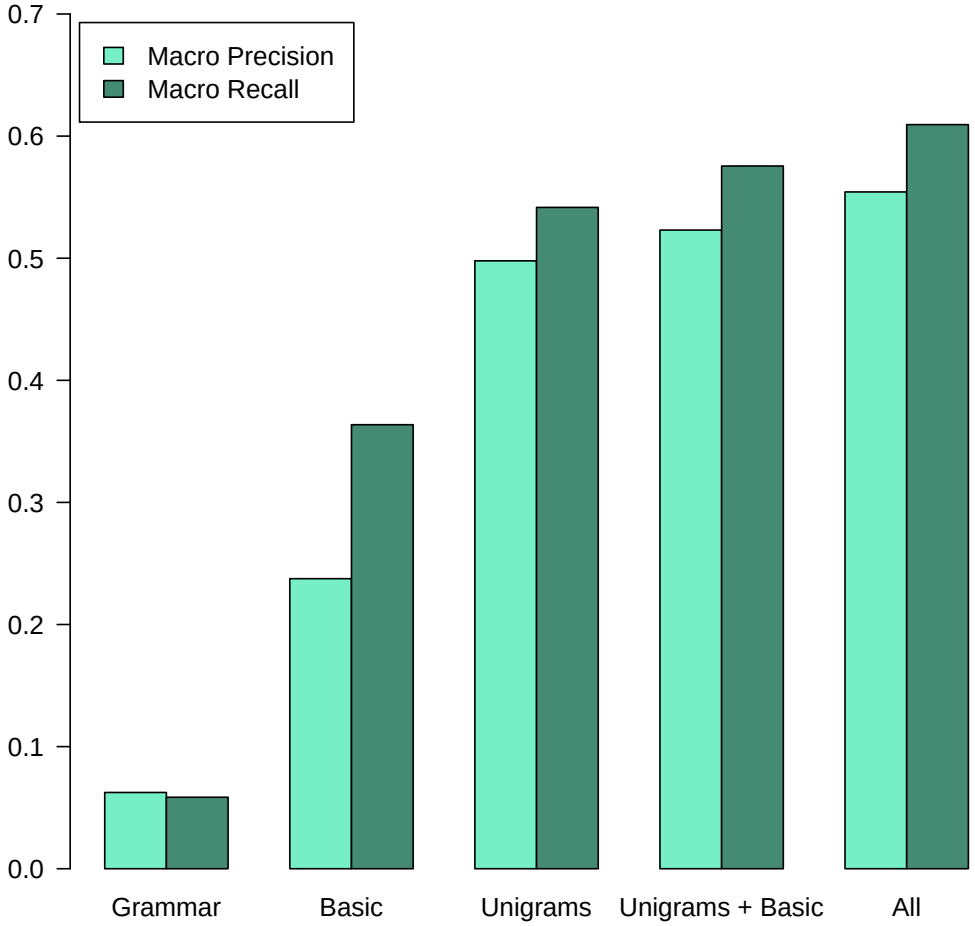
Performance of Base Classifiers (LargeValid)

Classifier	Vote Accuracy	Vote Count	Veto Accuracy	Veto Count
basic-stats	0.958	5141	1	252380
tokens-stats	0.985	1056	1	77492
grammar-stats	0.980	2576	1	89085
slang-words	0.819	94	0.997	9277
pronoun	-	0	1	85
stop-words	0.532	1924	0.998	107544
intro-outro	0.826	2101	0.998	102431
pure-unigrams	0.995	186	0.999	35457
bigrams	0.999	6239	1	281442

Thresholds appear to be far too strict

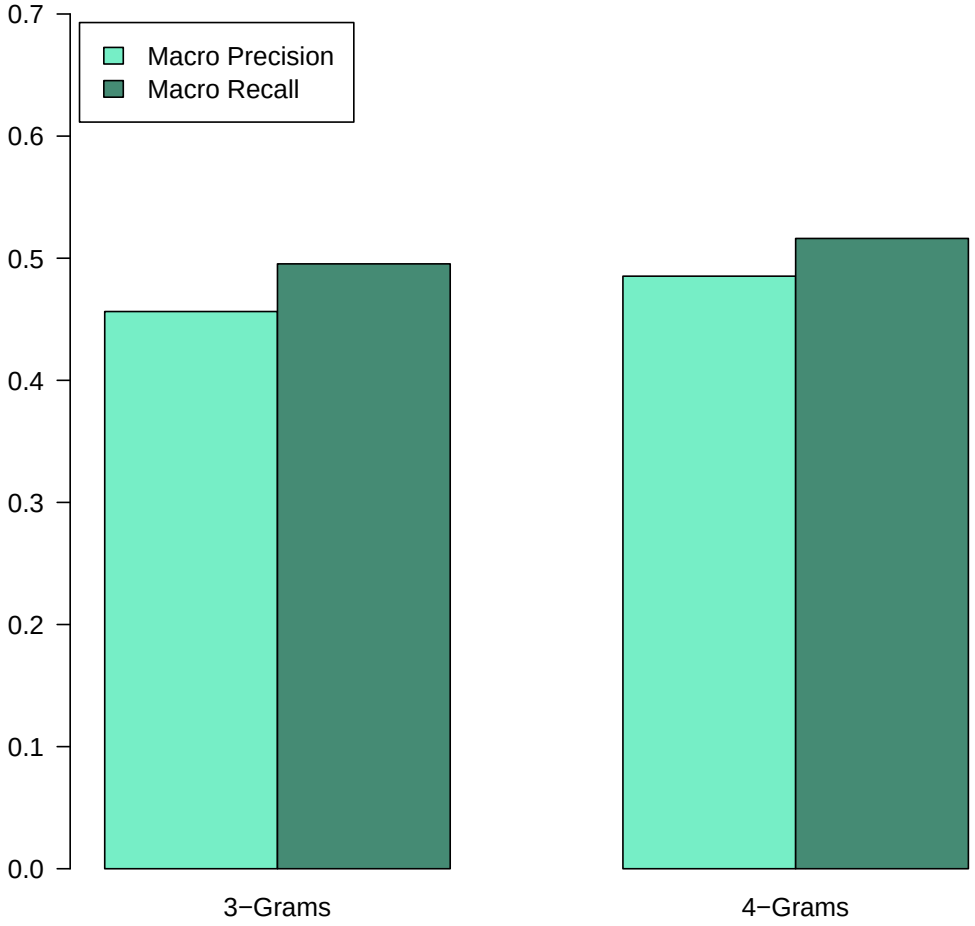
Evaluation 3/5

Performance of Selected Configurations (LargeValid)



Evaluation 4/5

Performance of Using Character n-Grams (LargeValid)



Evaluation 5/5

Performance of the System (Test)

Test Set	Micro Prec	Micro Recall	Micro F1	Rank
LargeTest	0.642	0.642	0.642	2
	-0.016	-0.016	-0.016	
LargeTest+	0.802	0.383	0.518	3
	+0.023	-0.088	-0.069	
SmallTest	0.685	0.685	0.685	5
	-0.032	-0.032	-0.032	
SmallTest+	1	0.095	0.173	8
	+0.176	-0.362	-0.415	

High precision, recall needs to be addressed

Conclusions

System overview

- Preprocessing pipeline tailored towards writing styles
- Large set of features and multiple feature-spaces
- Meta-classifier algorithm

Results

- "Topical" and layout features more important than "syntactical" features
- Room for improvements :)

The End

Thank you!

References

L. Breiman. Bagging predictors. *Mach. Learn.*, 24:123–140, August 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350.

L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324.

M. de Marneffe, B. MacCartney, and C. Manning. Generating typed dependency parses from phrase structure parses. In *LREC 2006*, 2006.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, 2008. ISSN 1532-4435.

R. Kern and M. Granitzer. Efficient linear text segmentation based on information retrieval techniques. In *MEDES '09: Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 167–171, 2009. doi: http://doi.acm.org/10.1145/1643823.1643854.

R. Kern and M. Granitzer. German Encyclopedia Alignment Based on Information Retrieval Techniques. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, editors, *Research and Advanced Technology for Digital Libraries*, pages 315–326. Springer Berlin / Heidelberg, 2010. doi: 10.1007/978-3-642-15464-5_32.

D. Klein and C. D. Manning. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL '03*, pages 423–430, 2003. doi: 10.3115/1075096.1075150.