Research Article

Research and Politics
April-June 2021: 1–10
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20531680211022206
journals.sagepub.com/home/rap

Transfer learning for topic labeling: Analysis of the UK House of Commons speeches 1935–2014

Hannah Béchara[1], Alexander Herzog[2], Slava Jankin[1] and Peter John[3]

Abstract

Topic models are widely used in natural language processing, allowing researchers to estimate the underlying themes in a collection of documents. Most topic models require the additional step of attaching meaningful labels to estimated topics, a process that is not scalable, suffers from human bias, and is difficult to replicate. We present a transfer topic labeling method that seeks to remedy these problems, using domain-specific codebooks as the knowledge base to automatically label estimated topics. We demonstrate our approach with a large-scale topic model analysis of the complete corpus of UK House of Commons speeches from 1935 to 2014, using the coding instructions of the Comparative Agendas Project to label topics. We evaluated our results using human expert coding and compared our approach with more current state-of-the-art neural methods. Our approach was simple to implement, compared favorably to expert judgments, and outperformed the neural networks model for a majority of the topics we estimated.

Keywords

Topic models, topic labeling, transfer learning, word embeddings, neural networks

Introduction

Political science scholars working with large quantities of textual data are often interested in discovering latent semantic structures in their document collections. Examples include legislative debates, policies, media content, manifestos, and open-ended survey questions. Political scientists increasingly use variations of probabilistic topic models (Blei et al., 2003) to summarize large text collections (see several recent examples: Baerg and Lowe, 2020; Bagozzi et al., 2018; Barnes and Hicks, 2018; Martin and McCrain, 2019; Mueller and Rauh, 2018; Munger et al., 2019; Pan and Chen, 2018). These models typically require the manual labeling of estimated latent dimensions. In practice, this means that researchers have to assign meanings to a list of words that these algorithms have identified as a latent "topic" – a requirement that is similar to labeling a dimension that emerges from a principal component analysis of some numerical data. This process of manual labeling is not scalable, may suffer from human bias, and is difficult to replicate.

In this article, we present a strategy for automatically labeling topics that is simple to implement, easy to replicate, and reduces the inherent human bias when labeling topics. Our strategy takes advantage of the fact that experts in our field have spent considerable time and resources to develop and refine codebooks for labeling text in different domains. Examples of such projects include the Comparative Agendas Project (CAP), work by the Manifesto Research Group, and the Congressional Bills Project. These codebooks contain predefined categories that are of interest to political scientists and typically include lengthy, written category descriptions to guide human coders. We argue that these well-defined codebooks

[1] Hertie School of Governance, Germany
[2] Clemson University, USA
[3] King's College London, UK

Corresponding author:
Peter John, Department of Political Economy, King's College London, Bush House (NE), 40 Aldwych, London, WC2B 4PH, UK.
Email: peter.john@kcl.ac.uk

contain a wealth of information that can be used to automatically transfer existing domain-specific knowledge to the process of topic labeling.

We illustrate the logic of our method with a large-scale topic analysis of the debates in the UK House of Commons from 1935 to 2014. We extracted 22 topics from this corpus, which we automatically labeled using the coding manual of the CAP (Bevan, 2014). We validated our results using human labeling of the topics by CAP expert coders (see Supplemental material). Our method applies more generally and could be easily extended to other areas with an existing domain-specific knowledge base, such as party manifestos, open-ended survey questions, social media analysis, and legal cases. Using our method, researchers in these fields can be more confident that the building blocks of their models are not an artifact of human coding decisions emerging from within the research process itself. In addition, by transferring labels from existing codebooks to estimated topics, our method allows for a tighter integration between the results of a text analysis and existing, domain-specific projects from which codebooks are drawn.

Related work

In the absence of roll-call data that can be used for ideal point estimation, scholars have turned to legislative speech to estimate policy positions, either by focusing on selected debates (e.g., Herzog and Benoit, 2015; Laver and Benoit, 2002) or through the analysis of all speeches during a legislative term (Lauderdale and Herzog, 2016). A parallel stream of the literature has used topic modeling to estimate the extent to which legislators speak on different topics (Quinn et al., 2010). Topic modeling is a class of models that estimate the underlying themes in a collection of documents. Originally proposed by Blei et al. (2003) in their seminal article on the latent Dirichlet allocation (LDA), various extensions of LDA have been developed (e.g., Blei and Lafferty 2006, 2007; Roberts et al., 2016; Teh et al., 2012).

Topic labeling is a key post-processing step of all probabilistic topic models. The topics that emerge from LDA and related methods are rankings of the words that appear in a corpus. Topic labeling is the manual step of assigning meanings to these word rankings. As a general rule, labels should be relevant, understandable, with high coverage inside topic, and discriminate across topics. Early research focused on generating labels by hand, using a set of top *n* words in a topic distribution (so called *cardinality*) learned by a topic model (Griffiths and Steyvers, 2004). This manual approach is not scalable, carries a high cognitive load in forming the topic concept and its interpretation (Aletras and Mittal, 2017), and also suffers from a potential bias of the human labeler (Lau and Baldwin, 2016).

An alternative approach is to implement a supervised topic modeling approach that limits the topics to a predefined set with their word distributions provided a priori (McAuliffe and Blei, 2008; Ramage et al., 2009). This approach is unable to pick up topics unknown beforehand (Wood et al., 2017). Keyword assisted topic models (Eshima et al., 2020) is a more recent development that allows seeding the topics with a dictionary of keywords, thereby constraining their generation and consequently composition. This supervised topic modeling approach can be used to generate meaningful topic labels via its supervision component. In contrast to this method, our proposed strategy does not constrain the topic estimation step, but instead uses external keyword lists to post hoc label topics.

Several automatic labeling approaches have been proposed in the literature that utilize external, contextual information, which is also the strategy we follow in this article. Mei et al. (2007) minimize the semantic distance between the topic model and the candidate label based on the phrases from inside documents. Lau et al. (2011) utilize various ranking mechanisms of the top *n* words and candidate labels from Wikipedia articles containing these terms. Most recently, Bhatia et al. (2016) have used word embeddings (Mikolov et al., 2013) to map topics and candidate labels derived from Wikipedia article titles, and then select topic labels based on cosine similarity and relative ranking measures.

Word embeddings pretrained on a large corpus, like Wikipedia, and deployed for topic labeling of PubMed abstracts as in Bhatia et al. (2016) are a simple form of general domain knowledge transfer. More generally, a machine learning framework captures the ability to transfer knowledge to new conditions, which is known as transfer learning (Pan and Yang, 2010). Our work builds on these earlier general approaches and develops a computationally quick and scalable topic labeling strategy that takes advantages of existing, domain-specific knowledge bases in political science.

Unsupervised topic modeling with transfer topic labeling

Our main idea is illustrated in Figure 1. The dotted box on the right-hand side illustrates traditional unsupervised topic modeling, which stops with estimated latent topics that need manual labeling. In our approach, we used outside expert codebooks to extract topic labels and associated keywords, which we then used to automatically label the estimated latent topics. Retaining human-in-the-loop or human interaction allows for adjustment of the labels for specific domains with sparse coverage in the source knowledge base.

In the remainder of this section, we demonstrate the utility of this approach with speeches from the UK House of Commons over the period 1935 to 2014. We first explain how we have estimated latent, dynamic topics from the speeches. We then discuss how we have used the

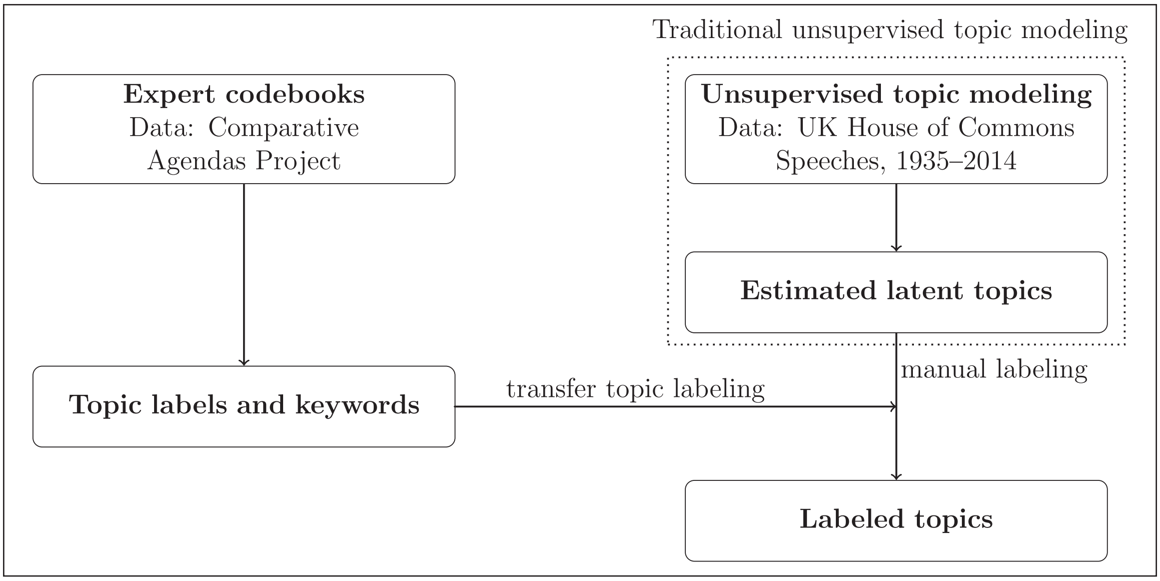
Traditional unsupervised topic modeling
Expert codebooks
Data: Comparative Agendas Project
Unsupervised topic modeling
Data: UK House of Commons Speeches, 1935–2014
Estimated latent topics
manual labeling
Topic labels and keywords
transfer topic labeling
Labeled topics

Figure 1. Illustration of the application of unsupervised topic modeling with transfer topic labeling.

codebooks from the CAP as an existing knowledge base to transfer topic labels.

Estimating dynamic topics from House of Commons speeches, 1935–2014

Our data consist of the complete record of debates from the UK House of Commons during the period 1935–2014. All debates and information about speakers were downloaded from TheyWorkForYou,[1] a transparency website that provides access to parliamentary records and information about Members of Parliament (MPs). All data were downloaded in XML format and were further processed in Python.

The full data set consists of about 4.3 million floor contributions with an average of 49,720 contributions per year (min = 17,280, max = 118,500, SD = 17,596) and a total of 117,914 unique words. Within each session, we combined each MP's contributions into a single text, excluding contributions that concern the rules of procedure or the business of the House, such as the reading of the parliamentary agenda or formal announcements. We also removed the traditional prayer at the beginning of each sitting and all contributions and announcements by the Speaker.

As part of the preprocessing, we applied stemming (reducing words to their root form), removed words that appeared fewer than 50 times and in fewer than 5 documents, removed punctuation, numbers, symbols, stopwords, hyphens, single letters, and a custom list of high-frequency terms.[2] The final data set from which we estimated topics included 47,524 documents (i.e., an MP's concatenated speeches during a session) and 19,185 unique words.

We used the dynamic topic model (DTM) by Blei and Lafferty (2006) to estimate topics from the speech data. Like any unsupervised topic model, DTM requires setting the number of topics a priori. We followed the standard in the literature and picked the number of topics based on semantic coherence and exclusivity (c.f., Roberts et al., 2016). Based on these two metrics, we selected a model with 22 topics.[3]

Extracting topic labels and keywords from expert codebooks

We used coding instructions from the CAP as our external source to extract topic labels and associated keywords. We selected the CAP because we expected the majority of parliamentary speeches to be on topics related to public policymaking. This attention to policy topics is the central interest to projects on the policy agenda, which have been active since the late 1990s (Baumgartner et al., 2013). What has been called the policy agendas code frame is a fuller articulation of policy topic ideas with a larger number of major topic codes, which aims at comprehensive coverage of any topic that is likely to appear.

While our demonstration of transfer topic labeling is limited to the CAP codebooks, we note that our method could be easily extended to other codebooks as long as they include written coding instructions or vignettes. Further, as we demonstrate below, our method for matching topic labels to estimated latent topics produces goodness-of-fit measures for each match, which allows for an evaluation of how well the topics derived from a codebook capture the estimated latent dimensions.

Table 1. Overview of Comparative Agendas Project topics.

Policy agenda topic	Top 10 words based on *tf-idf* weighting
Macroeconomic issues	tax, inflat, index, treasuri, fiscal, price, taxat, unemploy, bank, gold
Civil rights	discrimin, asylum, immigr, equal, right, citizenship, minor, age, refuge, freedom
Health	healthcar, care, health, medic, drug, coverag, nurs, provid, alcohol, mental
Agriculture	agricultur, farm, anim, food, livestock, produc, crop, erad, fisheri, diseas
Labour and employment	employ, labour, job, migrant, youth, worker, employe, workplac, work, train
Education and culture	educ, student, school, art, vocat, higher, secondari, teacher, grant, learn
Environment	water, pollut, environment, wast, hazard, conserv, emiss, climat, municip, air
Energy	electr, gas, energi, coal, oil, power, natur, nuclear, fuel, gasolin
Transportation	highway, transport, rail, truck, bus, road, ship, aviat, speed, air
Law and crime	crime, crimin, drug, justic, traffick, polic, juvenil, sentenc, court, offend
Social welfare	benefit, elder, volunt, social, food, welfar, incom, contributori, meal, lunch
Community development, planning and housing	hous, mortgag, urban, tenant, veteran, low, homeless, citi, rural, tenanc
Banking and finance	small, bankruptci, copyright, busi, patent, consum, mortgag, tourism, sport, mutual
Defence	defenc, weapon, arm, intellig, militari, forc, reserv, veteran, armi, war
Space science	scienc, space, radio, communic, satellit, tv, launch, telecommun, broadcast, research
Foreign trade	trade, export, tariff, import, invest, exchang, duti, competit, u.k, restrict
International affairs and foreign aid	european, soviet, east, u.n, africa, u.k, peac, polit, europ, treati
Government operations	postal, legislatur, execut, minist, employe, elect, census, elector, offici, prime
Public lands, water management	indigen, land, park, convey, histor, water, forest, monument, memori, reclam

The codebook for the UK Policy Agendas Project includes 19 major topics with subtopics.[4] For each subtopic, the CAP codebook provides written examples of what is being included in each category. For example, category "1. Macroeconomics – 100: General domestic macroeconomic issues" is described as follows:

> the government's economic plans, economic conditions and issues, economic growth and outlook, state of the economy, long-term economic needs, recessions, general economic policy, promote economic recovery and full employment, demographic changes, population trends, recession effects on regional and local economies, distribution of income, assuring an opportunity for employment to every person seeking work, standard of living.

Because the descriptions of CAP subtopics are relatively short, we combined all subtopics under a major topic label into a single document. We then applied *tf-idf* (term frequency–inverse document frequency) weighting to generate 19 weighted word lists (one for each major topic label), where the weight on each word reflected its importance to a topic label.[5] Table 1 provides an overview of the 19 topics together with their 10 highest ranked words.

Transfer topic labeling

We transferred topic labels from the CAP to the estimated latent topics through a pair-wise matching procedure that finds the most similar CAP topic word list for each latent dimension. For the CAP topics, the word lists were the aforementioned *tf-idf*-weighted word lists. For the DTM, we constructed one word list for each of the 22 estimated latent topics.[6]

We restricted the CAP word lists to the top 150 words to eliminate low-ranked words and to make sure that all word lists had the same length. For the estimated target topics, we limited matching to the top 20 words, which we found maximized the average Jaccard index in a parameter search considering word thresholds from 10, 15, and 50. However, we note that using fewer or more than 20 keywords yields similar results to those reported in this article.

To identify the best matching topics, we used the Jaccard similarity coefficient, which is a widely used set-based similarity measure. Jaccard is simple and often used in text-based similarity calculations. Furthermore, as we did not expect repetition to affect our similarity measure, Jaccard was an appropriate choice for this specific task.

The Jaccard similarity coefficient is defined as the measure of similarity between two sets. It is calculated by taking the size of the intersection of two sets divided by the size of their union. The Jaccard index is bound between 0 and 1, with higher numbers indicating a greater overlap between two sets, as demonstrated in equation (1).

$$Sim(s1, s2) = \frac{|s1 \cap s2|}{|s1 \cup s2|} \tag{1}$$

where $Sim(s1, s2)$ is the Jaccard similarity between sets of words $s1$ and $s2$.

We calculated the Jaccard index for each pair of word lists consisting of one CAP topic and one estimated DTM topic. This resulted in 19 unique matches based on the highest calculated Jaccard value, where the CAP label was transferred to the estimated DTM topic.

Results and evaluation

Table 2 provides an overview of the 22 estimated topics together with their Jaccard index, the matched CAP topic label, and the top 20 words from each DTM topic. As a validation exercise, we recruited a group of CAP experts to label the word lists for each topic according to the CAP categorization. Seventeen experts who participated in this exercise were provided with an online survey in which they were asked to pick two labels from the CAP codebook (most appropriate and second most appropriate) for each estimated topic. We assessed the quality of expert labeling using Fleiss's kappa measure of intercoder agreement. We also calculated the proportion of experts who agreed with the automatically selected topic label as their first or second choice. All measures are included in Table 2. We provide additional information on our expert coding validation exercise in the supplementary materials (see Section C).

The majority of experts agreed with the automatic approach on 12 topic labels. The clearest example being the topic of agriculture (#1) where transfer labeling and all the experts identified farming and agriculture-related terms. Further, four topics showed sufficiently large agreement between experts across two choices and automatic labeling (#13 government operations and #14 social welfare). The banking topic was labeled by a total of 12% of experts, but it also showed significant disagreement across experts with Fleiss's kappa at 0.3. For macroeconomics (#16) a majority of experts labeled it as transport, while 25% of the experts agreed with the automatic labeling of this topic as macroeconomics, which was their second choice (kappa = 0.46).

The remaining six entries in the table show complete disagreement between our automatic approach and experts. Not a single expert assigned the same label as the transfer-learning approach. These cases are difficult to explain as they both contain varied values for Jaccard and Fleiss's kappa. The topics morph into concerns about representation and territorial identity, moving away from these political topic. With transportation the match was for regional policies in Wales and England, which is an amalgam of keywords on transport. Another crossover was for social welfare that combined with legislative procedures, which reflects the extent to which MPs focus on social welfare in asking parliamentary questions. The importance of this topic is that it comes up four times with different word formations. The experts were possibly using the government operations label for catch-all procedural issues (e.g., in #18 and #22), or fitting a label to a topic that is not represented in CAP, like Northern Ireland (#23). In the latter case, the algorithm arguably more correctly applied the label of civil rights and minority issues. We provide additional validation results in the supplementary materials.

Finally, we note that three CAP categories from Table 1 were not matched: environment, space science, and public lands and water management.

Robustness study

Bhatia et al. (2016) presented a neural embedding approach that used Wikipedia titles to generate and rank topics. The authors compared their model to state-of-the-art systems and found that it is "simpler, more efficient, and achieves better results across a range of domains" (9). This approach is still generally considered to be state-of-the-art for automatic topic labeling. We therefore chose to apply this approach as our robustness study, and to label the parliamentary debates, using both the out-of-the-box (Wikipedia) labels and domain-specific labels. This method combines document (doc2vec) and word (word2vec) embeddings to select the most relevant labels for topics. The doc2vec embedding of a title is the embedding of the document the label is associated with. Its word2vec embedding is the result of generating word embeddings for the title.

Following Bhatia et al. (2016), for all experiments we used the Gensim (Řehůřek and Sojka, 2010) implementation of both doc2vec and word2vec. To this end, titles were treated as single tokens (e.g., concatenating financial crisis into financial_crisis) and then the text of all of the Wikipedia articles were greedily tokenized. The word embeddings for the tokens were built using the SkipGram algorithm (Mikolov et al., 2013). To generate the candidates, given a topic, the cosine similarity between the title embeddings (generated by either doc2vec or word2vec) and each of the word embeddings for the top-10 topic terms was calculated and aggregated by taking the arithmetic mean.

The titles that yielded the highest similarity scores were selected as the most relevant labels for the topic. The generated labels were ranked based on letter trigram overlap between a given topic label and the topic words (Kou et al., 2015).

In addition to an out-of-the-box implementation of Bhatia et al. (2016), we replicated the pipeline with domain-specific embeddings instead of the Wikipedia labels. We extracted the domain-specific labels using the UK Policy Agendas Codebook (PAC), treating each file in the codebook as a document whose label is the filename. We built the embeddings for each title, following the same process for the Wikipedia labels.

For the PAC, only eight of the labels had word2vec embeddings, therefore the final scores achieved by them were higher than the other labels. As a result, candidates

Table 2. Dynamic topic model topics with matched policy agenda topic labels and comparison to expert coding.

#	Topic label selected by	Topic label selected by	Prop. experts		Jaccard index	Fleiss's kappa	Top 20 words from estimated dynamic topics
	Transfer-learning approach	Experts	1st	2nd			
1	Agriculture	Agriculture	1.00	0	0.62	0.81	price, agricultur, food, suppli, ask, ration, milk, water, farmer, ministri, market, industri, fisheri, consum, sugar, beef, meat, fish, rural, increas
2	Labour and employment	Labour and employment	0.94	0	0.47	0.54	employ, industri, polic, men, labour, worker, union, work, unemploy, area, women, trade, law, wage, crime, home, court, factori, train, case
3	International affairs and foreign aid	International affairs and foreign aid	0.88	0.06	0.23	0.82	hous, european, question, matter, eu, committe, order, union, communiti, discuss, statement, europ, made, treati, constitut, countri, debat, point, answer, make
4	Defence	Defence	0.88	0	0.42	0.61	air, defenc, forc, ministri, civil, aviat, ireland, aircraft, aerodrom, servic, northern, broadcast, imperi, airway, afghanistan, televis, iraq, corpor, offic, fli
5	Community development, planning and housing issues	Community development, planning and housing issues	0.81	0.06	0.52	0.68	local, hous, author, council, build, road, work, rent, charg, plan, home, region, area, counti, rate, communiti, land, london, peopl, develop
6	Government operations	Government operations	0.75	0.12	0.27	0.40	scotland, scottish, state, vote, elect, elector, secretari, hous, parliament, commiss, regist, parti, assembl, system, ask, gallant, peopl, glasgow, awar, devolut
7	Foreign trade	Foreign trade	0.69	0.06	0.32	0.60	trade, hous, question, committe, industri, board, matter, export, countri, import, duti, answer, discuss, refer, presid, agreement, made, film, hope, british
8	Health	Health	0.56	0.44	0.52	0.52	school, educ, health, servic, care, author, hospit, evacu, nhs, children, patient, local, board, adopt, medic, peopl, teacher, area, univers, doctor
9	Transportation	Transportation	0.56	0.25	0.32	0.46	busi, london, steel, product, industri, ship, war, suppli, constitu, british, ministri, transport, research, peopl, aircraft, work, rail, vessel, firm, factori
10	Energy	Energy	0.56	0.19	0.52	0.43	agricultur, coal, industri, energi, land, farmer, board, oil, farm, miner, gas, subsidi, power, water, scheme, climat, british, committe, electr, carbon
11	Labour and employment	Labour and employment	0.50	0.12	0.13	0.54	question, pension, peopl, sir, figur, work, benefit, answer, increas, million, inform, rate, refer, repli, report, cent, part, gallant, committe, matter
12	Energy	Energy/Labour and employment[1]	0.38	0.19	0.37	0.48	coal, industri, employ, unemploy, board, area, mine, job, peopl, fuel, develop, train, electr, miner, transport, region, men, work, east, north
13	Government operations	Law, crime, and family issues	0.31	0.25	0.18	0.51	peopl, home, ask, point, speaker, hous, constitu, case, offic, polic, order, general, agre, debat, secretari, prison, man, awar, post, public
14	Social Welfare	Macroeconomics	0.25	0.06	0.42	0.18	secretari, state, tax, chancellor, peopl, benefit, pension, exchequ, cut, war, incom, social, hous, purchas, problem, profit, minist, compani, duti, govern
15	Banking, finance, and domestic commerce	Social welfare	0.06	0.06	0.23	0.30	pension, nation, price, unemploy, industri, case, assist, insur, increas, busi, benefit, compani, british, peopl, age, offic, man, widow, board, allow
16	Macroeconomics	Transportation	0	0.25	0.32	0.46	secretari, transport, state, railway, tax, road, industri, compani, price, bank, subsidi, commiss, vehicl, trade, nationalis, control, servic, chancellor, union, privat
17	Foreign trade	International affairs and foreign aid	0	0	0.37	0.82	countri, state, commonwealth, coloni, leagu, british, intern, unit, india, foreign, secretari, majesti, ask, syria, develop, german, south, peopl, rhodesia, world
18	Social welfare	Government operations	0	0	0.18	0.40	amend, claus, point, committe, hous, learn, move, order, debat, case, matter, word, beg, line, act, deal, provis, make, legisl, law
19	Social welfare	Education	0	0	0.27	0.53	ask, secretari, state, school, educ, awar, offic, statement, war, armi, servic, make, teacher, children, men, admiralti, view, step, forc, releas
21	International affairs and foreign aid	Transportation	0	0	0.18	0.46	ask, wale, welsh, assembl, secretari, road, transport, war, awar, state, view, north, east, railway, learn, author, step, region, number, local
22	Social welfare	Government operations	0	0	0.27	0.40	matter, question, case, sir, answer, sport, made, act, local, author, fund, inform, report, point, time, nation, person, concern, regul, servic
23	Civil rights, minority issues, and civil liberties	Government operations	0	0	0.23	0.40	ireland, northern, countri, point, peopl, polic, war, hous, speech, great, time, parti, irish, speaker, debat, issu, order, opposit, state, agreement

Note: Column "Prop. experts" is the proportion of experts who selected the same topic as the transfer-learning approach with their first or second choice.

[1] Experts were tied between topic label "Energy" and "Labour and employment." Fleiss's kappa reported in this row is the average of the kappas for each label.

were limited to these eight labels. Moreover, we found that the quality of the embeddings was quite poor. This is likely to be due to the small size of the codebook: the PAC files contained a total of only 15,188 words.

We evaluated these results by comparing the labels chosen by the models with labels chosen by two human annotators. The annotators were asked to choose the topics based on a list of labels generated using the UK PAC. We calculated interannotator agreement for the human labels based on a weighted Cohen's kappa, which used a predefined table of weights to measure the degree of disagreement, and found a good agreement (0.51).

Table 3 provides an overview of the 22 estimated topics, side by side with the topics estimated by the model proposed by Bhatia et al. (2016). In the first two columns of Table 3, we present the top three ranked labels produced by the two models described in this section. We compared the labels to those chosen by the human annotators, presented in the fourth column. Where only one label is shown means the two annotators chose the same label.

As we can see in the table, most of the labels were not meaningful and did not match up with those chosen by the human annotators. The labels in Column 2 are more directly comparable with those of the transfer-labeling approach (Column 3) and those chosen by the human annotators (Column 4), as they were based on the PAC topics.

The labels matched with the transfer topic labels showed the strongest agreement with the human annotators. When looking at only the top ranked label, this matching approach had a kappa of 0.31 agreement with the first human annotator and 0.23 with the second. This agreement went up when the top three highest ranking labels were taken into account. In total, one of the top three ranked labels matched a human annotator's choice for 16 out of 22 cases. In contrast, the PAC labels chosen by the neural model only matched the annotator's choice in 5 out of 22 cases. This was a large improvement over the labels chosen by the neural methods, even with the domain-specific embeddings.

Conclusion

Treating text as data is an approach of increasing importance in political science. Natural language processing techniques developed in the computing sciences are routinely added to methodological toolkits. Topic modeling is a favorite tool of document summarizing. Political scientists often have to be creative in interpreting and labeling estimated topics; yet such labeling is also often difficult to replicate – a *sine qua non* of modern political science.

Arguably, the task of coding a small number of topics, as in the examples above, could be designed with sufficient reliability and replicability built in. However, in practice we often estimate models with a much higher number of topics. Increasing the number of topics makes human labeling less scalable and arguably increases the effects of human biases in labeling. In order to demonstrate the scalability of our approach, we estimated the topics on the full data set of parliamentary debates between 1935 and 2014.

To address the deficiency of current labeling techniques, and to have a better way of accommodating change over time, we presented a new method for topic labeling. Our approach provided an automatic labeling method that transferred the wealth of substantive knowledge accumulated in political science into labeling topic models. By doing so, it would allow researchers to integrate the results of a topic model into their existing coding framework. Our approach is also fully transparent and replicable, which would allow the bringing of human expertise to bear on difficult cases.

As part of the robustness analysis, we compared this method to current state-of-the-art neural network methods. Even when trained with in-domain embeddings, these methods did not adequately match topics and failed to correlate with human judgment. Furthermore, these methods are computationally expensive and time-consuming. At the same time we found our proposed method to be much faster and capable of producing more accurate labels. A full manual analysis of these labels showed high correlation with human experts.

An important limitation of our approach is that its success depends on the linguistic similarity and overlap in vocabulary between the corpus, from which topics are estimated, and the knowledge base used to pick labels. The analysis presented in this article worked well because the topics and words we expected to find in legislative speech were well captured by the CAP coding framework. Our method would be less applicable when trying, for example, to pick CAP category labels for topics estimated from Twitter messages. Another limitation is that reference topics extracted from existing knowledge bases need to be sufficiently distinct. If this is not the case, one may find that a small number of "catch all" reference topics will dominate the matching procedure. A possible avenue for future research is developing metrics that will inform researchers a priori how well an existing knowledge base is suited to extracting reference topics for a particular corpus. Another avenue is developing transfer methods that combine the versatility and linguistic context of neural embeddings with the computational simplicity of the Jaccard index we used.

While our proposed approach is simple and specific to the political domain, it could be extended to other domains where extensive and rich codebooks are available, such as party manifestos, open-ended survey questions, social media data, legal documents, and other research domains where topic models have made advances in recent years.

Table 3. Topic labels matched by all three systems and compared with human matches.

#	Topic label selected by			
	Bhatia with Wiki embeddings	Bhatia with PAC embeddings	Transfer-labeling approach	Human annotators
1	employment payment income	transportation government operations foreign trade	Social welfare Labour and employment Community development	Social welfare Labour and employment

2	constitution
	paisley_(scottish_parliament_constituency)
	dumfries_(scottish_parliament_constituency)
		transportation
		foreign trade
		agriculture
			Government operations
			Civil rights
			Public lands and water management
				Civil rights
				Government operations

3	we do_something everything	government operations agriculture foreign trade	Social welfare Civil rights Community development	Law and crime Social welfare
4	ensure the_case addition	transportation foreign trade health	Law and crime Civil rights Defence	Law and crime
5	region community industry	transportation agriculture government operations	Macroeconomic issues Social welfare Labour and employment	Civil rights Government operations
6	transport_network transport train	foreign trade government operations energy	Transportation Social welfare Community development	Transportation
7	we everything fact	government operations health environment	Social welfare Community development Education and culture	Government operations
8	subject matter reason	foreign trade agriculture transportation	Social welfare Macroeconomic issues Civil rights	Government operations
9	ensure particular nature	transportation environment social welfare	Agriculture Macroeconomic issues Environment	Environment
10	ensure information finance	transportation foreign trade agriculture	Foreign trade Government operations Macroeconomic issues	Macroeconomics Labour and employment
11	addition country government	transportation government operations agriculture	Foreign Trade International affairs and foreign aid Social welfare	International affairs and foreign aid
12	money addition results	transport government operations agriculture	Community development Social welfare Transportation	Macroeconomics Social welfare
13	european integration question matter	transportation agriculture government operations	International affairs and foreign aid Civil rights Government operations	International affairs and foreign aid
14	mental health health care government	transportation health government operations	Health Social welfare Government operations	Health
15	teacher school college	transportation foreign trade government operations	Education and culture Social welfare Labour and employment	Education and culture
16	eventually fact particular	transportation agriculture government operations	Macroeconomic issues Community development Banking and finance	Macroeconomics Foreign trade
17	finance the financial money	transportation agriculture health	Macroeconomic issues Foreign trade Banking and finance	Banking and finance

(Continued)

Table 3. (Continued)

#	Topic label selected by			
	Bhatia with Wiki embeddings	Bhatia with PAC embeddings	Transfer-labeling approach	Human annotators
18	military addition devonshire regiment	transportation foreign trade agriculture	Defence Social welfare Community development	Defence
19	particular government addition	transportation agriculture foreign trade	Public lands and water management Civil rights Social welfare	Civil rights Defence
20	we question take	foreign trade transport agriculture	Government operations Public lands and water management Civil rights	Government operations
21	addition particular at the time	transport agriculture government operations	Foreign trade Social welfare Banking and finance	Education and culture Social welfare
22	industry pricing investment	transportation agriculture government operations	Energy Banking and finance Macroeconomic issues	Energy

PAC: UK Policy Agendas Codebook.

Acknowledgements

All authors have contributed equally. We thank Stefano Gurciullo for collaboration and research assistance. We are graateful to Christina Fernandes for editorial work on the manuscript.

Declaration of conflicting interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We thank University College London for financial support through its Grand Challenges.

ORCID iDs

Slava Jankin https://orcid.org/0000-0001-6915-177X
Peter John https://orcid.org/0000-0002-7934-1187

Supplemental material

See supplementary materials for appendices. The supplementary files are available at http://journals.sagepub.com/doi/suppl/10.1177/20531680211022206.
The replication files are available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DT3KUE

Notes

1. https://www.theyworkforyou.com/
2. We used the fairly extensive MySQL stopword list, which includes more than 500 words. Our custom list includes the following words: hon, mr, member, members, bill, minister, prime, government(s), friend, year(s), gentleman, gentlemen.
3. Further details and references regarding model selection are provided in the supplementary materials, Section A.
4. This codeframe for the UK is summarized in John et al. (2013) and on the UK project website (www.policyagendas.org.uk/).
5. Before calculating *tf-idf* weights, we applied the same pre-processing rules that we applied to the speech data to increase the similarity between the two vocabularies.
6. Additional information on our implementation of transfer topic labeling is provided in the supplementary materials, Section B.

Carnegie Corporation of New York Grant

This publication was made possible (in part) by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

References

Aletras N and Mittal A (2017) Labeling topics with images using a neural network. In: Jose J, et al. (eds) Advances in Information Retrieval. ECIR 2017. *Lecture Notes in Computer Science* 10193. Cham: Springer, 500–505.

Baerg N and Lowe W (2020) A textual Taylor rule: Estimating central bank preferences combining topic and scaling methods. *Political Science Research and Methods* 8(1): 106–122.

Bagozzi BE and Berliner D (2018) The politics of scrutiny in human rights monitoring: Evidence from structural topic models of US State Department human rights reports. *Political Science Research and Methods* 6(4): 661–677.

Barnes L and Hicks T (2018) Making austerity popular: The media and mass attitudes toward fiscal policy. *American Journal of Political Science* 62(2): 340–354.

Baumgartner FR, Green-Pedersen C and Jones BD (2013) *Comparative Studies of Policy Agendas*. Abingdon: Routledge.

Bevan S (2014) Gone fishing: The creation of the comparative agendas project master codebook. In: Baumgartner FR, Breunig C and Grossman E (eds) *Comparative Policy Agendas: Theory, Tools, Data*. Oxford: Oxford University Press, pp. 219–242.

Bhatia S, Lau JH and Baldwin T (2016) Automatic labelling of topics with neural embeddings. In: *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (eds Matsumoto Y, Prasad R), Osaka, Japan, December 11-17 2016, pp. 953–963.

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Blei DM and Lafferty JD (2006) Dynamic topic models. In: *Proceedings of the 23rd international conference on Machine learning* (eds Cohen W, Moore A), City, Country, June 2006, pp. 113–120.: ACM.

Blei DM and Lafferty JD (2007) A correlated topic model of science. *The Annals of Applied Statistics* 1(1): 17–35.

Eshima S, Imai K and Sasaki T (2020) Keyword assisted topic models. *arXiv:2004.05964*.

Griffiths TL and Steyvers M (2004) Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1): 5228–5235.

Herzog A and Benoit K (2015) The most unkindest cuts: Speaker selection and expressed government dissent during economic crisis. *Journal of Politics* 77(4): 1157–1175.

John P, Bertelli A, Jennings W, et al. (2013) *Policy Agendas in British Politics*. London: Palgrave MacMillan.

Kou W, Li F and Baldwin T (2015) Automatic labelling of topic models using word vectors and letter trigram vectors. In: *Proceedings of the eleventh Asian Information Retrieval Societies conference*, (eds Zuccon G, Geva S, Joho H, Scholer F, Sun A, Zhang P), Brisbane, Australia, 4 December 2015, pp. 253–264. Publisher location: AIRS.

Lau JH, Grieser K, Newman D, et al. (2011) Automatic labelling of topic models. In: *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human language technologies – Volume 1*, (ed Editor A), City, Country, xx–xx Month Year, pp. 1536–1545. Publisher location: Association for Computational Linguistics.

Lau JH and Baldwin T (2016) The sensitivity of topic coherence evaluation to topic cardinality. In: *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, (eds Knight K, Nenkova A, Rambow A), San Diego, USA, June 2016, pp. 483–487. Publisher location: Association for Computational Linguistics.

Lauderdale BE and Herzog A (2016) Measuring political positions from legislative speech. *Political Analysis* 24(3): 374–394.

Laver M and Benoit K (2002) Locating TDs in policy spaces: The computational text analysis of D'ail speeches. *Irish Political Studies* 17(1): 59–73.

Martin J and McCrain J (2019) Local news and national politics. *American Political Science Review* 113(2): 372–384.

McAuliffe JD and Blei DM (2008) Supervised topic models. In: *Advances in neural information processing systems, In Advances in Neural Information Processing Systems* (eds Schölkopf B, Platt J, Hofmann T), pp. 121–128, Cambridge, MA: MIT.

McAuliffe JD and Blei DM (2008) Supervised topic models. In: *Advances in neural information processing systems*. pp. 121–128.

Mei Q, Shen X and Zhai C (2007) Automatic labeling of multinomial topic models. In: *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, (eds Berkhin P, Caruana R, Xindong Wu), San Jose California USA, August, 2007, pp. 490–499. Publisher location: ACM.

Mikolov T, Sutskever I, Chen K, et al. (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26: arXiv:1310.4546.

Mueller H and Rauh C (2018) Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review* 112(2): 358–375.

Munger K, Bonneau R, Nagler J, et al. (2019) Elites tweet to get feet off the streets: Measuring regime social media strategies during protest. *Political Science Research and Methods* 7(4): 815–834.

Pan J and Chen K (2018) Concealing corruption: How Chinese officials distort upward reporting of online grievances. *The American Political Science Review* 112(3): 602–620.

Pan SJ and Yang Q (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10): 1345–1359.

Quinn KM, Monroe BL, Colaresi M, et al. (2010) How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54(1): 209–228.

Ramage D, Hall D, Nallapati R, et al. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1*, (eds Koehn P, Mihalcea RA), Singapore, August 2009, pp. 248–256. Publisher location: Association for Computational Linguistics.

Řehůřek R and Sojka P (2010) Software framework for topic modelling with large corpora. In: *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*, (eds Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D), Valletta, Malta, 22 May 2010, pp. 45–50. Publisher location: ELRA.

Roberts ME, Stewart BM and Airoldi EM (2016) A model of text for experimentation in the social sciences. *Journal of the American Statistical Association* 111(515): 988–1003.

Teh YW, Jordan MI, Beal MJ, et al. (2012) Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476): 1566–1581.

Wood J, Tan P, Wang W, et al. (2017) Source-LDA: Enhancing probabilistic topic models using prior knowledge sources. In: *2017 IEEE 33rd international conference on data engineering*, (eds Papakonstantinou Y, Yanlei Diao D), San Diego, CA, US, San Diego, CA, US, 19-22 April 2017, pp. 411–422. Publisher location: IEEE.