Day 7:Automated dictionary-based approaches

Kenneth Benoit

Essex Summer School 2012

July 17, 2012

Rationale for dictionaries

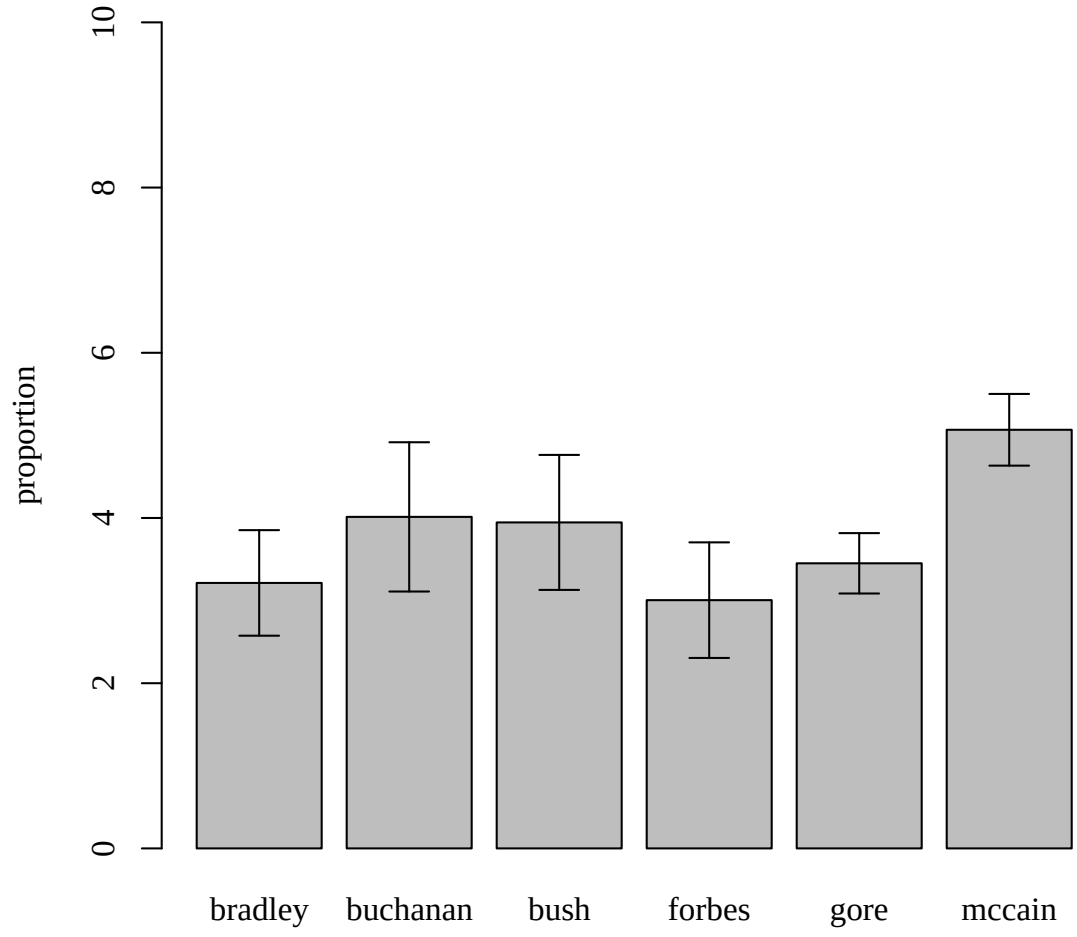
- Rather than count words that occur, pre-define words associated with specific meanings
- Another move toward the fully automated end of the text analysis spectrum, since involves no human decision making as part of the text analysis procedure
- Frequently involves lemmatization: transformation of all inflected word forms to their "dictionary look-up form" — more powerful than stemming
- Example: General Inquirer codes *I*, *me*, *my*, *mine*, *myself* as self, and *we*, *us*, *our*, *ours*, *ourselves* as selves

Well-known dictionaries: General Inquirer

- General Inquirer (Stone et al 1966)
- Maps texts to counts from an extensive dictionary
- Latest version contains 182 categories – the "Harvard IV-4" dictionary, the "Lasswell" dictionary, and five categories based on the social cognition work of Semin and Fiedler
- Examples: "self references", containing mostly pronouns; "negatives", the largest category with 2291 entries
- Uses stemming
- Also uses disambiguation, for example to distinguishes between *race* as a contest, *race* as moving rapidly, *race* as a group of people of common descent, and *race* in the idiom "rat race"
- Output example: `http://www.wjh.harvard.edu/~inquirer/Spreadsheet.html`

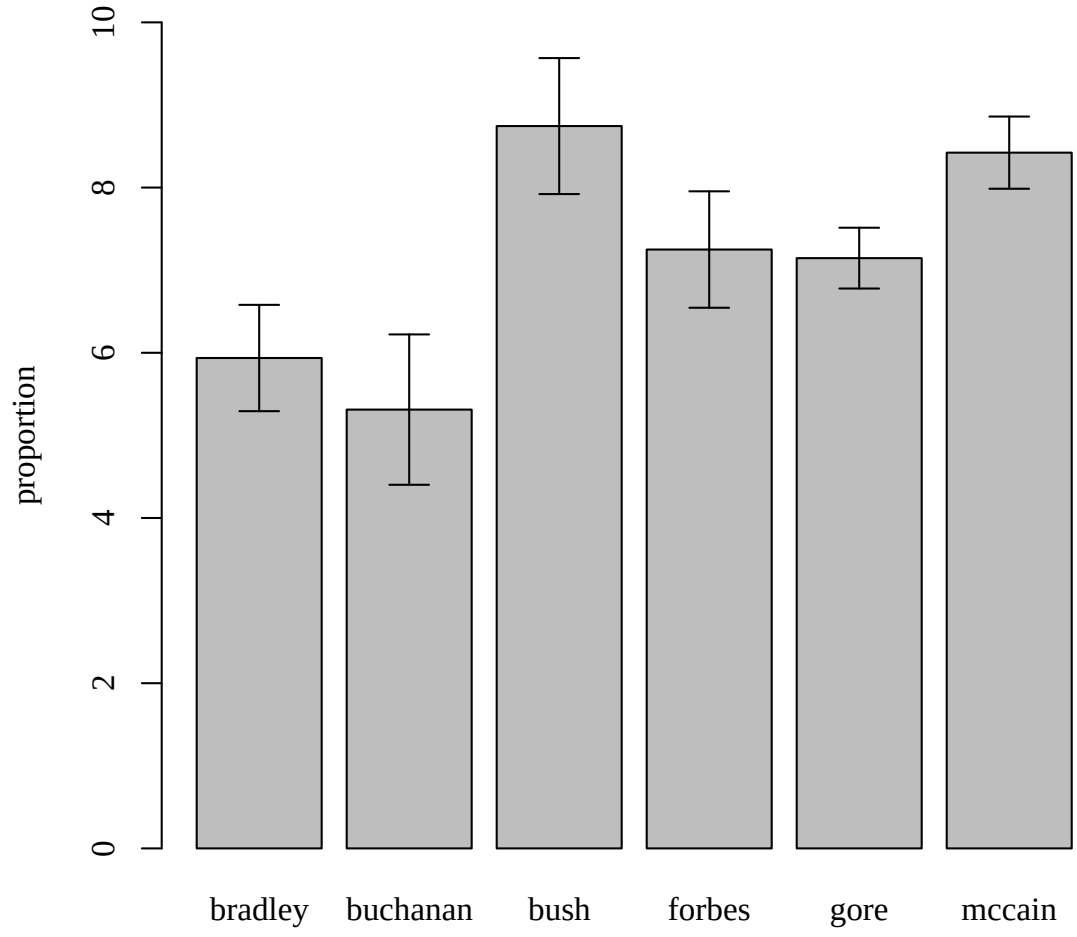
General Inquirer Applied to US Presidential Candidate Speeches (2000)

Negative language



General Inquirer Applied to US Presidential Candidate Speeches (2000)

Positive language



Well-known dictionaries: Regressive Imagery Dictionary

- Consists of about 3,200 words and roots, assigned to 29 categories of primary process cognition, 7 categories of secondary process cognition, and 7 categories of emotions
- designed to measure primordial vs. conceptual thinking
  - Conceptual thought is abstract, logical, reality oriented, and aimed at problem solving
  - Primordial thought is associative, concrete, and takes little account of reality – the type of thinking found in fantasy, reverie, and dreams
- Categories were derived from the theoretical and empirical literature on regressive thought by Martindale (1975, 1990)

Regressive Imagery Dictionary categories

- Full listing of categories

1 orality	21 brink-passage	41 aggression	62 novelty
2 anality	22 narcissism	42 expressive behaviour	63 negation
3 sex	23 concreteness	43 glory	64 triviality
4 touch	24 ascend	44 female role	65 transmute
5 taste	25 height	45 male fole	
6 odour	26 descent	46 self	
7 general sensation	27 depth	47 related others	
8 sound	28 fire	48 diabolic	
9 vision	29 water	49 aspiration	
10 cold	30 abstract thought	50 angelic	
11 hard	31 social behaviour	51 flowers	
12 soft	32 instrumental behaviour	52 synthesize	
13 passivity	33 restraint	53 streight	
14 voyage	34 order	54 weakness	
15 random movement	35 temporal references	55 good	
16 diffusion	36 moral imperative	56 bad	
17 chaos	37 positive affect	57 activity	
18 unknown	38 anxiety	58 being	
19 timelessness	39 sadness	59 analogy	
20 counscious	40 affection	61 integrative con	

- More on categories:
  `http://www.kovcomp.co.uk/wordstat/RID.html`

Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory
- Exact dictionary is proprietary (e.g. *secret*) but you can view a summary here:
  `http://www.liwc.net/descriptiontable1.php`

Example: Terrorist speech

	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Controls N = 17	p (two-tailed)
Word Count	2511.5	1996.4	4767.5	
Big words (greater than 6 letters)	21.2a	23.6b	21.1a	.05
Pronouns	9.15ab	9.83b	8.16a	.09
I (e.g. I, me, my)	0.61	0.90	0.83	
We (e.g. we, our, us)	1.94	1.79	1.95	
You (e.g. you, your, yours)	1.73	1.69	0.87	
He/she (e.g. he, hers, they)	1.42	1.42	1.37	
They (e.g., they, them)	2.17a	2.29a	1.43b	.03
Prepositions	14.8	14.7	15.0	
Articles (e.g. a, an, the)	9.07	8.53	9.19	
Exclusive Words (but, exclude)	2.72	2.62	3.17	
Affect	5.13a	5.12a	3.91b	.01
Positive emotion (happy, joy, love)	2.57a	2.83a	2.03b	.01
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.87b	.03
Anger words (hate, kill)	1.49a	1.32a	0.89b	.01
Cognitive Mechanisms	4.43	4.56	4.86	
Time (clock, hour)	2.40b	1.89a	2.69b	.01
Past tense verbs	2.21a	1.63a	2.94b	.01
Social Processes	11.4a	10.7ab	9.29b	.04
Humans (e.g. child, people, selves)	0.95ab	0.52a	1.12b	.05
Family (mother, father)	0.46ab	0.52a	0.25b	.08
Content				
Death (e.g. dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g. buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g. faith, Jew, sacred)	2.41	1.84	1.89	

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

Example: Laver and Garry (2000)

- A *hierarchical* set of categories to distinguish policy domains and policy positions – similar in spirit to the CMP
- Five domains at the top level of hierarchy
  - economy
  - political system
  - social system
  - external relations
  - a " 'general' domain that has to do with the cut and thurst of specific party competition as well as uncodable pap and waffle"
- Looked for word occurences within "word strings with an average length of ten words"
- Built the dictionary on a set of specific UK manifestos

Example: Laver and Garry (2000): Economy

Table 1 Abridged Section of Revised Manifesto Coding Scheme

1 ECONOMY
Role of state in economy

1 1 ECONOMY/+State+
Increase role of state

1 1 1 ECONOMY/+State+/Budget
Budget

1 1 1 1 ECONOMY/+State+/Budget/Spending
Increase public spending

1 1 1 1 1 ECONOMY/+State+/Budget/Spending/Health

1 1 1 1 2 ECONOMY/+State+/Budget/Spending/Educ. and training

1 1 1 1 3 ECONOMY/+State+/Budget/Spending/Housing

1 1 1 1 4 ECONOMY/+State+/Budget/Spending/Transport

1 1 1 1 5 ECONOMY/+State+/Budget/Spending/Infrastructure

1 1 1 1 6 ECONOMY/+State+/Budget/Spending/Welfare

1 1 1 1 7 ECONOMY/+State+/Budget/Spending/Police

1 1 1 1 8 ECONOMY/+State+/Budget/Spending/Defense

1 1 1 1 9 ECONOMY/+State+/Budget/Spending/Culture

1 1 1 2 ECONOMY/+State+/Budget/Taxes
Increase taxes

1 1 1 2 1 ECONOMY/+State+/Budget/Taxes/Income

1 1 1 2 2 ECONOMY/+State+/Budget/Taxes/Payroll

1 1 1 2 3 ECONOMY/+State+/Budget/Taxes/Company

1 1 1 2 4 ECONOMY/+State+/Budget/Taxes/Sales

1 1 1 2 5 ECONOMY/+State+/Budget/Taxes/Capital

1 1 1 2 6 ECONOMY/+State+/Budget/Taxes/Capital gains

1 1 1 3 ECONOMY/+State+/Budget/Deficit
Increase budget deficit

1 1 1 3 1 ECONOMY/+State+/Budget/Deficit/Borrow

1 1 1 3 2 ECONOMY/+State+/Budget/Deficit/Inflation

Example: Laver and Garry (2000)

```
ECONOMY / +STATE
    accommodation
    age
    ambulance
    assist
    ...

ECONOMY / -STATE
    choice*
    compet*
    constrain*
    ...
```

How to build a dictionary

- The ideal content analysis dictionary associates all and only the relevant words to each category in a perfectly valid scheme
- Three key issues:

Validity	Is the dictionary's category scheme valid?
Sensitivity	Does this dictionary identify *all* my content?
Specificity	Does it identify *only* my content?

How to build a dictionary

Assume you want to construct an entry for the category 'Terrorism'
Imagine two different dictionary entries:

- One contains all the words in the language (D1)
- The other contains the word 'terrorist' (D2)

D1 is *highly sensitive*: no language about terrorism is ever missed,
but *highly unspecific*: terrorism language is swamped
D2 is *highly specific*: the word occurs in discussions of terrorism,
but *highly insensitive*: much terrorism language is ignored
Of course, useful dictionaries lie in the middle

How to build a dictionary

Different problems arise with more than one category, e.g.

- 'Agricultural policy' vs 'National security'

Even if the categories *themselves* are exclusive there is always a chance a *word* suitable for one slips into the other category,
Or there are words that are used to describe both topics, e.g.

- 'revolution', 'outbreak', 'quarantine'

That is a fact not easily dealt with by CCA. An explicitly statistical framework is needed.

As Measurement

Translation. For each word:

	$P(\theta = \text{`Pro-State'} \mid \text{W})$	$P(\theta = \text{`Anti-State'} \mid \text{W})$
age	1	0
benefit	1	0
...	...	...
assets	0	1
bid	0	1
...	...	...

Using a dictionary

For each word W_i in a document

- If W_i is in category j, increment C_j
- Compute category proportions:

$$\hat{\theta}_i = \frac{C_i}{\sum_j C_j}$$

- The vector of category proportions is the content

Using a dictionary

A wrinkle in the interpretation: No category $K + 1$ to catch boring words —

> θ_i *is the proportion of category* i*, relative to other categories*

There is a category $K + 1$ to catch boring words —

> θ_i *is the proportion of the document devoted to category* i

A Sketch of the Statistical Framework

Assume $P(W \mid \theta)$ is

	θ	
	agriculture	security
nuclear	0	0.8
tractor	0.3	0
revolution	0.7	0.2
	1	1

A Sketch of the Statistical Framework

Bayes Theorem:

$$P(\theta \mid W) = \frac{P(W \mid \theta)P(\theta)}{P(W)}$$

So if $P(\theta = \text{`agriculture}') = 0.5$ then

	θ		
	agriculture	security	
nuclear	0	1	1
tractor	1	0	1
revolution	0.78	0.22	1

Proportions

Compute category proportions (as before):

$$\hat{\theta}_i = \frac{C_i}{\sum_j C_j}$$

C_i is a sum of $P(\theta = i \mid W)$s which can now be fractional

- e.g. two tokens of 'revolution' adds 1.56 to agriculture and 0.44 to security

Training, validation, and test sets

We can steal some useful terminology from Machine Learning:

Training set	documents you use to build the dictionary
Validation set	documents you use to tell how well you're doing
Test set	documents you use to quantify external validity

This scheme is intended to avoid 'over-fitting' — building a dictionary that is highly specific to a set of documents

A problem if you only sampled the population of texts, or want to use the dictionary on new data

Connecting dictionary content to substantive scales

- We're usually interested in category proportions per unit (usually document), e.g.
- *How much* of this document is about a given topic/category/affect?
- What is the score of a particular text *relative to a control group*?
- What is the *difference* of aggregated categories when compared?
- How does the of categories *change across time*?

Inference about content

Statistically speaking, the three types of measures are

- a proportion
- a difference of proportions
- a ratio of proportions

Under certain sampling assumptions we can make inferences about a population

Inference about proportions

The large sample standard error for the proportion $\hat{\theta}$ is

$$\hat{\sigma} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}}$$

where N is the length of the text. Works better when

$$N\hat{\theta} \text{ and } N(1-\hat{\theta}) > 10$$

Approximate 95% confidence interval is

$$\hat{\theta} \pm 1.96\hat{\sigma}$$

Inference about proportions

Example: in the 2001 Labour manifesto there are 879 matches to Laver and Garry's +state category

- 0.029 (nearly 3%) of the document's words
- 0.093 (about 9%) of words that matched *any* categories

The document has 30825 words, so the *first* proportion is estimated as

$$\hat{\theta}_{+\text{state}} = 0.029 \quad [0.027, 0.031]$$

What does this mean?

Inference about proportions

- Think of the party headquarters repeatedly *drafting* this manifesto
- The true proportion – the one suitable to the party's policies – is fixed but every draft is slightly different
- The confidence interval reflects the fact that we expect long manifestos to have more precise information about policy
- This interval is computed as if every word was a new (conditionally) independent piece of of information
  - That is probably not true, so it is probably *over*confident
- This is a quite general problem. . .

Reporting

Don't report proportions if you don't need to.

Rates are more intuitive

The rate of dictionary matches per B words is

$$\lambda_B = \theta B$$

which is a more interpretable proportion.

Different measures correspond to different choices of B.

Reporting

Not all choices are constant or comparable across languages, documents and topics

Quantity	B	Constant?
Proportion	1	Yes
Word count	N	No
Block	B	Yes
Sentence	?	No
Paragraph	?	No

Under what circumstances are these measures comparable?

Inference about differences

The large sample standard error for $\hat{\theta}_i - \hat{\theta}_j$ is

$$\hat{\sigma} = \sqrt{\frac{\hat{\theta}_i(1-\hat{\theta}_i)}{N} + \frac{\hat{\theta}_j(1-\hat{\theta}_j)}{N}}$$

where N is the length of the text. Works better when

$$N\hat{\theta} \text{ and } N(1-\hat{\theta}) > 10$$

Approximate 95% confidence interval is

$$\hat{\theta}_i - \hat{\theta}_j \pm 1.96\hat{\sigma}$$

Inference about differences

UK Conservatives tend to target rural voters.

How much more attention did they get from the Conservatives than from Labour in 2001?

Consider the (very small) category 'rural'

Conservatives match 29 words, Labour 31, but Labour's manifesto is much longer so

$$\hat{\theta}^{\text{LAB}} - \hat{\theta}^{\text{CON}} = -0.0012 \;\; [-0.0003, -0.002]$$

This difference is significant (though see caveats above).

Inference about ratios

Was the Conservative party in 1992 more or less for state intervention than New Labour in 1997?

Compare instances of +state and -state in the manifestos

Party	Counts		Proportion	
	+S	-S	+S	-S
Conservative	386	880	.013	.03
Labour	439	390	.025	.022

Risk Ratios

Compute two *risk ratios*:

$$RR_{+\text{state}} = \frac{P(+\text{state} \mid \text{cons})}{P(+\text{state} \mid \text{lab})}$$

$$RR_{\text{-state}} = \frac{P(\text{-state} \mid \text{cons})}{P(\text{-state} \mid \text{lab})}$$

and 95% confidence intervals

Risk Ratios

Standard error around estimated log RR is

$$\hat{\sigma} = \sqrt{\frac{1}{C_{\text{cons}}} - \frac{1}{N_{\text{cons}}} + \frac{1}{C_{\text{lab}}} - \frac{1}{N_{\text{lab}}}}$$

95% Confidence interval around log RR is

$$\log RR \ \pm \ 1.96\hat{\sigma}$$

Exponentiate the estimate and endpoints to get an interval for the risk ratio

Intepreting Risk Ratios

If $RR = 1$ then the category occurs at the same rate in labour and conservative manifestos

If $RR = 2$ then the conservative manifesto contains *twice* as much +state language as the labour manifesto

If $RR = .5$ then the conservative manifesto contains *half* as much +state language as the labour manifesto

If the confidence interval for RR contains 1 then we *no evidence* that +state and -state occur at different rates

Risk Ratios

	Risk Ratio
-S	1.35 [1.2, 1.5]
+S	0.53 [0.46,0.6]

Conservative manifesto generates 35% more -state words

- 35% = 100(1.35 - 1)%

Labour manifesto generates 89% more +state words

- 0.53 means *fewer* so
- 89% = 100(1/0.53 - 1)% more

Confidence interval suggests the increase is not less than 66% or more than 117%

More complex models

- More complex models are possible, when word rate occurrence is modeled more directly
- Example: Word rate occurrence could be Poisson distributed, and the dictionary approach simply selects specific words by pre-identified features
- From the quantitative matrix of (for instance) dictionary word occurrences by document, it would be possible to apply more advanced scaling or measurement methods
- But our next generalization will not involve modelling word rates by focusing on their stochastic process, but rather focusing on a relative probability model of word occurrence given a specific orientation