A Word-Count Based Classifier of Politicians in the Australian Federal Parliament (1901–2018)

Rohan Alexander, Australian National University
Patrick Leslie, Australian National University

EPSA 2019
Belfast, Northern Ireland
22 June 2019

Overview

Data

Model

Results

Conclusion

Motivation

“If New South Wales is to join the Federation, she should demand Proportional Representation in the Senate as well as in the lower house, so as to avoid the mischief which must necessarily arise... from the inordinate and arrogant demands of the smaller states.”

The Cumberland Argus and Fruitgrowers' Advocate, 2 October 1897

“[I]n a few years time... [t]he fact will then be that the Australian people as a whole elect... 36 Senators, and for convenience of election, Australia is roughly divided into six divisions. Thus the people as a whole will elect their representatives, and the people as a whole will elect their Senators. If the system has a fault, it is that the Senate is but a duplicate of the House of Representatives.”

The Dungog Chronicle, 8 July 1898

Questions

- To what extent is the Senate still, to the extent that it ever was, a ‘states’ house’?
- Is the Senate comparatively dominated by the parties?
- How has the influence of states and parties changed over time?

The Senate chamber in Old Parliament House.
By Unknown - http://catalogue.nla.gov.au/Record/4199499/Copyright

Approach

- Codify two data sources to create a new dataset of who said what in the Australian Federal Parliament between 1901 and 2018.
- Analyse politician-specific word counts using elastic net regularisation.
- We then examine the accuracy of our model.

The Senate chamber in New Parliament House.
By JJ Harrison (https://www.jjharrison.com.au/).

Findings

- Our model is better able to classify both state and party later in our sample than earlier.
- It is better able to classify party in the Senate.
- It is better able to classify state in the House of Representatives.

Contributions

Data
We bring an essentially-complete new corpus of who said what in the Australian Federal Parliament on an individual basis.

Methods
Our approach of using word counts from parliamentary speeches combined with regularised regression is parsimonious and could be easily applied to answer a variety of other questions.

Australian political knowledge
We show: the extent to which the Senate has drifted from its ostensible reason for being to now just be a variant of the House of Representatives with a different electoral system; and the increasingly paramount importance of political parties.

Overview

Data

Model

Results

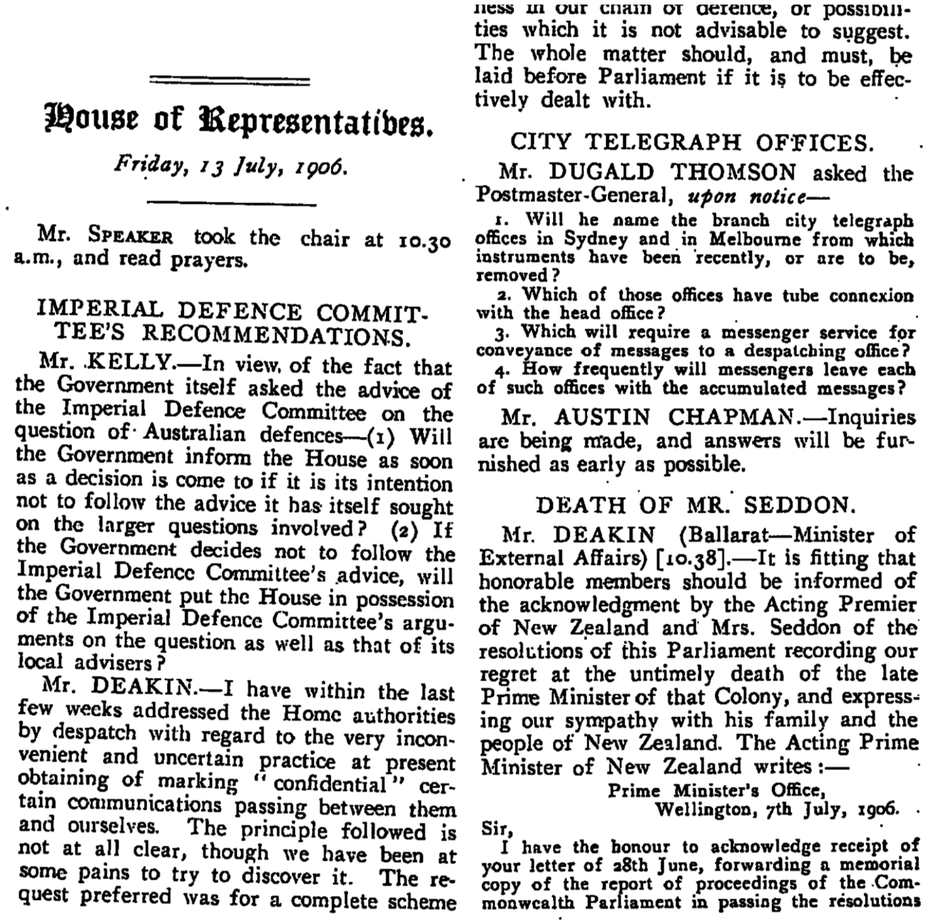
Conclusion

Data

Hansard PDFs are available since Federation (1901). XML available, but incomplete. No turnkey Hansard corpus for Australian researchers, yet.

Creating a corpus required a large PDF-parsing and data-cleaning exercise. We end up with 7,934 days in House of Representations and 6,746 days in Senate, across 118 years.

Our CSV corpus (c.4GB) is available for other researchers.



House of Representatives.

Friday, 13 July, 1906.

Mr. SPEAKER took the chair at 10.30 a.m., and read prayers.

IMPERIAL DEFENCE COMMITTEE'S RECOMMENDATIONS.

Mr. KELLY.—In view of the fact that the Government itself asked the advice of the Imperial Defence Committee on the question of Australian defences—(1) Will the Government inform the House as soon as a decision is come to if it is its intention not to follow the advice it has itself sought on the larger questions involved? (2) If the Government decides not to follow the Imperial Defence Committee's advice, will the Government put the House in possession of the Imperial Defence Committee's arguments on the question as well as that of its local advisers?

Mr. DEAKIN.—I have within the last few weeks addressed the Home authorities by despatch with regard to the very inconvenient and uncertain practice at present obtaining of marking "confidential" certain communications passing between them and ourselves. The principle followed is not at all clear, though we have been at some pains to try to discover it. The request preferred was for a complete scheme

...ness in our chain of defence, or possibilities which it is not advisable to suggest. The whole matter should, and must, be laid before Parliament if it is to be effectively dealt with.

CITY TELEGRAPH OFFICES.

Mr. DUGALD THOMSON asked the Postmaster-General, *upon notice*—

1. Will he name the branch city telegraph offices in Sydney and in Melbourne from which instruments have been recently, or are to be, removed?
2. Which of those offices have tube connexion with the head office?
3. Which will require a messenger service for conveyance of messages to a despatching office?
4. How frequently will messengers leave each of such offices with the accumulated messages?

Mr. AUSTIN CHAPMAN.—Inquiries are being made, and answers will be furnished as early as possible.

DEATH OF MR. SEDDON.

Mr. DEAKIN (Ballarat—Minister of External Affairs) [10.38].—It is fitting that honorable members should be informed of the acknowledgment by the Acting Premier of New Zealand and Mrs. Seddon of the resolutions of this Parliament recording our regret at the untimely death of the late Prime Minister of that Colony, and expressing our sympathy with his family and the people of New Zealand. The Acting Prime Minister of New Zealand writes:—

Prime Minister's Office,
Wellington, 7th July, 1906.

Sir,

I have the honour to acknowledge receipt of your letter of 28th June, forwarding a memorial copy of the report of proceedings of the Commonwealth Parliament in passing the resolutions

First page of Hansard for 13 July 1906 in the House of Representations.

Overview

Data

Model

Results

Conclusion

Model

Our model considers state as a multinomial outcome of word choice:

$$s_i = \beta_1 w_{i,1} + \beta_2 w_{i,2} + \ldots + \beta_n w_{i,n}$$

In this set-up, s_i is the state of a particular politician i, and $w_{i,1}$ is the count of the number of times politician i used the word or phrase w_1. Our party model is the same, expect that p_i, which is the party of a particular politician i replaces s_i. For the party model, there are two possible classifications and so the model is binomial. For the states model, there are six possible classifications and so the model is multinomial.

Overview

Data

Model

Results

Conclusion

Proportion of parties classified correctly

79 per cent of politicians are classified correctly by party.

Chamber	Party	Classified ALP	Classified LNP	Total
House	ALP	180	126	306
	LNP	25	311	336
Senate	ALP	96	30	126
	LNP	7	108	115

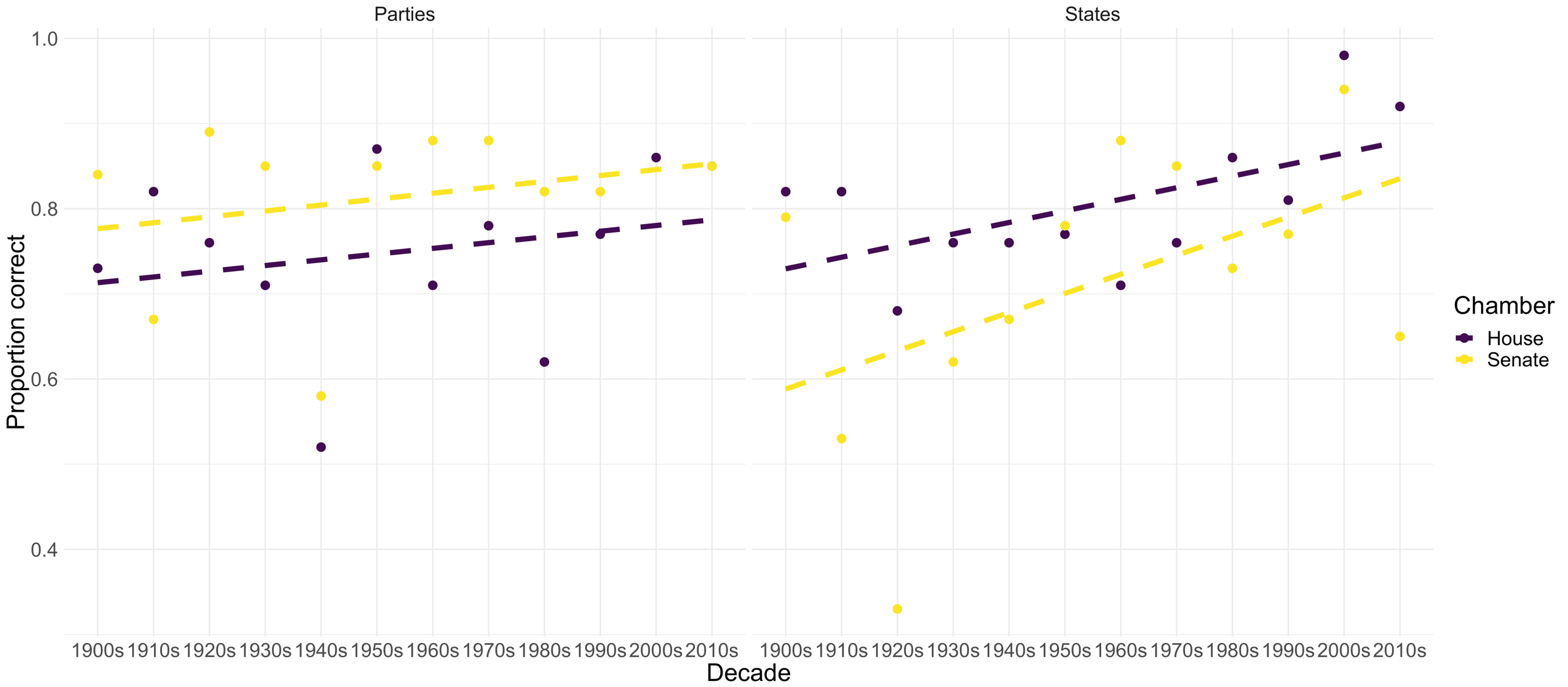
Proportion of states classified correctly

80 per cent of politicians are classified correctly by state.

Chamber	State*	Incorrect	Correct	Total	Prop. correct
House	NSW	38	181	219	0.83
	⋮	⋮	⋮	⋮	⋮
	WA	20	43	63	0.68
Senate	NSW	8	30	38	0.79
	⋮	⋮	⋮	⋮	⋮
	WA	14	26	40	0.65

*QLD, VIC, SA, and TAS removed for space reasons.

Proportion classified correctly over time



Estimated party classification probabilities

Estimated state classification probabilities

Overview

Data

Model

Results

Conclusion

Summary

1. Collect, parse, and clean Australian Hansard PDFs to construct a corpus with around 15,000 days spread across 118 years and 2,000 politicians.

2. Train a supervised machine learning model to forecast state and party

3. Evaluate how accurate that model is and how that changes over time.

4. Find that our model is better able to classify state and party later in our sample than earlier. It is better able to classify party in the Senate, and state in the House of Representatives.

Weaknesses

Data
Even after cleaning the dataset remains imperfect and is more fit-for-purpose than of broad applicability.

Parties
Concerns about the appropriateness of some of the coding of historical parties.

Model
Using a quantitative approach to analyse language is simplistic and does not account for many aspects of discourse. This is especially the case in our model which considers only word count. The model also performs poorly when there is only a small amount of text.

The Harold Holt Memorial Pool.
Photo by Kbpool2012, from Wikimedia Commons.

“A Word-Count Based Classifier of Politicians in the Australian Federal Parliament (1901–2018)”
Rohan Alexander and Patrick Leslie

Email: rohanalexander@anu.edu.au.

Twitter: @rohanalexander.

Paper is early in life-cycle and available on request and your comments are very welcome.

Data available for download and use, but maybe contact me if you need to know where the bodies are buried.

Acknowledgements: Thank you to Monica Alexander, Edward Howlett, Marija Taflaga, and John McAndrews for their help.

Slides theme based on Nathan Lane, see https://slides.com/nathanlane/kdi#/.