CIS 530 Fall 2015 Project Report

Authorship Attribution

Mao-Hsu Chen

December 17th, 2015

Overview

- Given training data labeled as 1 if the author is Gina Kolata and as 0 otherwise
- Authorship attribution task
    - Distinguish the authorship of new excerpts in the unlabeled testing data

Overview

- Given training data labeled as 1 if the author is Gina Kolata and as 0 otherwise
- Authorship attribution task
    - Distinguish the authorship of new excerpts in the unlabeled testing data

→ A semi-supervised binary classification task

Main idea

Stylometric features

- Differentiate Kolatas style of writing from those of other authors
- Topic/genre-independent
- Contextual/structural information matters more than content information

Main idea

Stylometric features

- Differentiate Kolatas style of writing from those of other authors
- Topic/genre-independent
- Contextual/structural information matters more than content information
- Potential style markers
    - Lexical: **most frequent words, function words** (Chung and Pennebaker, 2007)
    - Character: **character n-grams**
    - Syntactic: **part-of-speech**

Parameter manipulation

- The first dozens of most common words of a corpus are usually dominated by closed class words, many of which are function words, whereas open class words become majority after a few hundred words (Stamatatos, 2009).

Parameter manipulation

- The first dozens of most common words of a corpus are usually dominated by closed class words, many of which are function words, whereas open class words become majority after a few hundred words (Stamatatos, 2009).

→ How many most frequent non-function words should be included in the feature space?

→ How many *n*-grams should be used as features?

→ How big is *n*?

Approach

- Start with the baseline SVM using the top-1000 most common words in the training data as features.
  - Use `libsvm` library for the binary-class prediction SVM with a linear kernel
- Supplement with other stylometric features
  - Function words: `stopwords.txt` from HW4
  - Character *n*-grams: $n = 3$, **4**, or 5
  - POS tags: Stanford Part of Speech tagger with the mapping `en-ptb.map` to Google Universal tagset from HW3
- 9-fold cross-validation to avoid overfitting the training data

Feature vector of the final system

3 sets of features:

1. (Stop words) $\cup$ (top-1000 most common words)
2. Top-15,000 4-grams
3. 12 POS tags: % of each tag within an excerpt

For each of the first two feature sets:

- Logarithmic relative frequency was calculated for each term frequency
    - $F_{log}(w_k, d_i) = log(1 + f(w_k, d_i)/f(d_i))$
- Inverse document frequency weighting
- Euclidean (L_2) normalization of the resulting excerpt vector
    - $d_i^* = \dfrac{d_i}{\| d_i \|_{L_2}}$, where $\| x \|_{L_p} = (\sum_i |x_i|^p)^{1/p}$ is the p-norm

Experiments

Leaderboard team name: *lingolingoling*

Accuracy	Train	Test	Parameter anipulation					
			norm	n-gram	# of n-grams	# of mfw	stop words	POS
other 1	86.6011%	-	L2	-	-	100	-	-
other 2	88.5412%	-	L2	-	-	1000	-	-
other 3	87.5011%	-	L2	3	1000	-	-	-
other 4	**88.2936%**	-	L2	**4**	1000	-	-	-
other 5	87.6413%	-	L2	5	1000	-	-	-
other 6	79.1381%	-	na	4	1000	1000	-	-
other 7	79.1877%	-	L1	4	1000	1000	-	-
1st sub.	**88.970%**	87.7415%	**L2**	4	1000	1000	-	-
other 8	88.7890%	-	L2	4	1000	-	v	-
other 9	89.0614%	-	L2	4	1000	500	v	-
other 10	**89.3586%**	-	L2	4	1000	**1000**	v	-
other 11	89.6805%	-	L2	4	1500	1000	v	-
other 12	90.5142%	-	L2	4	2000	1000	v	-
2nd sub.	91.3398%	90.9361%	L2	4	3000	1000	v	-
3rd sub.	-	93.7593%	L2	4	15000	1000	v	-
4th sub.	-	93.4621%	L2	4	20000	1000	v	-
5th sub.	-	94.4279%	L2	4	20000	1000	v	v
Final	93.8414%	93.4621%	L2	4	15000	1000	v	v

Table: Performance and parameter manipulation of all the experiments.

Parameter manipulation of k & n

Accuracy	Train	Parameter manipulation			
		norm	n-gram	# of ngrams	# of mfw
other 1	86.60%	L2	-	-	100
other 2	**88.54%**	L2	-	-	**1000**
other 3	87.50%	L2	3	1000	-
other 4	**88.29%**	L2	**4**	1000	-
other 5	87.64%	L2	5	1000	-

Table: Parameter manipulation of the top-k most frequent words (mfw) and the number n for ngrams.

Parameter manipulation of normalization, k & stop words

Accuracy	Train	Test	Parameter manipulation		
			norm	# of mfw	stop words
other 6	79.14%	-	na	1000	-
other 7	79.19%	-	L1	1000	-
1st sub.	**88.97%**	87.74%	**L2**	1000	-
other 8	88.79%	-	L2	-	v
other 9	89.06%	-	L2	500	v
other 10	**89.36%**	-	L2	**1000**	**v**

Table: Parameter manipulation of the normalization method, the top-k most frequent words (mfw), and the use of the stop words list. Note that all these attempts used top-1000 most frequent 4-grams as features.

Parameter manipulation of # of ngrams & POS

Accuracy	Train	Test	Parameter manipulation # of ngrams	POS
other 11	89.68%	-	1500	-
other 12	90.51%	-	2000	-
2nd sub.	91.34%	90.94%	3000	-
3rd sub.	-	**93.76%**	**15000**	-
4th sub.	-	93.46%	20000	-
5th sub.	-	**94.43%**	20000	v
Final	93.84%	93.46%	15000	v

Table: Parameter manipulation of the number of ngrams and the use of POS tags. Note that all these attempts used top-1000 most frequent 4-grams and the stop words list as features and L2 normalization.

Discussion and Analysis

- With the capability of managing a large number of dimensions, SVM is ideal for the classification here when the number of stylometric features was unknown.
- SVM: the more features are included the better the performance would be, as exemplified in our experimental results.
- The achieved high performance (over 90%) only implies an integral success of all features used but fails to identify the contribution of individual feature set.
- There always exists a difference between the training and the testing data sets, especially when the latter is unlabeled, which is often the case in real-world application.

References

Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, pages 343–359, 2007.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

Thank you