

# Transfer Topic Labeling with Domain-Specific Knowledge Base

An Analysis of UK House of Commons Speeches  
1935-2014

[arXiv:1806.00793](https://arxiv.org/abs/1806.00793) [cs.CL]

Slava Jankin Mikhaylov

8 November 2018

1. Related work

2. Transfer topic labeling

3. House of Commons Debates use case (MVP)

- Unsupervised topic modeling (DTM)
- Expert codebooks (CAP)
- Transferring the labels (text matching)
- Evaluation

4. Results

# Motivation

	Topic1	Topic2	Topic3	Topic4	Topic5
prob.1	states	peace	nations	people	united
prob.2	united	international	united	united	nations
prob.3	international	country	international	countries	international
prob.4	peace	african	development	peoples	security
prob.5	countries	africa	security	independence	european

# Motivation

	Topic1	Topic2	Topic3	Topic4	Topic5
prob.1	states	peace	nations	people	united
prob.2	united	international	united	united	nations
prob.3	international	country	international	countries	international
prob.4	peace	african	development	peoples	security
prob.5	countries	africa	security	independence	european
prob.6	nuclear	community	cooperation	states	rights
prob.7	republic	republic	world	struggle	human
prob.8	world	government	peace	africa	must
prob.9	nations	united	countries	world	council
prob.10	soviet	development	economic	nations	also



# Motivation

	Topic1	Topic2	Topic3	Topic4	Topic5
prob.1	states	peace	nations	people	united
prob.2	united	international	united	united	nations
prob.3	international	country	international	countries	international
prob.4	peace	african	development	peoples	security
prob.5	countries	africa	security	independence	european
prob.6	nuclear	community	cooperation	states	rights
prob.7	republic	republic	world	struggle	human
prob.8	world	government	peace	africa	must
prob.9	nations	united	countries	world	council
prob.10	soviet	development	economic	nations	also
prob.11	security	organization	efforts	government	new
prob.12	relations	security	new	south	union
prob.13	weapons	people	states	international	europe
prob.14	disarmament	nations	council	peace	law
prob.15	union	must	human	republic	world

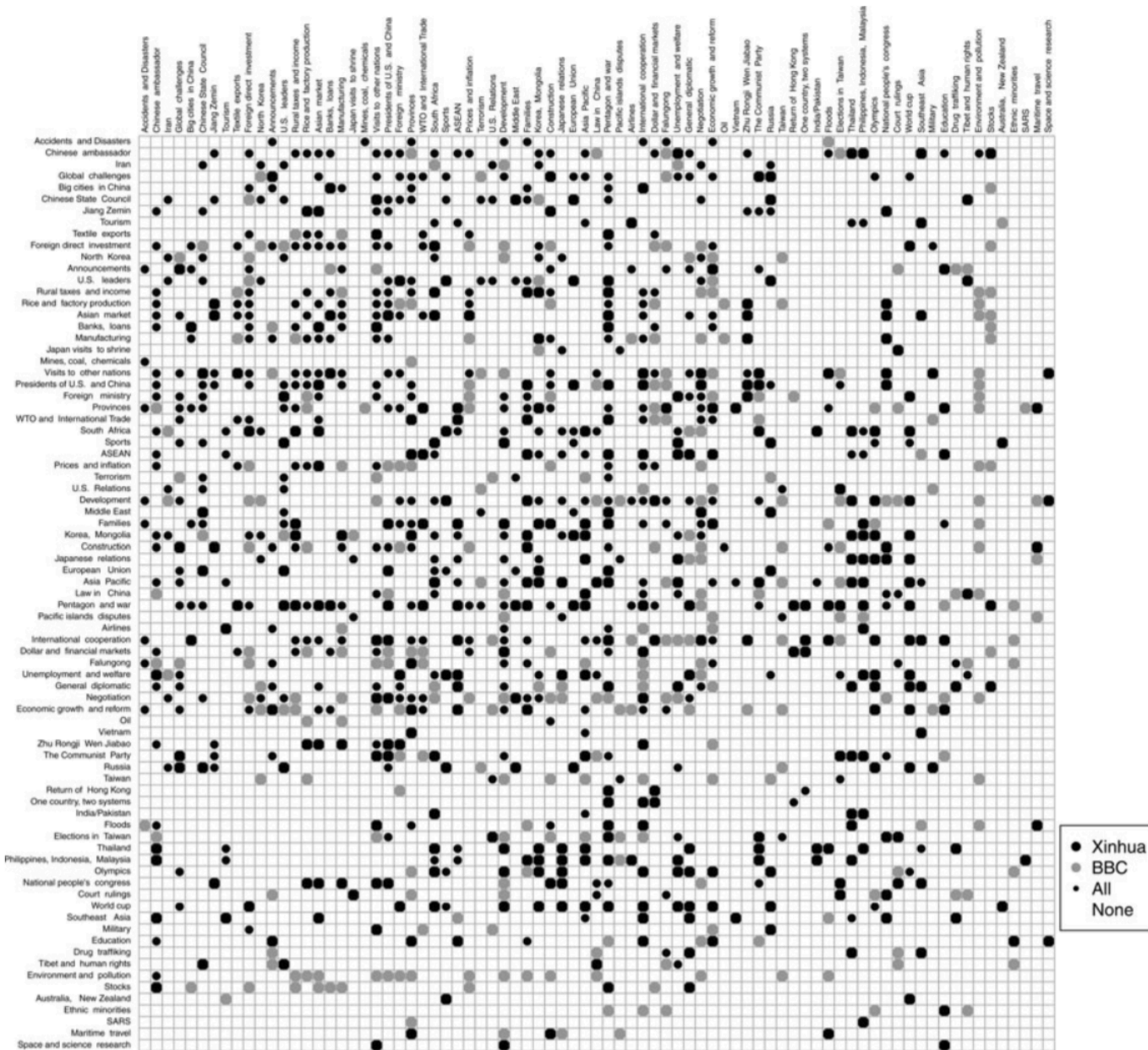


Figure 9. Correlation between topics, Xinhua versus BBC.

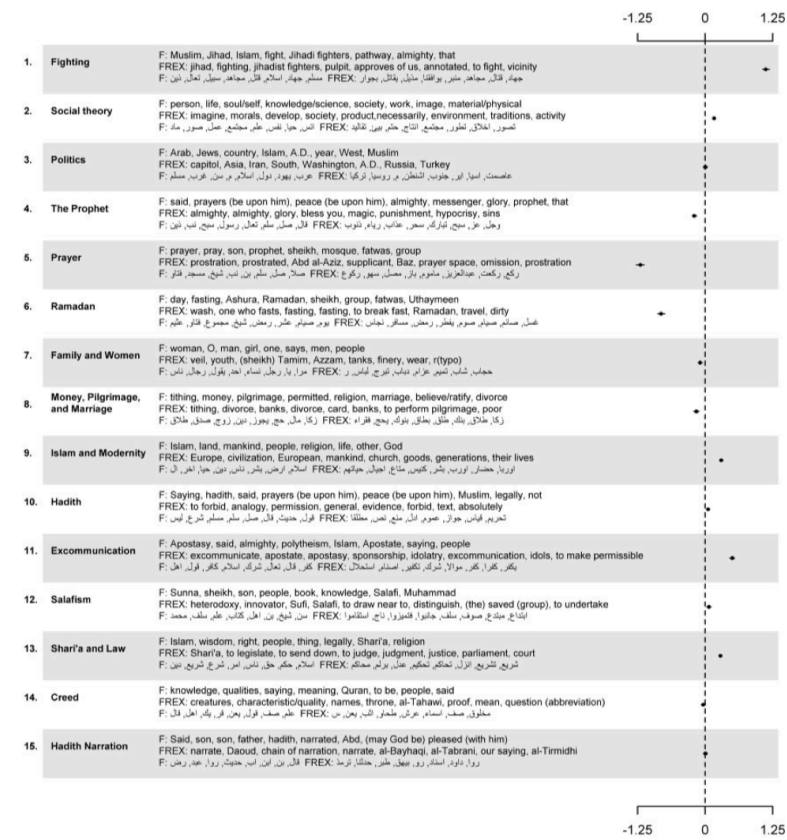
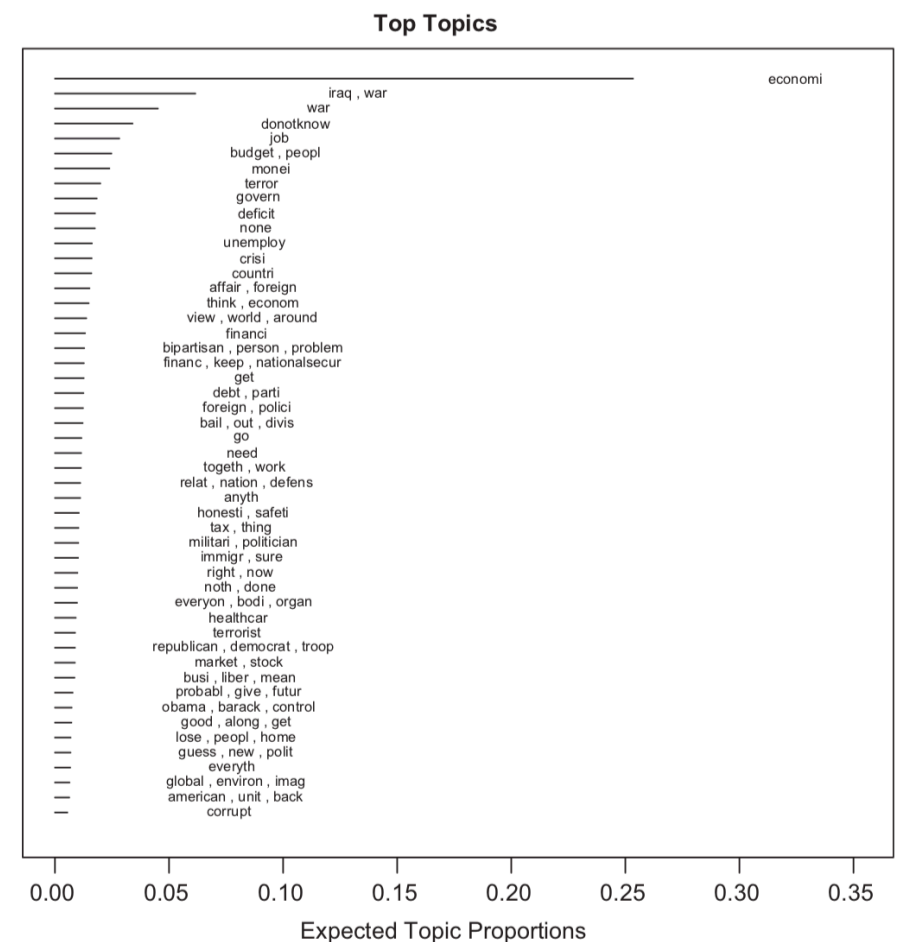


Fig. 1 Coefficients and standard errors for a 15-topic Structural Topic Model with Jihadi/not-Jihadi as the predictor of topics in Arab Muslim cleric writings. The words used to label each topic are shown on the left. "F:" indicates words that occur most frequently in each topic. "FREX:" indicates words that are frequent and exclusive to each topic. The Arabic words are in their stemmed form.

FIGURE 17 STM Topics from ANES Most Important Problem

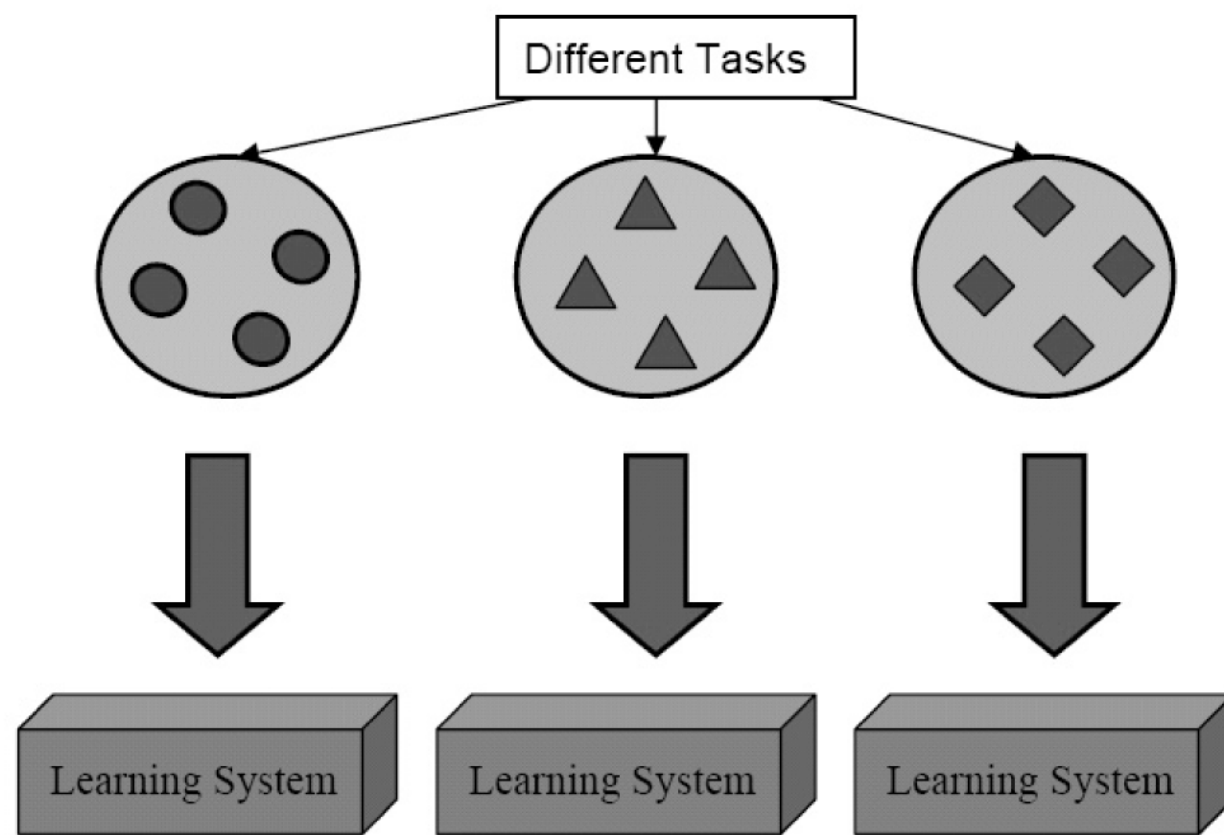


# Related Work

- **Preference of actors:** Wordscore, Wordfish, Wordshoal.
- **Political attention of actors:** Quinn et al. (2010), Grimmer (2009).
- **Concept drift:** Gama et al. (2014), Lowe et al. (2011), Benoit et al. (2016), Blei and Lafferty (2006).
- **Topic labeling:** Blei, Ng, Jordan (2003), Mcauliffe and Blei (2008), Lau et al. (2011), Bhatia et al. (2016).
- **Expert coding:** Comparative Agendas Project, Manifesto Project, European Election Study, Comparative Constitutions Project, CIRI Human Rights Data Project.

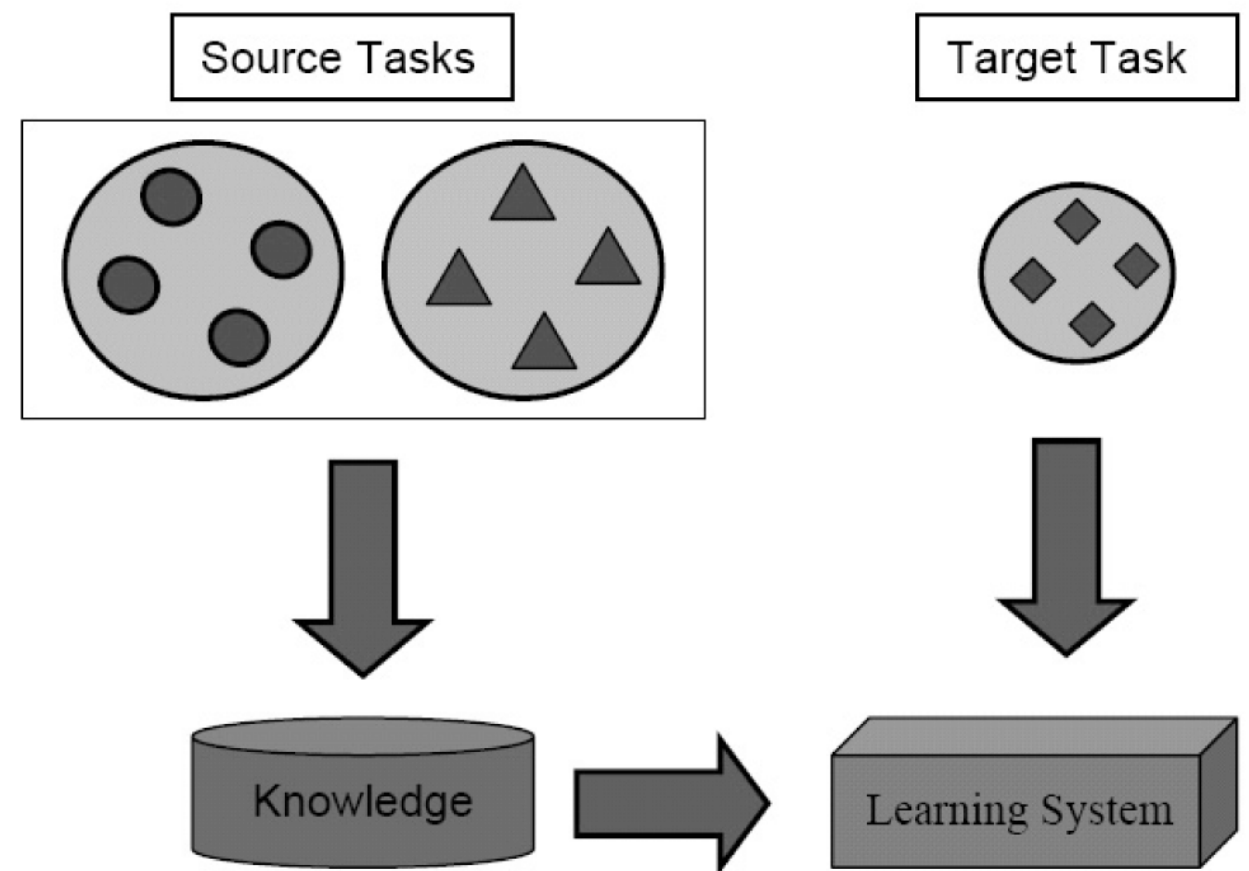
# Transfer Learning

Learning Process of Traditional Machine Learning



(a)

Learning Process of Transfer Learning

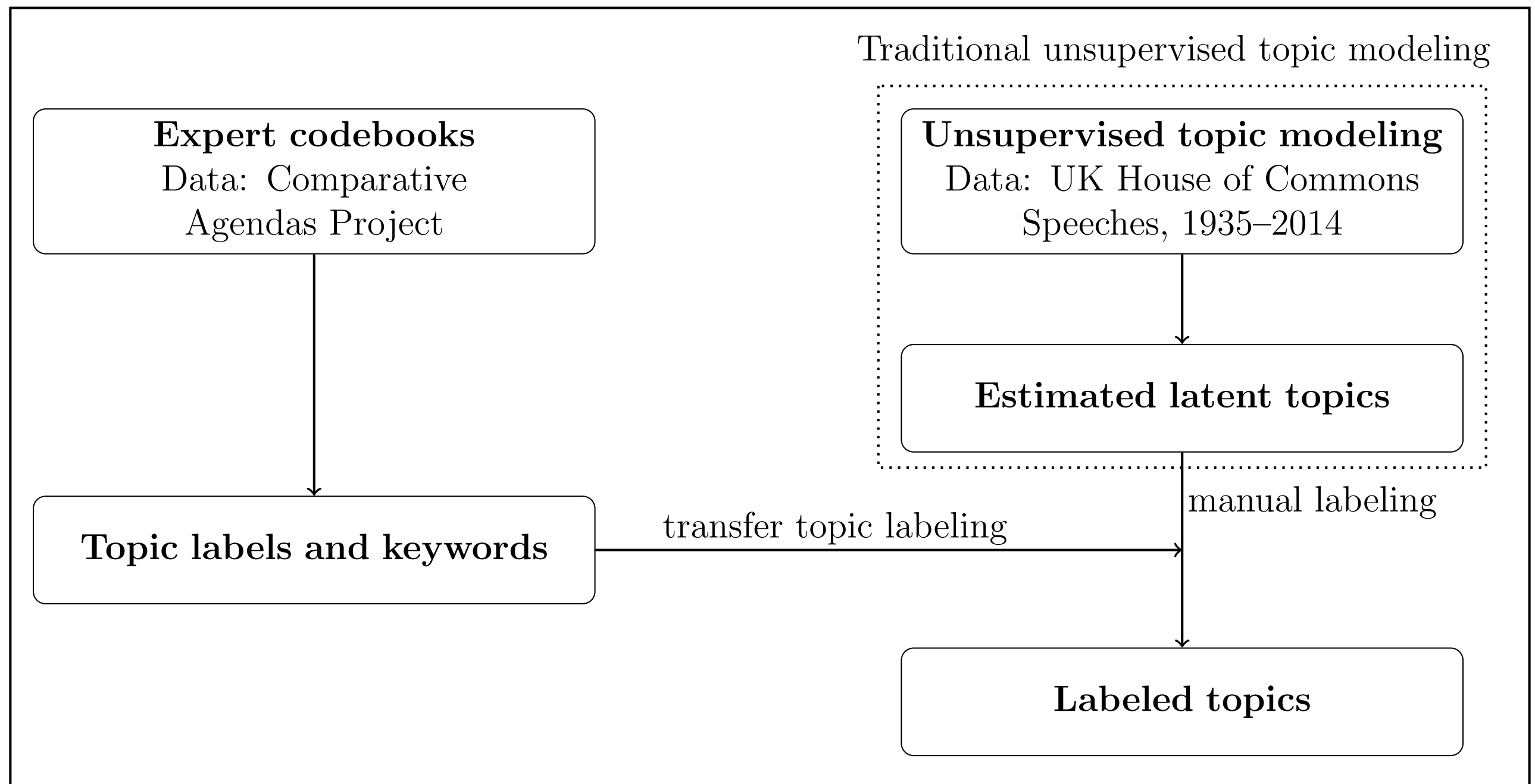


(b)

Source: Pan and Yang, 2010

# Transfer Topic Labeling

Unsupervised topic modeling with transfer topic labeling



# UK House of Commons Speeches, 1935-2014

- 4.3 million floor contributions from 1935-2014 (79 legislative sessions), downloaded from [www.theyworkforyou.com](http://www.theyworkforyou.com)
- Preprocessing:
  - combined each MP's contributions into a single text for each legislative session
  - stemming
  - removed stopwords, punctuation, and numbers
  - removed words that appeared less than 50 times and in fewer than 5 documents
- Final data set includes 47,524 documents, 19,185 unique words, 3,557 unique MPs

# Transfer Topic Labeling

**Expert codebooks**

(Comparative Agendas Project)



**Topic labels and keywords**

**Unsupervised topic modeling**

(UK House of Commons Debates,  
1935-2014)



**Estimated latent topics**



**Labeled topics**





# Estimation

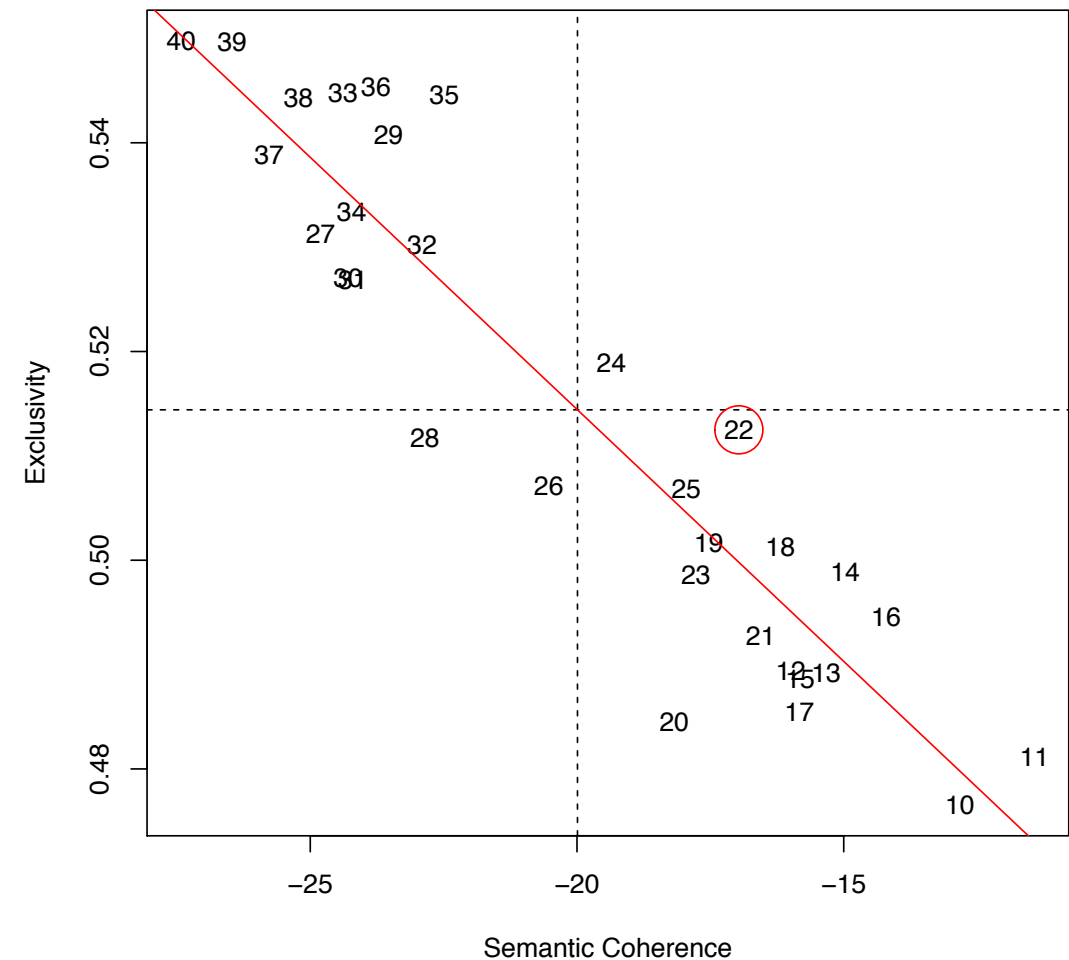
- Original Dynamic Topic Model (Blei and Lafferty 2006), C code.
- Estimated 31 models, varying  $k$  from 10 to 40
- Used compute resources on Amazon Web Services (AWS) to estimate all 31 models in parallel



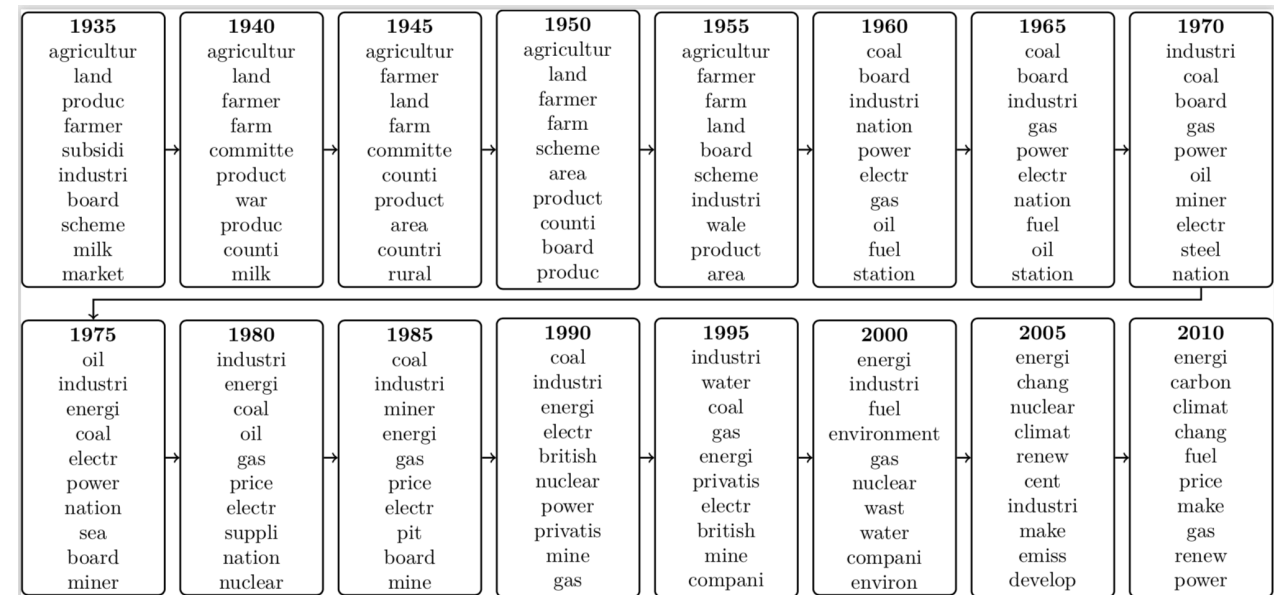
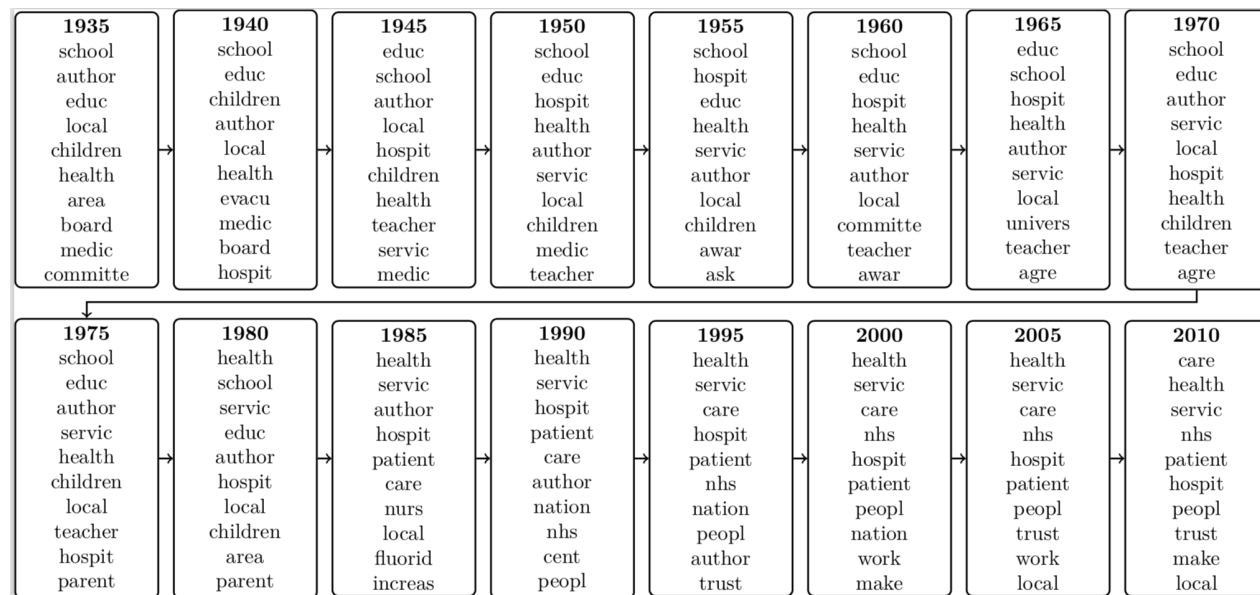
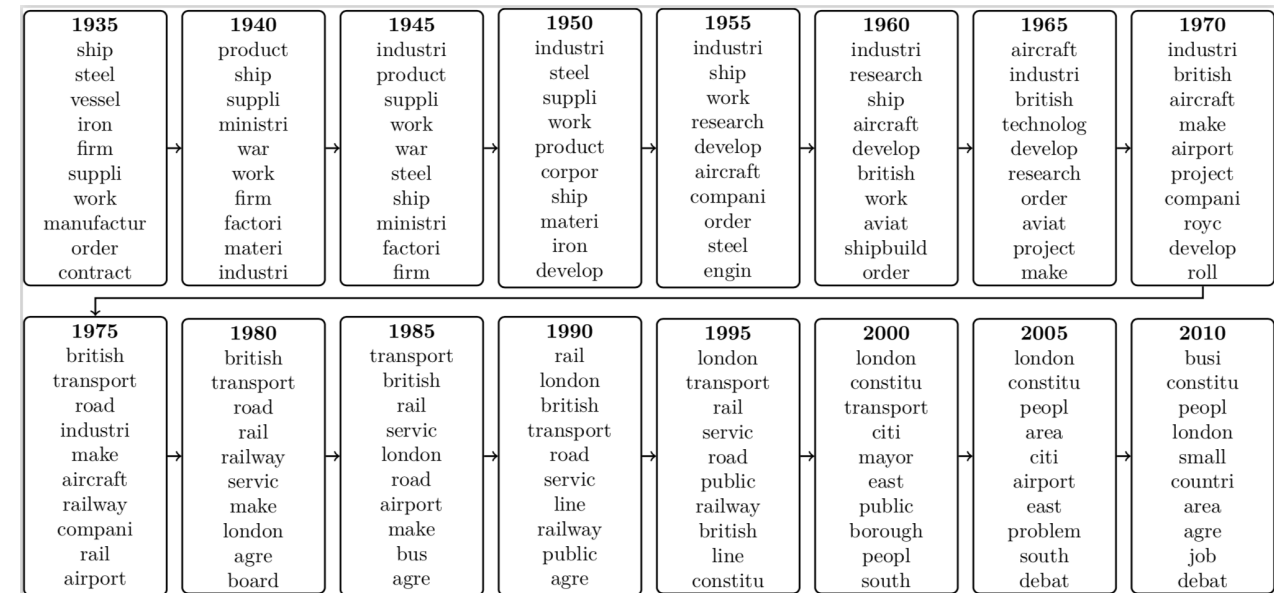
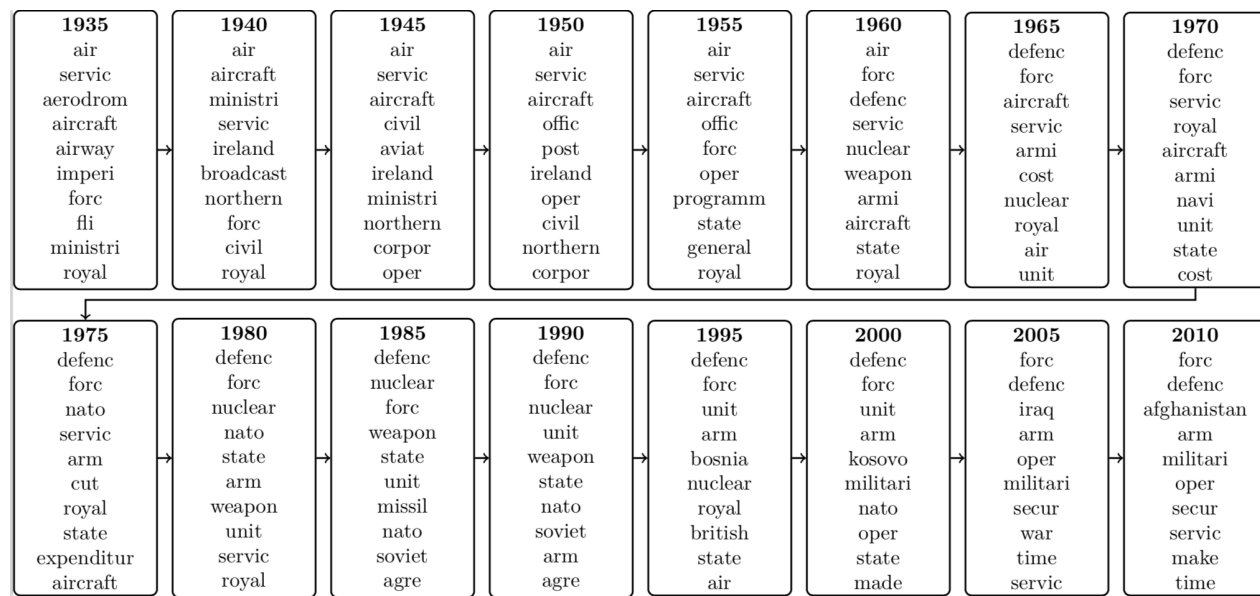


# Model Selection

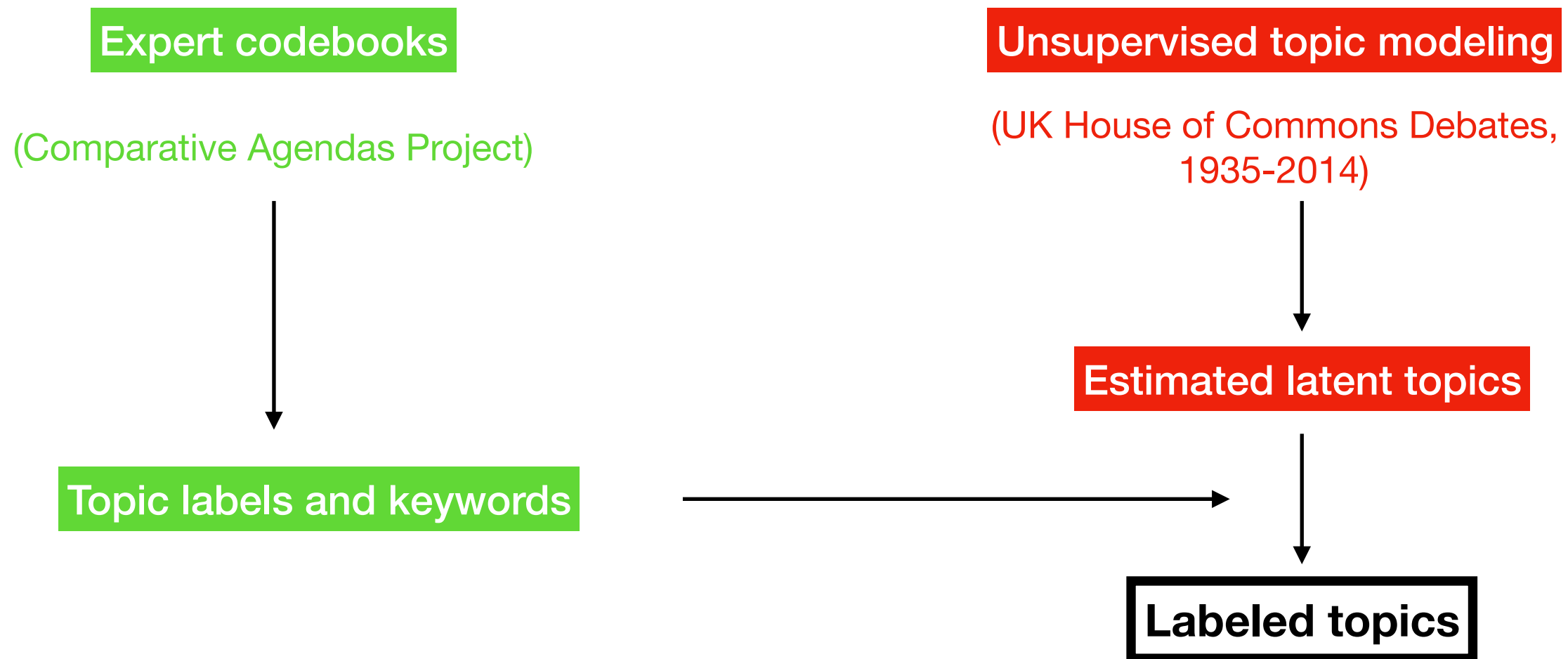
- Selected the “best” model based on **semantic coherence** and **exclusivity** (Mimno et al., 2011, Roberts et al., 2014, 2016)
- **Semantic Coherence:** Measures the extent to which top words in a topic jointly appear in documents.
- **Exclusivity:** Measures the extent to which words with high probability in one topic have low probabilities in other topics.



# Estimated Topics



# Transfer Topic Labeling



# Comparative Agendas Project

## 1. Macroeconomics

### **100: General domestic macroeconomic issues**

Examples: 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.

### **101: Inflation, prices, and interest rates**

Examples: inflation control and reduction, anti-inflation programs, calculation of inflation statistics and price index statistics, consumer price index (“retail price index”), food prices, cost of living, interest rates, government reports on inflation, effects of inflation on business, general economic statistics, delegation of responsibility for setting interest rates to the central bank (i.e. Bank of England).

### **103: Unemployment rate**

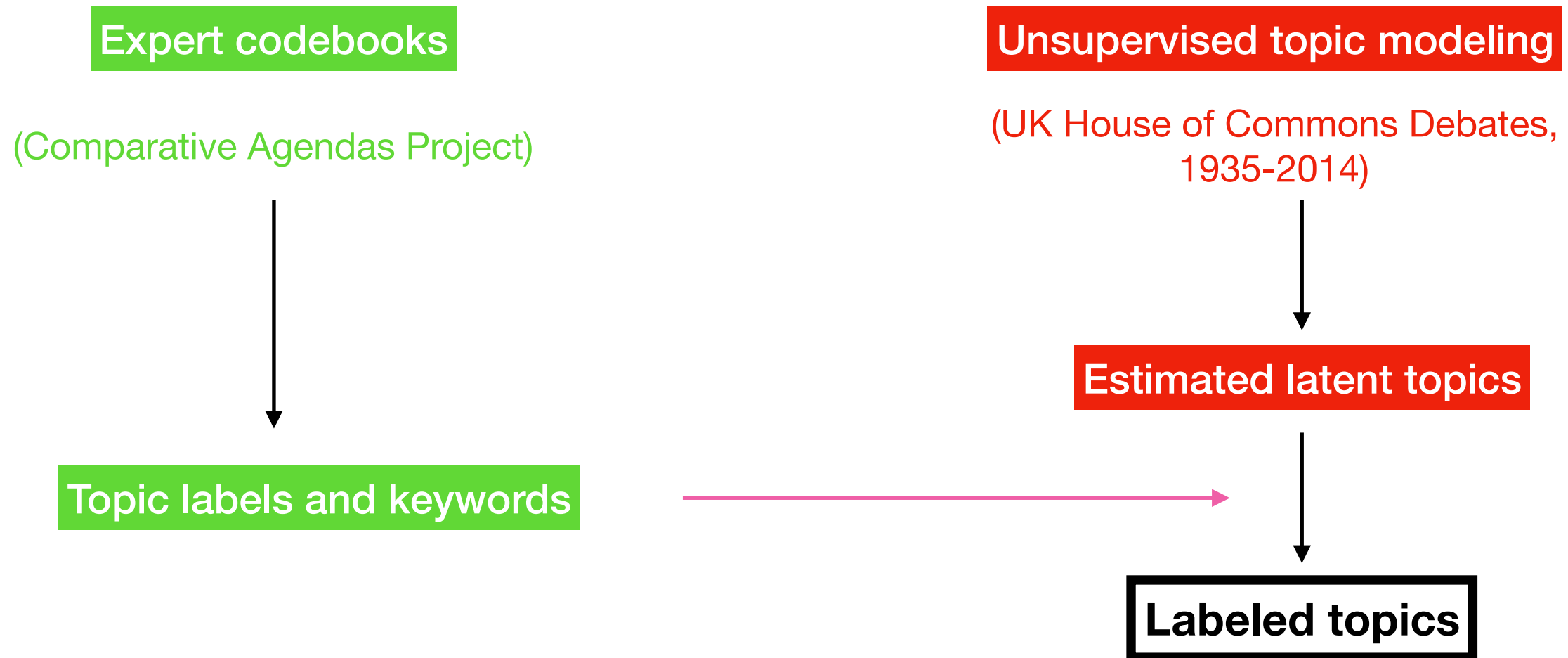
Examples: unemployment and employment statistics, economic and social impact of unemployment, national employment priorities, employment and labour market development, government reports on unemployment.

...

# Comparative Agendas Project

Policy agenda topic	Top ten words based on <i>tf-idf</i> 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
Labor 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
Defense	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

# Transfer Topic Labeling



# Transfer Labeling

## Jaccard index

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

**A:** Set of top  $k$  words from an estimated topic

**B:** Set of weighted keywords from a Comparative Agendas Project topic

# Example

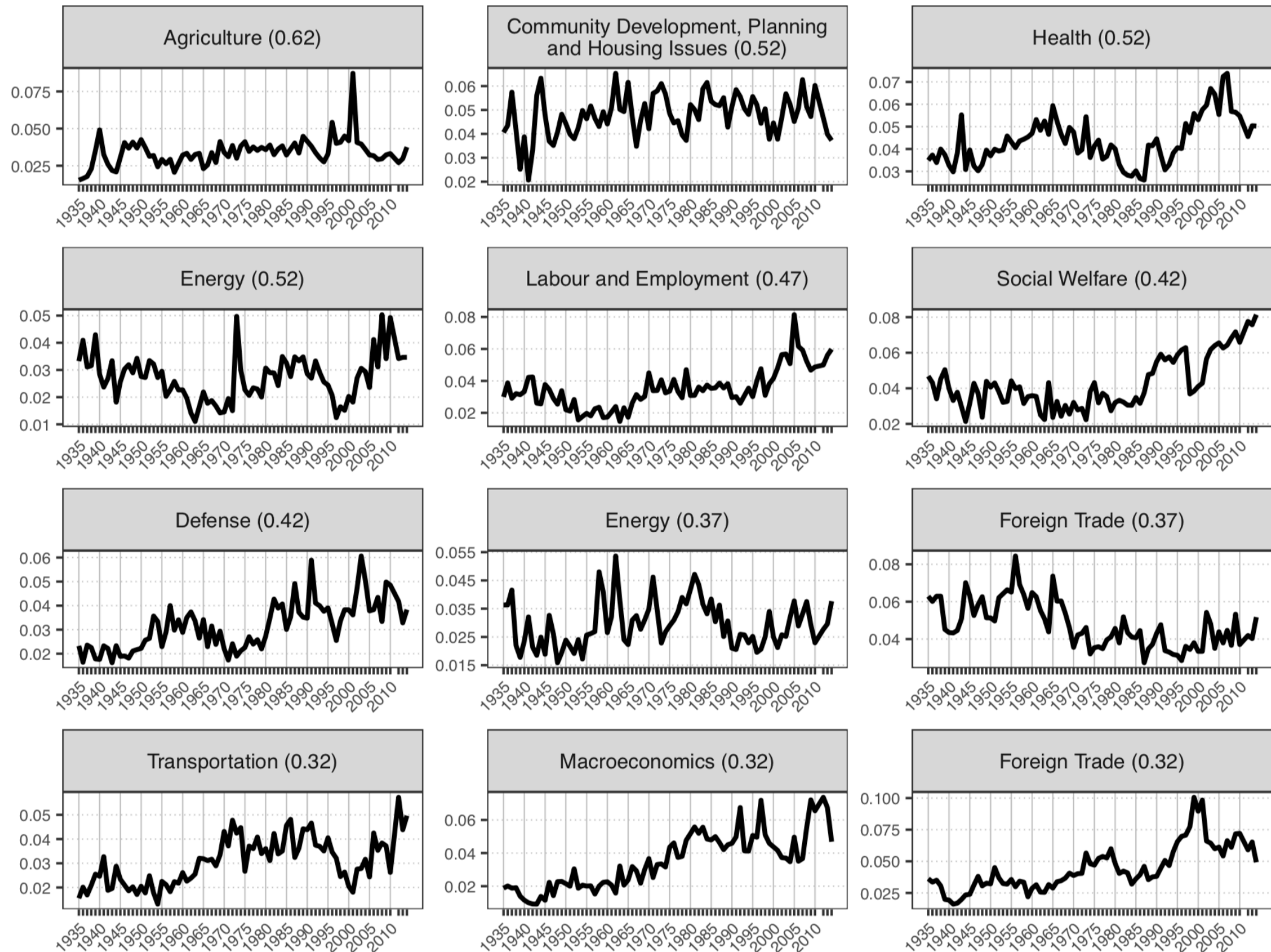
DTM topic	CAP topic "Agriculture"	CAP topic "Health"
food price agricultur farmer import	food price agricultur farmer crop	healthcar health drug coverag price
$J(A, B) = \frac{ A \cap B }{ A \cup B } = \frac{4}{6} = \frac{2}{3}$		

DTM topic	CAP topic "Agriculture"	CAP topic "Health"
food price agricultur farmer import	food price agricultur farmer crop	healthcar health drug coverag price

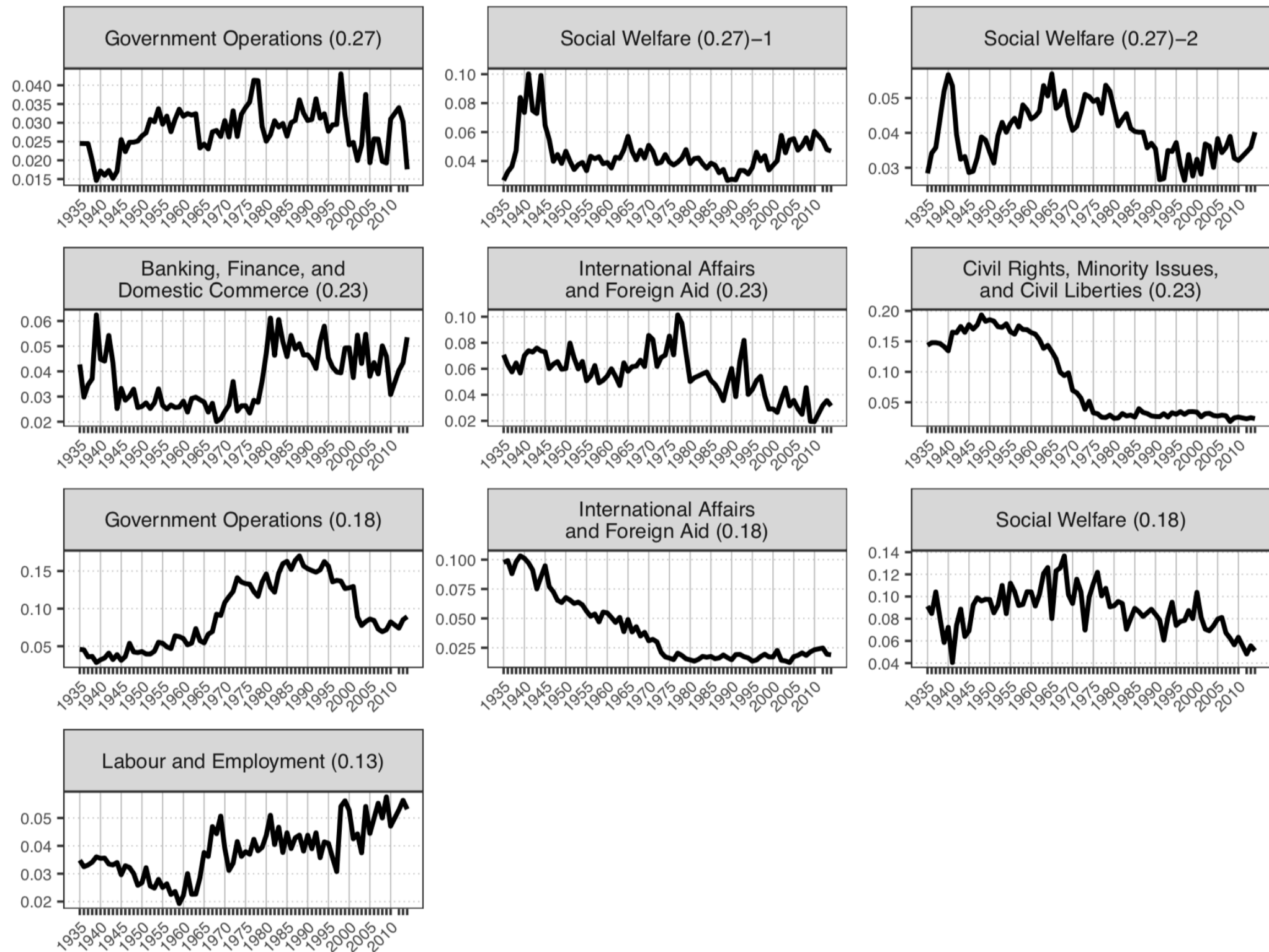
DTM topic	CAP topic "Agriculture"	CAP topic "Health"
food price agricultur farmer import	food price agricultur farmer crop	healthcar health drug coverag price
$J(A, B) = \frac{ A \cap B }{ A \cup B } = \frac{4}{6} = \frac{2}{3}$		$J(A, B) = \frac{ A \cap B }{ A \cup B } = \frac{1}{9}$



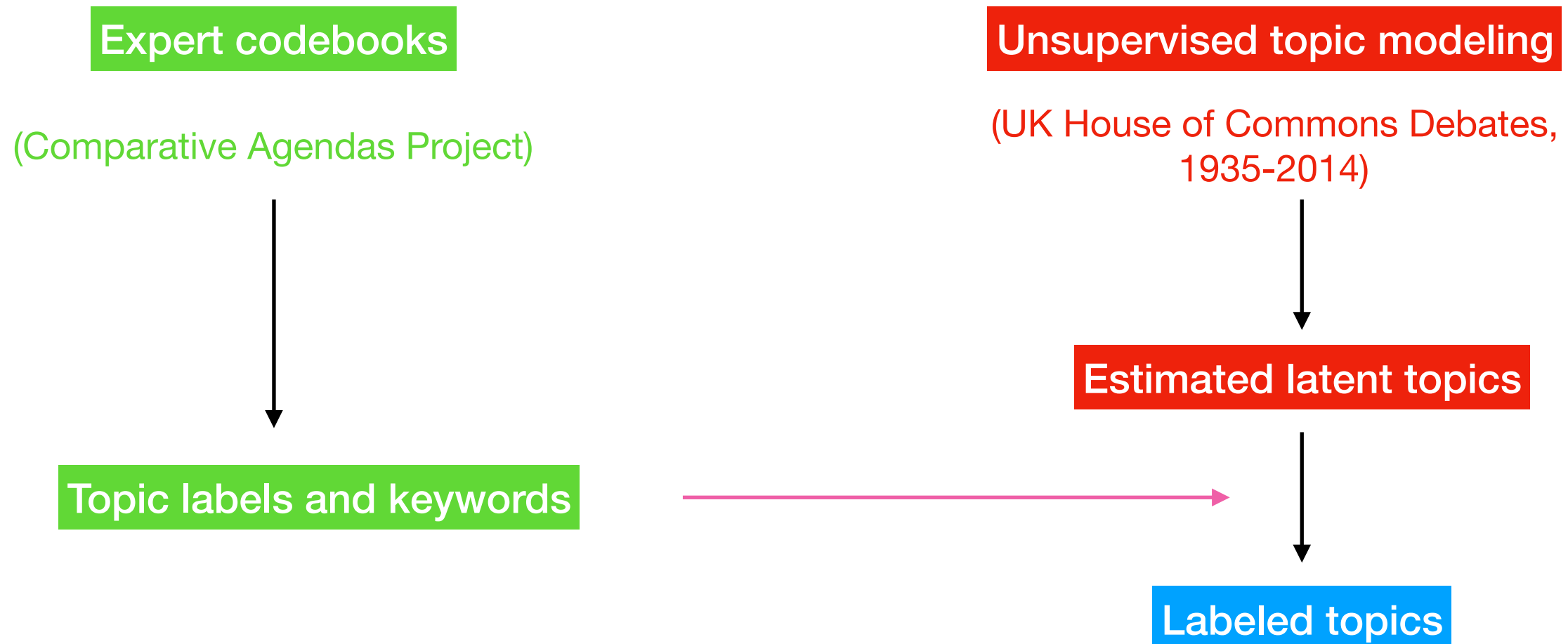
# High matching topics



# Low matching topics



# Transfer Topic Labeling



# Evaluation

## Coding of UK House of Commons Speech Topics


2 / 27

### Instructions

- Some words that you will see have been reduced to their word stem. For example, you will see the word "educ", which may represent "education", "educate", "educating", and similar words with the same stem. Another example is "hospit", which can mean "hospital", "hospitalise", and related words.
- You can click on a policy topic to open the official coding instructions from the UK Policy Agendas Project with examples for each category.

Prev Page

Next Page

Powered by  
  
See how easy it is to [create a survey](#).

## Coding of UK House of Commons Speech Topics

3 / 27

\* Which of the following major policy topics best describes the words below?

"busi, london, steel, product, industri, ship, war, suppli, constitu, british, ministri, transport, research, peopl, aircraft, work, rail, vessel, firm, factori"

- ☐ 1. [Macroeconomics](#)
- ☐ 2. [Civil Rights, Minority Issues, and Civil Liberties](#)
- ☐ 3. [Health](#)
- ☐ 4. [Agriculture](#)
- ☐ 5. [Labour and Employment](#)
- ☐ 6. [Education](#)
- ☐ 7. [Environment](#)
- ☐ 8. [Energy](#)
- ☐ 10. [Transportation](#)
- ☐ 12. [Law, Crime, and Family Issues](#)
- ☐ 13. [Social Welfare](#)
- ☐ 14. [Community Development, Planning and Housing Issues](#)
- ☐ 15. [Banking, Finance, and Domestic Commerce](#)
- ☐ 16. [Defense](#)
- ☐ 17. [Space, Science, Technology and Communications](#)
- ☐ 18. [Foreign Trade](#)
- ☐ 19. [International Affairs and Foreign Aid](#)
- ☐ 20. [Government Operations](#)
- ☐ 21. [Public Lands, Water Management, Colonial and Territorial Issues](#)

\* How well does the policy topic you selected describe the list of words?

- ☐ Extremely well
- ☐ Very well
- ☐ Somewhat well
- ☐ Not so well
- ☐ Not at all well

## Coding of UK House of Commons Speech Topics

3 / 27

\* Which of the following major policy topics best describes the words below?

"busi, london, steel, product, industri, ship, war, suppli, constitu, british, ministri, transport, research, peopl, aircraft, work, rail, vessel, firm, factori"

- ☐ 1. [Macroeconomics](#)
- ☐ 2. [Civil Rights, Minority Issues, and Civil Liberties](#)
- ☐ 3. [Health](#)
- ☐ 4. [Agriculture](#)
- ☐ 5. [Labour and Employment](#)
- ☐ 6. [Education](#)
- ☐ 7. [Environment](#)
- ☐ 8. [Energy](#)
- ☐ 10. [Transportation](#)
- ☐ 12. [Law, Crime, and Family Issues](#)
- ☐ 13. [Social Welfare](#)
- ☐ 14. [Community Development, Planning and Housing Issues](#)
- ☐ 15. [Banking, Finance, and Domestic Commerce](#)
- ☐ 16. [Defense](#)
- ☐ 17. [Space, Science, Technology and Communications](#)
- ☐ 18. [Foreign Trade](#)
- ☐ 19. [International Affairs and Foreign Aid](#)
- ☐ 20. [Government Operations](#)
- ☐ 21. [Public Lands, Water Management, Colonial and Territorial Issues](#)

\* How well does the policy topic you selected describe the list of words?

- ☐ Extremely well
- ☐ Very well
- ☐ Somewhat well
- ☐ Not so well
- ☐ Not at all well

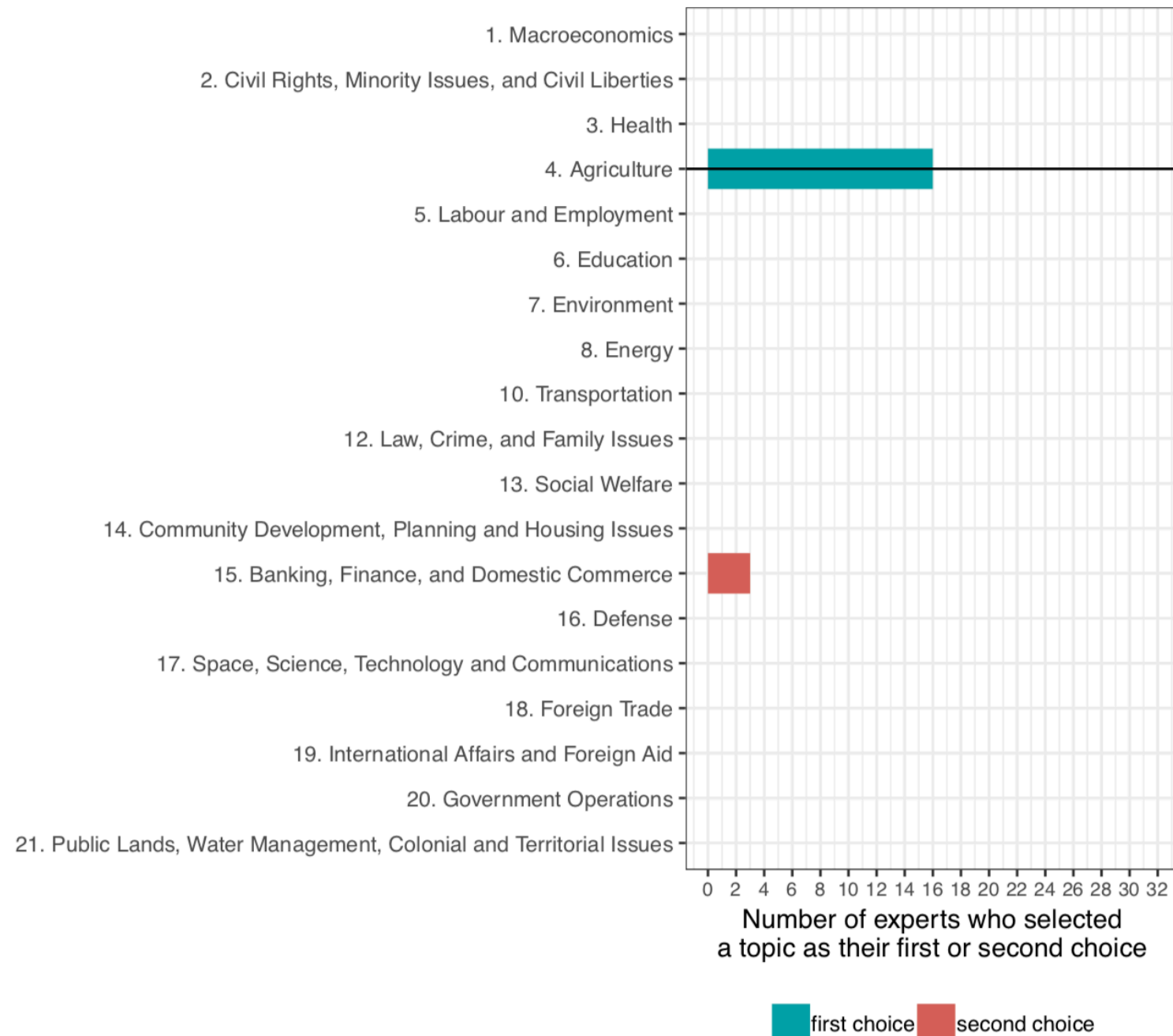
# Results

#	Topic Label Selected by Transfer-Learning Approach	Topic Label Selected by Experts	Prop. 1st	Experts 2nd	Jaccard Index	Fleiss' Kappa	Top 20 Words from Estimated Dynamic Topics
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	Defense	Defense	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 <sup>1</sup>	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

# SOTA(?)

Topic Label: Agriculture

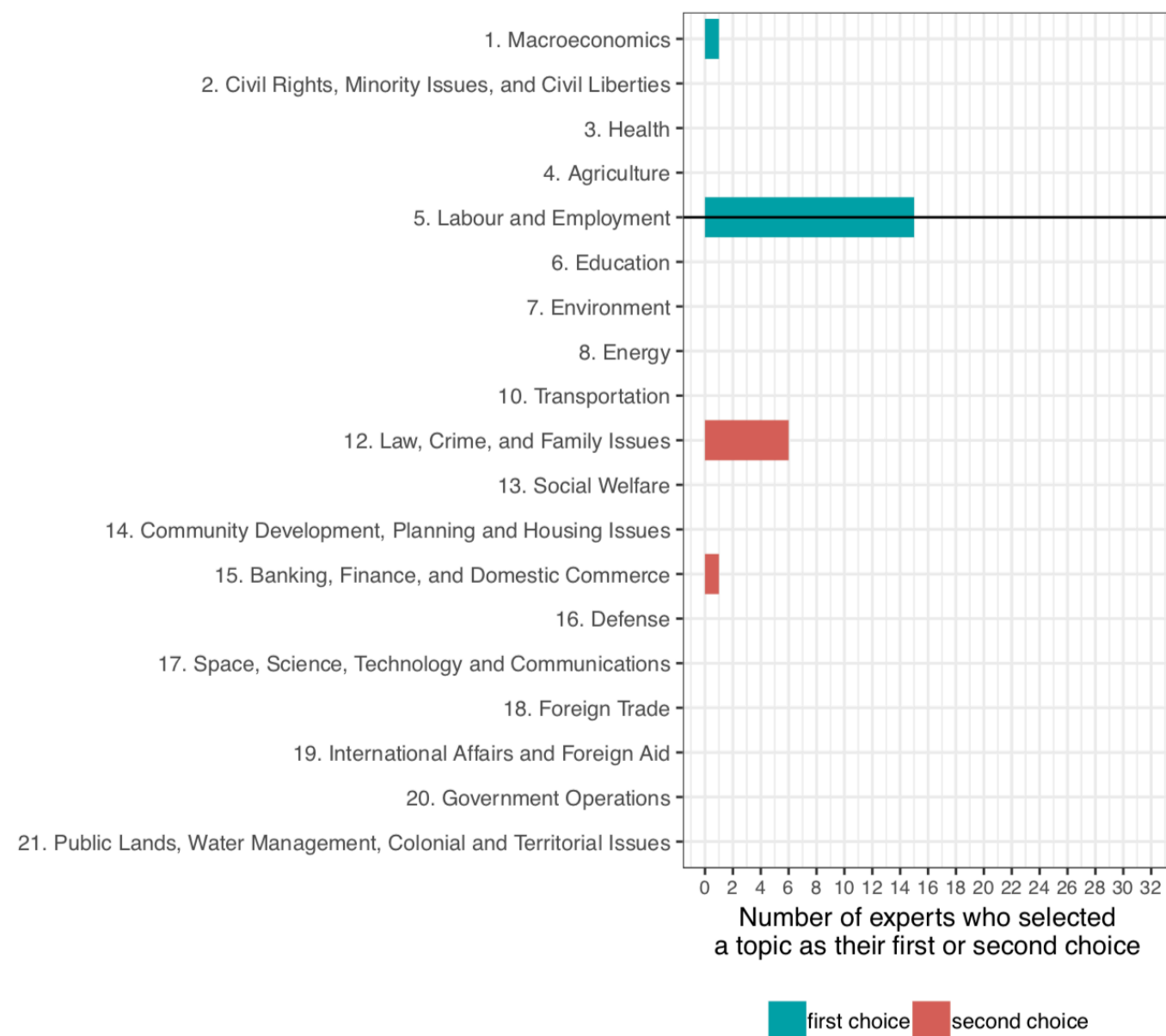
Top 20 words: "price, agricultur, food, suppli, ask, ration, milk, water, farmer, ministri  
market, industri, fisheri, consum, sugar, beef, meat, fish, rural, increas"



# SOTA(?)

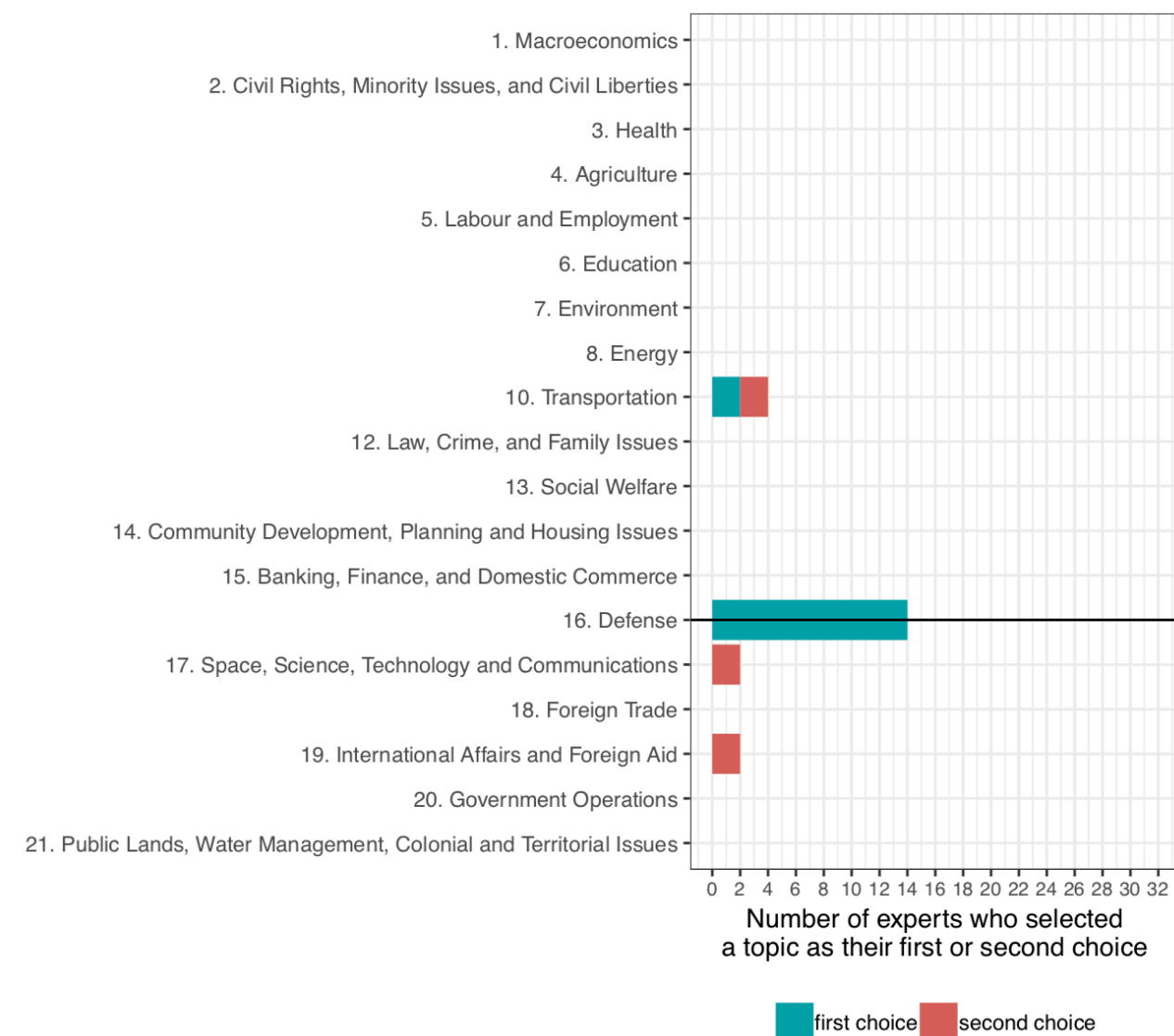
Topic Label: Labour and Employment

Top 20 words:"employ, industri, polic, men, labour, worker, union, work, unemploy, area women, trade, law, wage, crime, home, court, factori, train, case"



Topic Label: Defense

Top 20 words:"air, defenc, forc, ministri, civil, aviat, ireland, aircraft, aerodrom, servic northern, broadcast, imperi, airway, afghanistan, televis, iraq, corpor, offic, fli"

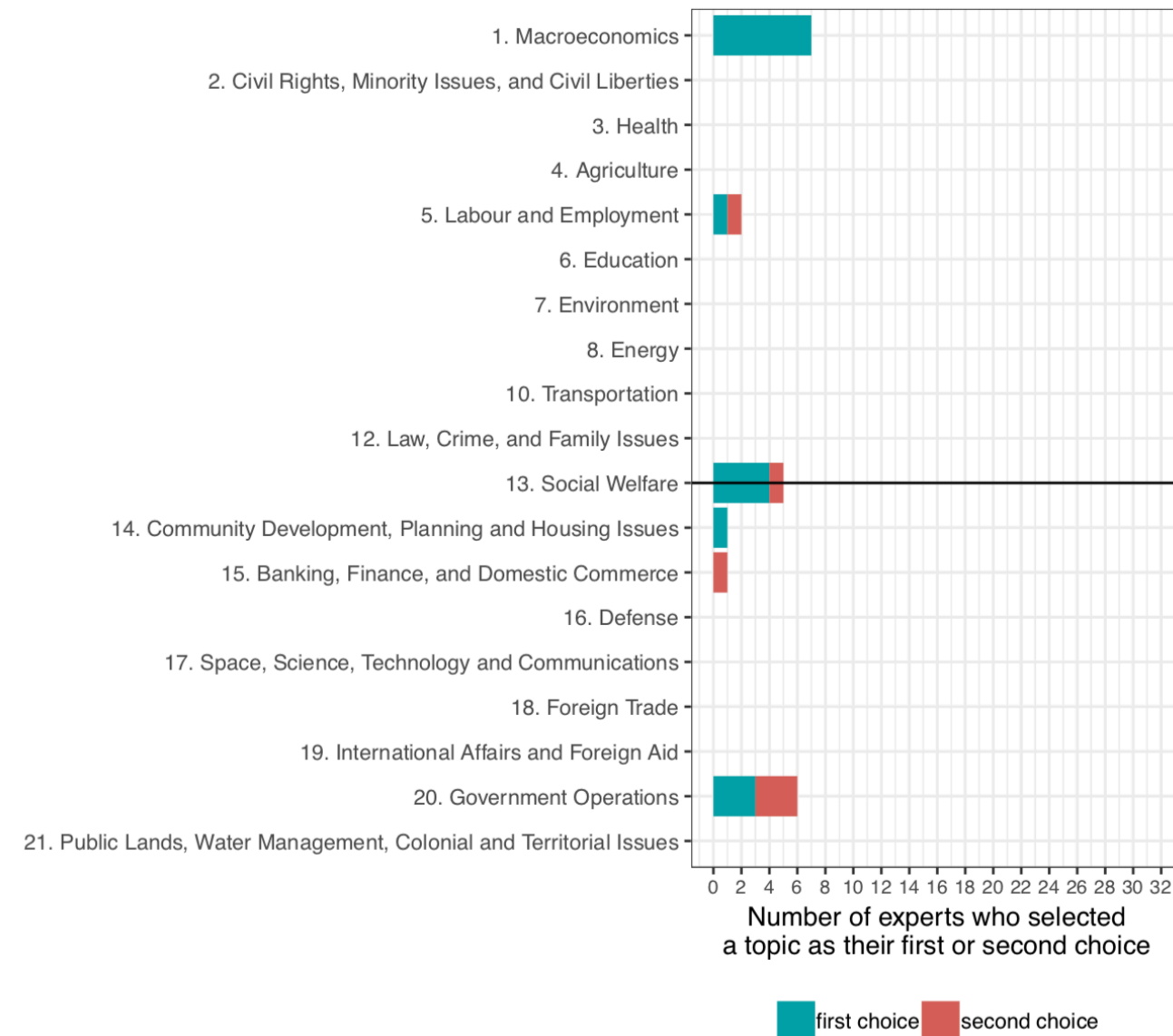




# SOTA(?)

Topic Label: Social Welfare

Top 20 words:"secretari, state, tax, chancellor, peopl, benefit, pension, exchequ, cut, war incom, social, hous, purchas, problem, profit, minist, compani, duti, govern"



Topic Label: Civil Rights, Minority Issues, and Civil Liberties

Top 20 words:"ireland, northern, countri, point, peopl, polic, war, hous, speech, great time, parti, irish, speaker, debat, issu, order, opposit, state, agreement"





# Lessons and next steps

- Cardinality and human bias are important in human topic labeling with consequences for downstream tasks
- Experts may not be SOTA (similar to CMP work)
- Cannot fully automate (yet)
- Next, annual rather than aggregated wordlists for topic label matching (without human evaluation  $79 \times 22 = 1,738!$ )

**Thank you!**

**s.mikhaylov@essex.ac.uk**