More than Unigrams Can Say: Detecting Meaningful Multi-word Expressions in Political Text

Kenneth Benoit

November 28, 2019

Outline

1. My background and perspective on this problem
2. Characterizing the problem
3. What are “meaningful multi-word expressions”
4. Detecting MWEs
5. Using MWEs to improve bag-of-words
6. Practical delivery of the solution

ME

Kenneth Benoit

- PhD in political science, specialization in statistics
- Department of Methodology
- “Computational social science”
  - research and PhD supervision in applications in data science to the social world
  - teach "Data for Data Scientists", "Quantitative Text Analysis", "Computer Programming", "Introduction to Machine Learning", among others
- R package author (**quanteda** and related packages)

THE PROBLEM

The problem: lots of MWEs in domain-specific text

Phrase	German equivalent	Left prefers to
Income tax	Einkommensteuer	Raise
Payroll tax	Lohnsteuer	Raise
Sales tax	Umsatzsteuer	Lower
Value added tax	Mehrwertsteuer	Lower
Flat tax	Abgeltungssteuer	Abolish
Carbon tax	Kohlenstoffsteuer	Raise
Inheritance tax	Erbschaftssteuer	Raise
Capital gains tax	Wertzuwachssteuer	Raise
Corporate tax	Körperschaftssteuer	Raise
Property tax	Vermögenssteuer	Raise
Real estate transfer tax	Grunderwerbsteuer	Raise
Motor vehicle tax	Kraftfahrzeugsteuer	Not mention
Employer's National Insurance Contribution	Sozialversicherungsbeiträge	Raise

Table 1: *Tax-related multi-word expressions in English and German.*

Domain-specific terminology is rife with MWEs - up to 40%

a worst case

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

a worst case

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

meaning: "the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef"

even worse?

Austrittsvertragsratifizierungsgesetzentwurf

even worse?

Austrittsvertragsratifizierungsgesetzentwurf

meaning: “withdrawal agreement bill”

Especially true in politics (and economics)

	Robertson			Safire		
	N	%	Examples	*N*	%	Examples
Unigrams	300	54%	Watergate	645	33%	bork
Bigrams	199	36%		806	42%	
A-N	116		agrarian parties	338		Young Turks
N-N	69		cabinet government	314		gunboat diplomacy
Other	14		politically correct	154	*	bridge building
Trigrams	38	7%		236	12%	
A-A-N	3		single transferable vote	8		redheaded Eskimo bill
A-N-N	6		additional member system	10		yellow dog democrat
N-A-N	0		--	1		*illegitimi non carborundum*
N-N-N	2		war crimes tribunals	6		Rose Garden rubbish
N-P-N	13		equality of opportunity	65		milk for Hottentots
Other	11		*raison de guerre*	13		buck stops here
> 3-grams	16	3%	vanguard of the proletariat	247	13%	chicken in every pot
Total entries	553	100%		1934	100%	

Sources: Robertson, David. 2004. The Routledge dictionary of politics. Routledge;
Safire, William. 2008. Safires political dictionary. Oxford University Press.

Problem: BOW is wrong

- violates conditional independence assumption
  - probability of observing one word significantly increases the probability of observing a second
  - causes underestimation of uncertainty
- conflates different feature associations
  - `national`, `insurance`, `security`, `socialist` or `national_insurance`, `national_security`, `National_Socialist` ?
  - double weighting affects averaging-based models for two-word terms, such as `European Union`

How to solve this?

1. NOT: "simply include all ngrams"

How to solve this?

1. NOT: "simply include all ngrams"
2. Determine a functional method to detect MWEs in political corpora
   - through association measures
   - through filtering: stopwords, parts-of-speech (POS)
   - predictive methods, against human annotation of two baseline corpora

How to solve this?

1. NOT: "simply include all ngrams"
2. Determine a functional method to detect MWEs in political corpora
   - through association measures
   - through filtering: stopwords, parts-of-speech (POS)
   - predictive methods, against human annotation of two baseline corpora
3. Apply the method to a massive set of political text, to develop a (comprehensive) standard list

How to solve this?

1. NOT: "simply include all ngrams"
2. Determine a functional method to detect MWEs in political corpora
   - through association measures
   - through filtering: stopwords, parts-of-speech (POS)
   - predictive methods, against human annotation of two baseline corpora
3. Apply the method to a massive set of political text, to develop a (comprehensive) standard list
4. Use the MWEs instead of unigram tokenization in applications

How to solve this?

1. NOT: "simply include all ngrams"
2. Determine a functional method to detect MWEs in political corpora
   - through association measures
   - through filtering: stopwords, parts-of-speech (POS)
   - predictive methods, against human annotation of two baseline corpora
3. Apply the method to a massive set of political text, to develop a (comprehensive) standard list
4. Use the MWEs instead of unigram tokenization in applications
5. Show it makes a difference.

WHAT ARE (MEANINGFUL) MWEs?

Defining a "collocation"

There are both linguistic and statistical criteria.

- Linguistic: MWE is a meaningful sequence of words that can have a meaning as a unit, rather than a string of individual words
- Statistical: a series of tokens whose collocated occurrence is not by chance

Here, however, we focus on *statistical* criteria for MWE candidate detection, and linguistic criteria for filtering meaningful MWEs being MWE

- In essence, based on co-occurrence of words: a sequence of K successive words is a candidate for MWE if occurs sufficiently often in the corpus
- Not sufficient, but necessary for an expression being MWE in the linguistic sense

Taxonomy of MWEs (Sag et al 2002)

Category	Subcategory	Examples
Fixed expressions	Proper names	Labour Party, New York City
	Foreign terms	*coup d'état*, *habeas corpus*
	Fixed phrases	banana republic, off the record
Semi-fixed expressions	Idioms	gunboat diplomacy, fat cat, pork barrel
	Compound nominals	attorney general, Member of Parliament
Institutionalized phrases		child benefit, alternative minimum tax

Table 2: *Examples of political MWEs according to Sag et al. (2002)'s typology.*

Define: “meaningful”

- **fixedness of a phrase**: *hung parliament* qualifies because we do not say “a parliament that is hung”
- **orthographic lexicalisation**: some words have taken the “German route”, e.g. “dataset” indicates that *data set* is a MWE
- **non-compositionality**: when you cannot detect a phrases meaning from a simple combination of the meaning of its component words, e.g. *hanging chad*, *first lady*
- **proper nouns**: almost always indicate MWEs, such as *Native American* or *Supreme Court*

Statistical definition of a "collocation"

For a given value of K, turn the corpus into a dataset of observed K-word sequences.

1. For each candidate expression in turn (e.g. every K-word sequence which appears in the corpus), calculate the value of some statistic θ defined in such a way that higher values of θ are regarded as stronger evidence that the expression is MWE
2. Order candidate expressions by their values of θ
3. Make decisions about which expressions will be treated as MWEs, e.g. all above some cut-off for θ or (more likely) human review and decision-making
4. Treat selected expressions as single words in subsequent text analysis

Statistical definition of a "collocation" (cont)

For expressions of different lengths, start with some maximum value $K = K_{max}$ and proceed toward smaller K. In other words, a K-word expression declared to be MWE is treated as a single word when we examine $(K-1)$-word expressions, and thus in effect removed from consideration.

How to choose θ

The main focus of the paper, however, is on choosing the statistic θ.

- Many possibilities have been proposed in the literature, but not always considered systematically, from statistical first principles
- we argue that this is best done drawing on some general ideas from models for categorical data
- a statistical *definition* of an MWE can be given in terms of a single quantity, the highest-order interaction parameter in a saturated loglinear model for a K-way contingency table defined by the appearances of the candidate expression and its sub-expressions in the corpus
- This parameter (λ) can itself be used as a statistic θ

DETECTING MWEs

Contingency tables for bigrams

In very basic terms, for bigrams only: tabulate every token against every other token as pairs, and compute for each pair:

	token2	¬token2	Totals
token1	n_{11}	n_{12}	n_{1p}
¬token1	n_{21}	n_{22}	n_{1p}
Totals	n_{p1}	n_{p2}	n_{pp}

(Previous) statistical association measures

where m_{ij} represents the cell frequency expected according to independence:

G^2 likelihood ratio statistic (Dunning 1993), computed as:

$$2 * \sum_i \sum_j (n_{ij} * log \frac{n_{ij}}{m_{ij}}) \tag{1}$$

χ^2 Pearson's χ^2 statistic, computed as:

$$\sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \tag{2}$$

Statistical association measures (cont.)

pmi point-wise mutual information score, computed as $\log n_{11}/m_{11}$

dice the Dice coefficient, computed as

$$\frac{n_{11}}{n_{1.} + n_{.1}} \tag{3}$$

POS filtering

- With the exception of some middle-word prepositions, we removed all MWEs containing stopwords (about 80% in our applicaitons)

POS filtering

- With the exception of some middle-word prepositions, we removed all MWEs containing stopwords (about 80% in our applicaitons)
- Justeson and Katz (1995) found that the following parts of speech contained relevant MWEs:
  - bigram MWEs: NOUN-NOUN and ADJECTIVE-NOUN
  - trigram MWEs: N-N-N, ADJ-ADJ-N, ADJ-N-N, N-ADJ-N, and N-PREP-N
  - we also included all exclusively NP (proper noun) MWEs, like *Scottish National Party*
- Note that advanced taggers can also identify named entities and noun phrases (e.g. **spacy**)

Our implementation

`quanteda::textstat_collocations()`

- sliding window of size n is used to scan the token sequences. These are tabulated (parallelized), and 0.5 added to counts as continuity correction factor
- uses a bitwise encoding method:
  For an n-gram $X_1, X_2, ..., X_n$, if $n = 3$, we use $m_{j_1 \dots j_K}, K = 3$ to denote the count of the trigram
  $X_1 = x_1 \wedge X_2 = x_2 \wedge X_3 = x_3$.
  $j_i = 1$ if $X_i = x_i$, otherwise $j_i = 0$
- Example:
  - m_{111} count $X_1 = \textit{United} \wedge X_2 = \textit{State} \wedge X_3 = \textit{Congress}$
  - m_{010} counts $X_1 \neq \textit{United} \wedge X_2 = \textit{State} \wedge X_3 \neq \textit{Congress}$

Our implementation (cont.)

So λ can be expressed as:

$$\lambda = \sum_{i=1}^{K} (-1)^{K - b_{j_1 \ldots j_K}} * \log m_{j_1 \ldots j_K} \tag{4}$$

Details: $K = 2$

Suppose we examine a corpus of text which has been turned into a dataset of observed K-word sequences $\mathbf{z}_1, \ldots, \mathbf{z}_{N*}$.

Our target expression is $\mathbf{x} = (x_1, x_2)$, and the comparisons between $\mathbf{x}$ and the sequences $\mathbf{z}_j$ observed in the corpus are summarised in a 2×2 contingency table.

Denote the dimensions of the table so that the probabilities p_i are written as $p_{c_1 c_2}$ for $c_1, c_2 = 0, 1$.

These are the probabilities that neither word of a $\mathbf{z}_j$ matches the corresponding word of $\mathbf{x} = (x_1, x_2)$ (probability p_{00}), the first word matches but the second does not (p_{10}), the second word matches but the first does not (p_{01}), and that an observed expression matches the target exactly (p_{11}).

Details: $K = 2$ (cont.)

The log-linear formulation can be written as

$$\log p_{c_1 c_2} = \lambda_0 + \lambda_1 I(c_1 = 1) + \lambda_2 I(c_2 = 1) + \lambda\, I(c_1 c_2 = 1) \quad (5)$$

where $\lambda = \log[(p_{00} p_{11})/(p_{01} p_{10})]$ is the log odds ratio (log-OR) which desctribes the association between the two dimensions of the table.

$\lambda = 0$ if the words x_1 and x_2 occur independently in the corpus as first and second words of two-word sequences

By contrast, $\lambda > 0$ if the words x_1 and x_2 occur together (and in this order) more often than would be expected.

POS filtering and expectations of meaningful MWEs

- With the exception of some middle-word prepositions, we removed all MWEs containing stopwords (about 80% in our applicaitons)

POS filtering and expectations of meaningful MWEs

- With the exception of some middle-word prepositions, we removed all MWEs containing stopwords (about 80% in our applicaitons)
- Justeson and Katz (1995) found that the following parts of speech contained relevant MWEs:
  - bigram MWEs: NOUN-NOUN and ADJECTIVE-NOUN
  - trigram MWEs: N-N-N, ADJ-ADJ-N, ADJ-N-N, N-ADJ-N, and N-PREP-N
  - we also included all exclusively NP (proper noun) MWEs, like *Scottish National Party*

POS filtering and expectations of meaningful MWEs

- With the exception of some middle-word prepositions, we removed all MWEs containing stopwords (about 80% in our applicaitons)
- Justeson and Katz (1995) found that the following parts of speech contained relevant MWEs:
  - bigram MWEs: NOUN-NOUN and ADJECTIVE-NOUN
  - trigram MWEs: N-N-N, ADJ-ADJ-N, ADJ-N-N, N-ADJ-N, and N-PREP-N
  - we also included all exclusively NP (proper noun) MWEs, like *Scottish National Party*
- we tagged the text prior to tokenization, so that the tagger could use context

POS filtering and expectations of meaningful MWEs

- With the exception of some middle-word prepositions, we removed all MWEs containing stopwords (about 80% in our applicaitons)
- Justeson and Katz (1995) found that the following parts of speech contained relevant MWEs:
  - bigram MWEs: NOUN-NOUN and ADJECTIVE-NOUN
  - trigram MWEs: N-N-N, ADJ-ADJ-N, ADJ-N-N, N-ADJ-N, and N-PREP-N
  - we also included all exclusively NP (proper noun) MWEs, like *Scottish National Party*
- we tagged the text prior to tokenization, so that the tagger could use context
- note: the tagger is often wrong

```
library("quanteda")
data(data\_corpus\_sotu, package = "quanteda.corpora")

toks <- tokens(data\_corpus\_sotu) %>%
    tokens\_remove("\\p{P}", padding = TRUE, valuetype = "regex") %>%
    tokens\_remove(stopwords("en"), padding = TRUE)

colls <- textstat\_collocations(toks, size = 2)
head(colls, 10)
collocation count count_nested length   lambda         z
1       united states  4811           0      2 9.533739 161.26344
2           last year   575           0      2 4.833398  98.77367
3        last session   427           0      2 6.629301  95.14509
4         fiscal year   840           0      2 7.861374  95.00841
5  federal government   477           0      2 4.636497  85.58259
6     american people   438           0      2 4.615388  84.95583
7             june 30   324           0      2 9.544416  84.09833
8         health care   237           0      2 7.230485  83.40335
9     social security   226           0      2 7.264191  79.87448
10     annual message   200           0      2 7.915638  79.02214
```

```
library("spacyr")

toks2 <- spacy_parse(data_corpus_sotu) %>%
    as.tokens(include_pos = "pos") %>%
    tokens_select("/(NOUN|ADJ)$", valuetype = "regex", padding = TRUE)

colls2 <- textstat_collocations(toks2, size = 2)
head(colls2, 15)
collocation count count_nested length   lambda        z
1              last/adj year/noun   606           0      2 5.065243 103.7828
2           last/adj session/noun   425           0      2 6.850330  96.5312
3            FISCAL/adj YEAR/noun   828           0      2 7.835043  94.4376
4        american/adj people/noun   437           0      2 4.749478  86.5269
5           HEALTH/noun CARE/noun   238           0      2 7.516710  84.2561
6            PUBLIC/adj DEBT/noun   284           0      2 6.084872  79.6998
7         ANNUAL/adj MESSAGE/noun   199           0      2 7.985613  79.1101
8              past/adj year/noun   316           0      2 5.716268  78.4098
9           PUBLIC/adj LANDS/noun   235           0      2 5.912245  72.6576
10       fellow/adj citizens/noun   159           0      2 7.157765  62.4847
11            last/adj annual/adj   158           0      2 5.842831  61.0288
12     LOCAL/adj GOVERNMENTS/noun   123           0      2 6.314859  60.2746
13          INDIAN/adj TRIBES/noun    93           0      2 7.949873  58.7688
14 favorable/adj consideration/noun 106           0      2 6.914765  57.2924
15        ECONOMIC/adj GROWTH/noun   114           0      2 6.157860  57.0053
```

Next steps

- Massive mining of political corpora
- Human verification of scored and filtered MWEs
- Payoff: domain-specific MWE "dictionaries" for pre-processing texts; OR
- Verified method for detecting MWEs for specific (new) domains

Initial corpora we've mined

Corpus	Description	Documents	Total words
US Presidential	Inaugural addresses 1789-2013; State of the Union addresses since 1985-2015	88	314,031
UK Manifestos	UK Manifestos 1945-2010	115	1,296,228
Irish Manifestos	Irish Manifestos 1992-2004	30	384,757
US Manifestos	US Party Platforms 1844-2004	88	743,718
UK Parliament	Hansard, from Eggers and Spirling (2014)	1,264,675	282,513,998
Irish Parliament	Full text 1919-2013, from Herzog and Mikhaylov (2013)	4,443,714	484,101,243
Amicus briefs	*Grutter/Gratz v. Bollinger*, from Evans et al. (2007)	102	602,469
Supreme Court Briefs	All briefs 1948–2012; from Sim, Routledge and Smith (2015)	40,672	396,744,956
Supreme Court opinions	Opinions 1948–2012 (Sim, Routledge and Smith, 2015)	8,486	65,248,384
Total		5,757,970	1,231,949,784

Table 5: *Description of corpora analyzed for collocations.*

POS and stopword filtering on US presidential corpus

POS Pattern	Examples
US Presidential Speeches	
A-N	middle class, economic growth, nuclear weapon(s), national security, natural gas, private sector, public transport, human rights
NP-NP	United States, Federal Government, Vice President, Al Qaida, Middle East
N-N	health care, health insurance, tax credit, child care, climate change, minimum wage, trade union(s), arms control
Other	chief executive (A-A), clean energy (V-N), equal rights (V-N)*
A-N-N	private health insurance, free trade agreement, political action committee(s)
N-P-N	Members of Congress, war on terror, rule of law, violence against women
A-A-N	gross national product, Native American reservations, alternative minimum tax, rural electric cooperatives, strategic nuclear weapons
N-N-N	health care system, social security benefits, capital gains tax, third world countries
NP-NP-NP	United States Congress, Strategic Defense Initiative, New York City
N-A-N	--
Other	research and development, step by step (V-P-N), weapons of mass (N-P-A), office of the

USING MWEs

PRACTICAL DELIVERY:
MWEs for the masses

Deliverable: Domain-specific dictionaries

From mining, filtering, and verifying numerous domain-specific corpora, not just politics.

- Examples: Legal, business, economic, finance, medicine
- Generally no penalties for being inclusive: "stare decisis" will not occur in non-legal texts, for instance, and therefore will not adversely affect results.

Deliverable: Domain-specific dictionaries

From mining, filtering, and verifying numerous domain-specific corpora, not just politics.

- Examples: Legal, business, economic, finance, medicine
- Generally no penalties for being inclusive: "stare decisis" will not occur in non-legal texts, for instance, and therefore will not adversely affect results.
  Very rarely do "false positive" collocations occur, such as:
  - The *first lady*, was happy over the successful Mars landing.
  - She was the *first lady* to make a successful Mars landing.
- And any "damage" from false positives likely to be less than the damage from ignoring MWEs

Tools (implementing the method)

R package **quanteda**:

- `textstat_collocations()`
- `textstat_compound()`
- dictionary and "lookup" methods optimized for MWEs
- all parallelized (in C++)
- integration with NLP tools such as **spacy**