Centre for
Data Analytics

Insight

Diachronic Text Classification with Character, Word, and Syntactic N-grams

Terry Szymanski and Gerard Lynch
Univeristy College Dublin

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN SFI Science Foundation Ireland

Related Work

- Stylometric text classification[1]
- Word epoch disambiguation[2]
- Temporal text ranking[3][4]
- Identifying period-specific language[5]
- Direct lookup[6]

---

[1] Argamon-Engelson et al. 1998. Style-based Text Categorization: What Newspaper am I Reading?

[2] Mihalcea and Nastase. 2012. Word Epoch Disambiguation: Finding How Words Change Over Time.

[3] Niculae et al. 2014. Temporal Text Ranking and Automatic Dating of Texts

[4] Zampieri et al. 2015. AMBRA: A Ranking Approach to Temporal Text Classification

[5] Salaberri et al. 2015. IXAGroupEHUDiac: A Multiple Approach System towards the Diachronic Evaluation of Texts

[6] Tan and Ordan. 2015. USAAR-CHRONOS: Crawling the Web for Temporal Annotations

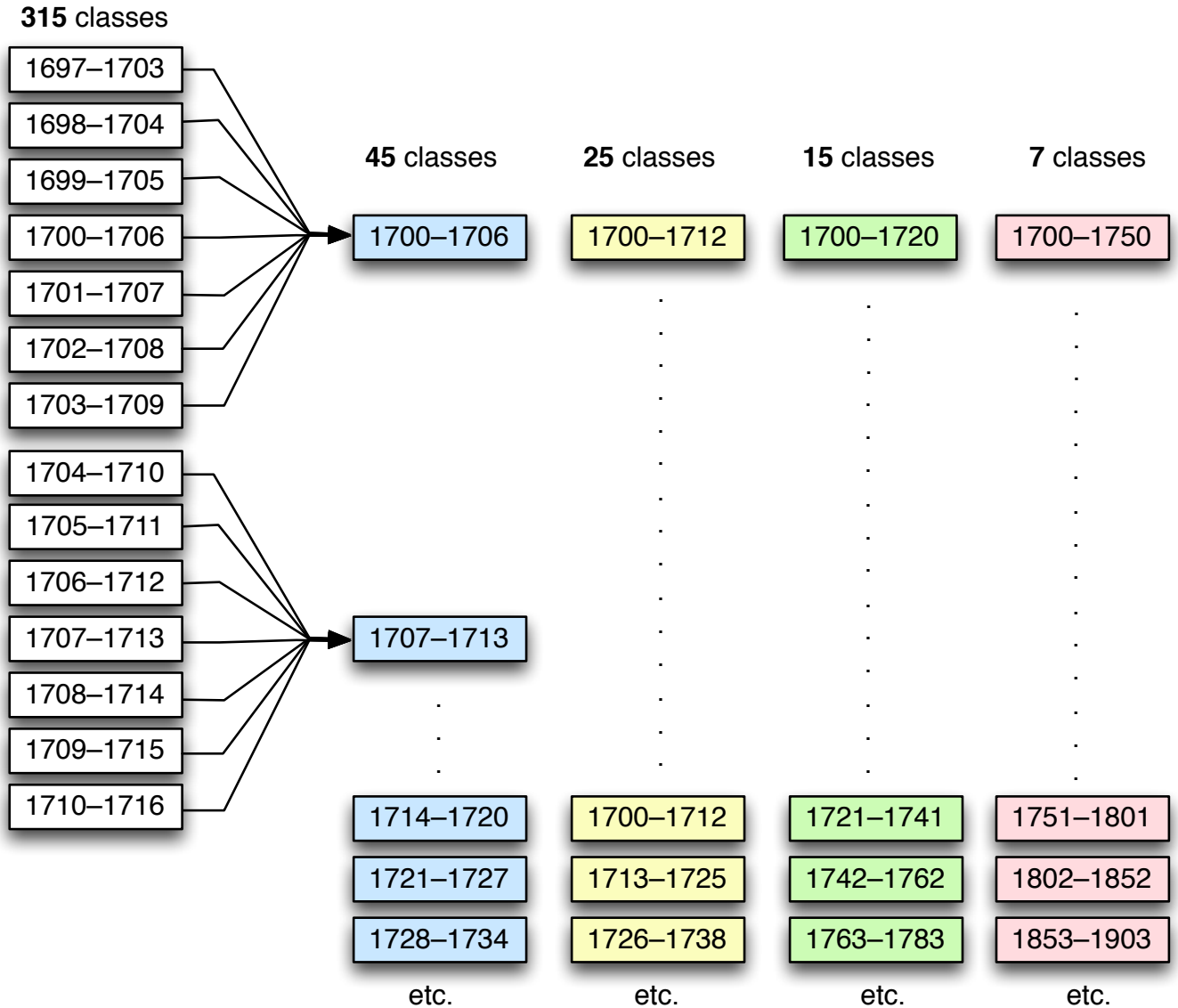
Overview

1. **Focus on language style**
   (Subtask 2: texts with specific time language usage)
2. **Treat dating as multiclass classification**
   (Weka SMO defaults: 1-vs-1, polynomial kernel)
3. **Use features extracted from text**
   (orthographic, lexical, syntactic)
4. **Augment with features from external data**
   (Google syntactic n-grams database)

Year-range Classes

Insight

- Label each text using non-overlapping year-ranges / classes.



Five Sets of Features

CPWS features extracted from text, parsed with Stanford CoreNLP, filtered by a minimum frequency cutoff.

Character n-grams	e.g.	`e` ; `&_c` ; `u_<COMMA>_<SPACE>`
Part-of-speech n-grams	e.g.	`NN` ; `VBZ_VBD` ; `RB_._<COMMA>`
Word n-grams	e.g.	`million` ; `towards_the of_his_majesty's`
Syntactic phrase-structure rules	e.g.	`S→S_,_NP_VP_.` `RB→'now'`

Google features from *nodes* dataset, counts smoothed over five neighboring counts and normalized by year. $p(y|\mathbf{w})$ normalized to [0,1] for each text.

Google syntactic n-grams	e.g.	`here/RB/nsubj/0` `elizabethans/NNPS/pobj/0`

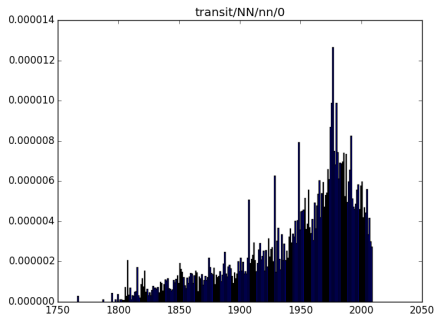
Google Syntactic N-grams

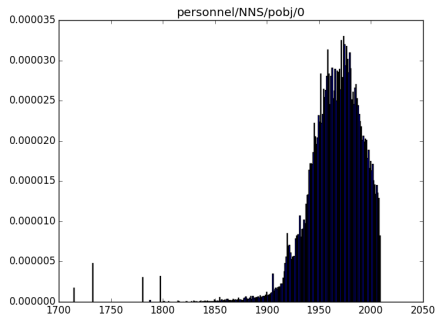
Insight

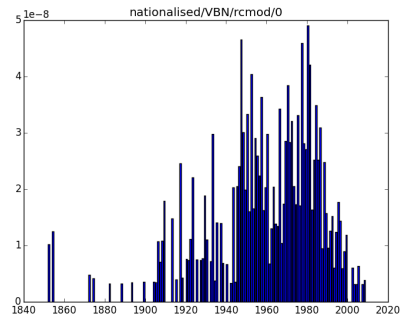
Naive Bayes: estimate $p(y|\mathbf{w})$ for each year. Assume $p(y) = C$.

$$y^* = \arg\max_y \frac{p(\mathbf{w}|y)p(y)}{p(\mathbf{w})} \approx \arg\max_y \prod_i p(w_i|y)$$

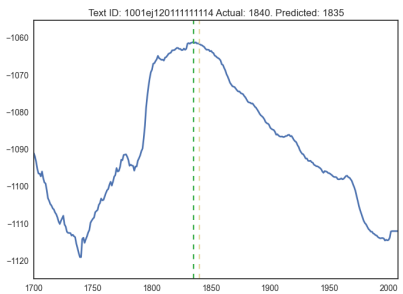
$p(w|y)$ – probability of an individual syntactic n-gram:

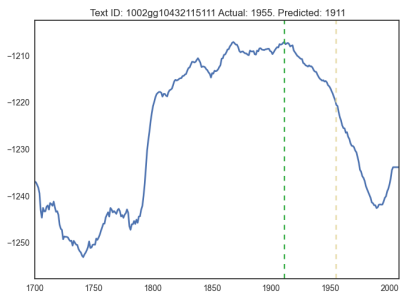


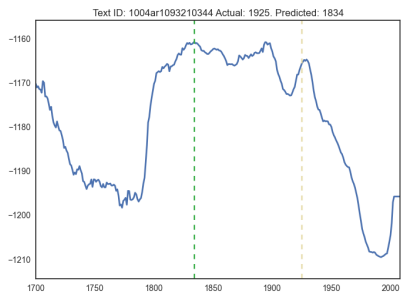




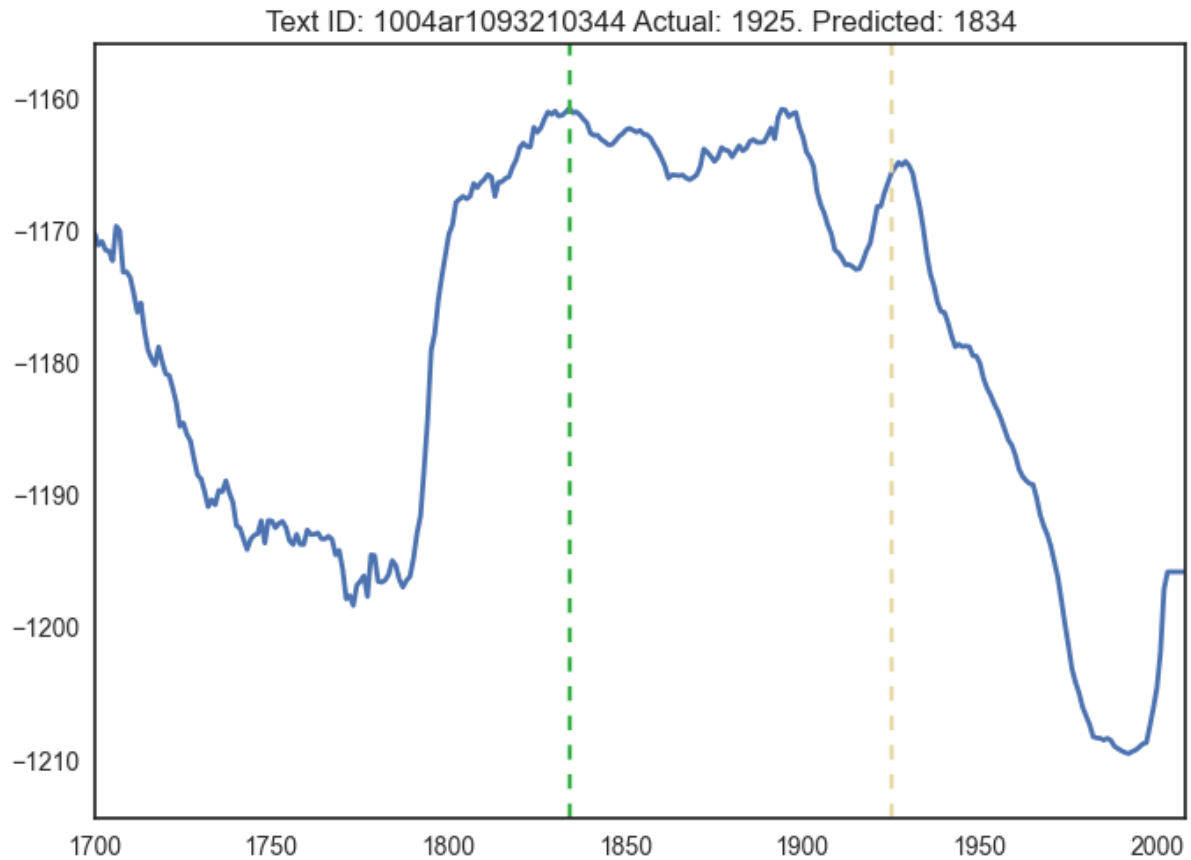
$\prod_i p(w_i|y)$ – probability of a text:







Naive Bayes Predictions



Mrs. Lambart tells about her grandfather, one of the last of the Roumanian Boyars, Konstantin Kretzulesco. He was a scholar and an eccentric ; one of his eccentricities " took the form of an undeviating rectitude of conduct "Roumanian politics were " Oriental in their convolutions." During one political upheaval he was offered the crown. But he refused it for reasons which he considered good and sufficient. He appeared on the balcony, rather bewildered and rather impatient at being so violently recalled to the World of Reality. What was all this noise about ? Why did they disturb him ? And when it was explained : " My children, I have no time to spare for you and your affairs. You have interrupted me. I am at this moment engaged in reading the Bible in English. Ah, what a Poem ! What a Monument of Literature ! Do not detain me. I can do nothing for you. Leave me, leave me." They left him accordingly, and he returned to his study of the Scriptures in the language of the Elizabethans.

Cross-Validation Accuracy

Percent of labels predicted correctly.

	System	(k=45) 6-Year	(k=25) 12-Year	(k=15) 20-Year	(k=7) 50-Year
Task data only:	$\texttt{Baseline}_Y$	10.4	12.6	20.5	36.6
	$\texttt{Char}_{svm}$	36.1	38.4	47.9	64.5
	$\texttt{POS}_{svm}$	24.6	26.8	36.3	53.6
	$\texttt{Word}_{svm}$	26.1	29.6	37.2	54.6
	$\texttt{Syn}_{svm}$	23.4	26.3	38.5	54.6
	$\texttt{CPWS}_{svm}$	**36.9**	**40.1**	**50.7**	**67.8**
External data only:	$\texttt{Baseline}_R$	2.2	4.0	6.7	14.3
	$\texttt{Google}_{nb}$	**10.9**	**18.7**	**31.7**	**52.4**
All data:	$\texttt{Google}_{svm}$	26.7	31.0	43.1	65.0
	$\texttt{CPWS+G}_{svm}$	**41.5**	**45.9**	**55.3**	**73.3**

System Comparison

Task 7 Systems:

	F (6-year)			M (12-year)			C (20-year)		
Team	acc	prec	err	acc	prec	err	acc	prec	err
BaseL	.224	.000	NA	.391	.000	NA	.542	.000	NA
IXA	.262	.038	41	.428	.068	65	.622	.098	75
AMBRA	.605	.143	23	.768	.143	29	.868	.292	29
UCD	**.759**	**.463**	**14**	**.847**	**.473**	**19**	**.910**	**.542**	**19**

acc : task7 weighted score

prec : % correct labels

err : mean distance from gold in years

Mihalcea et al. 2012 Word-Epoch Disambiguation:

	(50-year)	
	baseline	system
WED	.43	.62
UCD	.37	.73

Conclusions

- **Muticlass classification** *seems* to work better than regression, ordinal classification, or ranking.
- **Character n-grams** are highly effective features for diachronic classification.
- **External features** from a large, cross-domain dataset can improve classification, but:
- The **prior distribution** over date-labels has a significant domain-specific effect.

Future Work

AKA stuff we didn't have time to do:

- different features (normaliziation, stylometric features, higher-order syntactic n-grams)
- different learners (regression, ordinal classification)

The long view:

- Date texts of unknown provenance
- Identify and explain time-specific features
  - (Ideally) linguistically satisfying, interesting, unexpected phenomena

Thank you!

Top CPWS features ranked by information gain

Insight

- *'d* is an archaic past tense verb marker.
- Many features have no obvious significance, but are highly frequent.

Rank	Attribute	Type	Rank	Attribute	Type
1	NN	P-1	10.9	t	C-1
2	i	C-1	11.7	l	C-1
3.5	␣	C-1	13.3	o	C-1
3.9	. → .	S	13.8	.	W-1
5	ROOT → S	S	15.7	[' d]	C-2
5.8	a	C-1	15.9	[␣ .]	C-2
7.3	e	C-1	16.3	r	C-1
8.1	.	C-1	18.6	[JJ NN]	P-2
8.5	n	C-1	19.3	JJ	P-1
10.7	s	C-1	19.8	c	C-1

Cross-Validation Accuracy