

Computational Extraction of Social and Interactional Meaning from Speech

HW3 Genre Classification

July 29th, 2011

Performance

Accuracy	Classifier	Features
87.00	ME	Bag-of-word, pos-tag
86.99	SVM	Log-probability of word-type
84.96	ME	Bag-of-word
83.97	ME	Bag-of-word, pos-tag, type-token ratio
82.79	NB	Word-count, pos-tag
82.75	NB+GT	Bag-of-word (from Jinho's data)
82.61	NB+GT	Bag-of-word
82.61	TF-IDF	Bag-of-word
82.61	TF-IDF	Bag-of-word
81.00	ME	Word-count, pos-count, punctuation-count
79.69	NB	Bag-of-word
76.30	NB	Bag-of-word (5-10%), word-length
74.58	NB	Bag-of-word
73.17	NB	Bag-of-word (top 2000)
72.10	SVM	Bag-of-word (cutoff)
71.14	NB	#(punctuation, words-in-line, references, keywords)

More

- Machine learning tools
 - Rainbow (6 people)
 - Mallet (4 people)
 - NLTK (2 people)
 - SVM-light (1 person)
 - Weka (1 person)
 - R (2 person)
- Error Analysis
 - 2 people did error analysis on the automatic output of my system.

Features

- Bag-of-word
 - The occurrence of each word type.
 - If a word type appears in the document, assign a value of 1; otherwise, assign a value of 0.
- Bag-of-word + stopwords
 - Discard word types that are not so helpfu.
 - e.g., a, the, do, it, they, ...

Features

- Word-count
 - The count of each word type.
 - Instead of assigning a binary value, assign its count.
- Word-count + cutoff
 - Discard word types that appear less than the cutoff.
- Word-count + percentile
 - Keep only word types within certain percentiles.
 - e.g., discard top 10% of word-types.
 - e.g., discard bottom 50% of word-types.

Features

- Probability of word-type
 - Probability of a word-type given a document.
 - $P(w_i|d) = \#(w_i) / \sum_{k \in d} \#(w_k)$
- Conditional probability of word-type
 - Probability of a word-type given all documents.
 - $P(w_i|d_j) * P(d_j)$
 - $P(d_j) = \#(d_j) / \sum_{k \in D} \#(d_k)$

Features

- POS tags
 - Give (word, word_pos) as a feature
 - Give (word, pos) as a feature
- Token type ratio
 - Hypothesis: polished writing such as news articles would tend to include more diverse vocabulary than user-generated reviews and blogs.

	AA	AB	BC	BN	CTS	NW	WB
Mean	0.73	0.71	0.44	0.45	0.34	0.49	0.76
SD	0.11	0.12	0.04	0.04	0.02	0.05	0.16

Features

- Others
 - Word length
 - # of words in line
 - # of references
 - # of keywords
 - # of each punctuation

Confusion Matrix

- NB: 82.79%

	AA	AB	BC	BN	CTS	NW	WB	Total	%
AA	110	0	0	0	0	0	0	110	100.00%
AB	6	104	0	0	0	0	0	110	94.55%
BC	0	0	76	39	1	0	0	116	65.52%
BN	0	1	24	66	1	0	0	92	71.74%
CTS	0	0	0	0	101	0	0	101	100.00%
NW	4	13	1	0	0	82	8	108	75.93%
WB	11	6	6	2	4	0	72	101	71.29%
	110	104	76	66	101	82	72	738	82.79%

Confusion Matrix

- ME: 84.96%

	AA	AB	BC	BN	CTS	NW	WB	Total	%
AA	108	2						110	98.18%
AB	12	98						110	89.09%
BC			76	40				116	65.52%
BN			45	47				92	51.09%
CTS					101			101	100.00%
NW		3				105		108	97.22%
WB	9						92	101	91.09%
	108	98	76	47	101	105	92	738	84.96%

Confusion Matrix

- SVM: 86.99%

	AA	AB	BC	BN	CTS	NW	WB	Total	%
AA	109	0	0	0	0	0	1	110	99.09%
AB	9	95	1	0	0	0	5	110	86.36%
BC	0	0	88	27	0	1	0	116	75.86%
BN	0	0	41	48	0	3	0	92	52.17%
CTS	0	0	0	0	101	0	0	101	100.00%
NW	0	1	1	0	0	106	0	108	98.15%
WB	2	3	0	0	0	1	95	101	94.06%
	109	95	88	48	101	106	95	738	86.99%