Taming the Firehose: Thematically summarizing very large news corpora using topic modeling

Philip A. Schrodt

Parus Analytics
Charlottesville, Virginia, USA
schrodt735@gmail.com
philipschrodt.org

Presentation at PolMeth XXXV: 2018 Conference of the
Society for Political Methodology
Brigham Young University
18 July 2018

Trigger warning!!!

This is an application-focused paper!

Papers with a focus on application: these are papers that do not develop new methodology, and instead employ existing methods creatively to answer substantive questions

`https://www.cambridge.org/core/membership/spm/conferences/polmeth-2018`

Plus the little matter of. . .

open access resources vs. paywalled journals

The problem: Drinking from a firehose

A core tool of international political analysis is a chronology of who did what to whom. Historically these were constructed by subject matter experts reading available material, picking out the major themes of the interactions.

Contemporary analysts, however, are faced with two problems:

- The set of potentially important state and non-state actors being monitored is far larger and diverse than in the past, particularly compared to the Cold War when most US analytical efforts were focused on a single highly centralized and bureaucratized actor, the Soviet Union. Then considered an adversary: how quaint.
- The information available on these actors is now vastly greater, and cannot be read or even summarized by analysts. But it's mostly machine-readable.

The proposed solution: Automated chronology generators

Proposals in “artificial intelligence” go back to at least the 1980s, typically conceptualized as a [proprietary, thoroughly black-boxed, and mind-boggingly expensive] “analyst’s workstation.”

The dream has continued to this day: for example this was one of the original foci of the 2008-2011 DARPA Integrated Conflict Early Warning System (ICEWS) project.

Yes, to this day: two weeks ago IARPA issued a BAA for some components of such a system: if you are interested in teaming on this, see me. Particularly if you can prime.

Approaches that don't work very well

Keyword searching:

- too many false positives
- difficulties in accommodating stylistic differences (particularly synonyms) in heterogeneous corpora

Example-based document similarity:

- generating examples is time consuming
- texts describing interactions are frequently sentence-length rather than document-length
- only works if you know what you are looking for

Event data:

- requires *a priori* coding ontologies and dictionaries
- many analysts and policy-makers remain uncomfortable with statistical summaries and need to look at the texts

Topic modeling solves most of these problems

There has been a dramatic increase in the use of topic modeling approaches in the past two decades, particularly following the development of latent Dirichlet allocation methods (LDA;Blei et al). But some issues remain:

- These use only "bag of words" vectors so the absence of semantic and grammatical information means topics may not focus on interactions
- Because of its dependence on numerical optimization, LDA does not generate a unique set of topics in most applications using large corpora
- LDA tends to generate some nonsense topics that analysts may or may not tolerate

Approach used in this system

- Pre-filter for interactions using an event data coder
- Do an assortment of pre-processing
- Estimate multiple LDA models using the open-source `gensim` package (Python)
- Aggregate similar topics within and between models based on correlations between their keywords and classified sentences
- Generate summaries and chronologies by theme

Text corpus

- Roughly 2-million sentences on Middle East politics for 2017.
- All text is in English but about 70% of these are machine-translated from Arabic: these two sets are stylistically *very* different
- Application 1: Analyze events relevant to a single state—Qatar and Yemen in this example—for a single month (October 2017)
- Application 2: Analyze the entire corpus

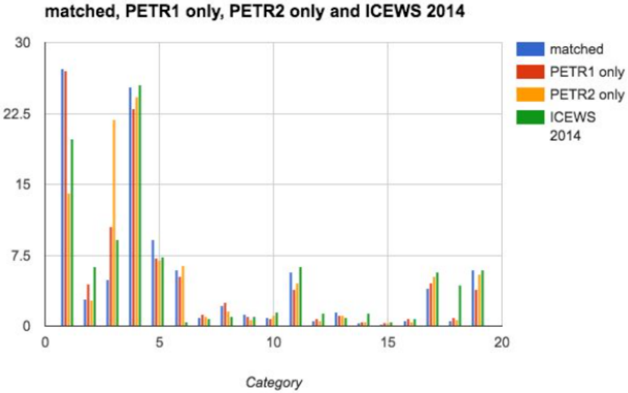
Pre-filtering

Pre-filtering is done with a proprietary political event coder descended from the coder used to generate the ICEWS data

This reduces the corpus to about 750,000 sentences

For the nation-month cases, sentences are selected if they contain the name of the country anywhere in the text

Open source coders are likely to produce similar results



Preprocessing

- Remove a relatively small number of stopwords—see Spirling et al for illustrations on why this may be consequential—plus the country names in the nation-month cases
- Standardize multi-word entities such as `United States`, `Saudi Arabia` and `United Nations`, resolve demonyms (`American`, `Yemeni`, `Qatari`), and deal with other common idioms such as the use of a capital city (`Washington`, `Riyadh`) to refer to a government.
- Remove numbers and punctuation
- For speed and memory considerations, generate dictionaries using first 2048 records after removing low-frequency words

Except for the pre-processing, the analysis is unsupervised and does not involve `a priori` dictionaries or ontologies

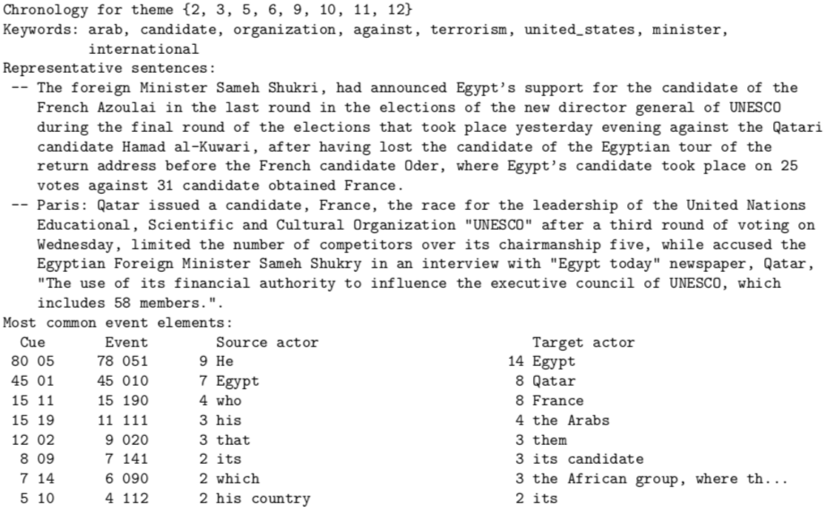
Processing

Realistically, just get the code from me, though the paper provides some detail. But briefly

- Estimate multiple models using `gensim.models.LdaModel()`. Assign texts to themes using the `gensim` similarity metrics
- Aggregate similar themes both within a single model and across multiple estimated model
- Use the `gensim.summarize()` function to try to get a sense of the thematic content of those clusters

This runs relatively quickly—ones to tens of minutes—using modest computing resources (individual cloud computing instances, a.k.a. cheap desktops, not supercomputers) though not sufficiently fast for real-time interaction via a dashboard

Output



```
Chronology for theme {2, 3, 5, 6, 9, 10, 11, 12}
Keywords: arab, candidate, organization, against, terrorism, united_states, minister,
          international
Representative sentences:
 -- The foreign Minister Sameh Shukri, had announced Egypt's support for the candidate of the
    French Azoulai in the last round in the elections of the new director general of UNESCO
    during the final round of the elections that took place yesterday evening against the Qatari
    candidate Hamad al-Kuwari, after having lost the candidate of the Egyptian tour of the
    return address before the French candidate Oder, where Egypt's candidate took place on 25
    votes against 31 candidate obtained France.
 -- Paris: Qatar issued a candidate, France, the race for the leadership of the United Nations
    Educational, Scientific and Cultural Organization "UNESCO" after a third round of voting on
    Wednesday, limited the number of competitors over its chairmanship five, while accused the
    Egyptian Foreign Minister Sameh Shukry in an interview with "Egypt today" newspaper, Qatar,
    "The use of its financial authority to influence the executive council of UNESCO, which
    includes 58 members.".
Most common event elements:
  Cue       Event       Source actor                           Target actor
 80 05     78 051      9 He                                 14 Egypt
 45 01     45 010      7 Egypt                               8 Qatar
 15 11     15 190      4 who                                 8 France
 15 19     11 111      3 his                                 4 the Arabs
 12 02      9 020      3 that                                3 them
  8 09      7 141      2 its                                 3 its candidate
  7 14      6 090      2 which                               3 the African group, where th...
  5 10      4 112      2 his country                         2 its
```

What works

- Both expected (violence in Syria, Yemen) and unexpected (Arabic press controversy over UNESCO election in October-2017) credible themes emerge
- The relative importance of themes can be assessed by the number of times it is found in the multiple estimates
- Assignment of sentences to themes is plausible most of the time, though definitely not all of the time
- Pre-filtering for interactions yields texts that do, in fact, look like chronologies
- In the country-month case, many themes focus on interactions with specific states, which is what one expects in human-generated themes

What doesn't work so well

- Except for a few very conspicuous themes—violence in Syria, Arab reaction to US embassy in West Jerusalem—most of the themes in the general case are vague: stopword list may be too limited
- `gensim.summarize()` function fails in a surprising number of cases, and a better algorithm may be needed here
- Classifying sentences outside of the nation-month estimated—useful for detecting precursors—hasn't worked well, though this may be due to implementation issues on my side; this also seems very sensitive to a hyperparameter in the `gensim` similarity function
- Stylistic differences between translated Arabic and native English—starting with the simple issue of sentence length—are almost certainly affecting results

Next steps

- Experimentation with hyperparameters to adjust precision/recall tradeoffs
- Differentiation—and ideally, visualization using, e.g. correspondence analysis or t-SNE—of texts which are central vs those peripheral to the thematic clusters
- Daily summarization in the general case, where the chronologies currently run to tens of megabytes
- Bayesian seeding of topics using examples and/or keywords and/or over-weighting terms such as country/leader names

Thank you

Email:
schrodt735@gmail.com

Slides:
`http://eventdata.parusanalytics.com/presentations.html`