scelsi_2021_principled_analysis_of_energy_discourse_across_domains_with_thesaurus_based_automatic_topic_labeling

Year

2021

Author(s)

Scelsi, Thomas and Arranz, Alfonso Martinez and Frermann, Lea

Title

Principled Analysis of Energy Discourse across Domains with Thesaurus-based Automatic Topic Labeling

Venue

ALTA

Topic labeling

Fully automated

Focus

Primary

Type of contribution

Novel approach

Underlying technique

Thesaurus-based automatic labeling

Topic labeling parameters

Nr of retained label in the Thesaurus: 40

Nr of considered topic dimension (terms): 10

Glove embedding dimension: 50

Label generation

We propose a simple solution by leveraging structured, and broadly domain-relevant Thesauri as our label inventory.

Specifically, we use the EuroVoc thesaurus, compiled by the European Union which covers all areas of European Parliament discussion (including energy policy), noting that our methods extend to any thesaurus.

We propose methods for filtering the resource to a focused set of labels; and mapping induced topics to one or more thesaurus labels.

Formally, we describe the set of EuroVoc labels as L. Each label $l \in L$ represents a set of keyphrases (consist of one or more tokens. e.g. *mining industry*) v in the EuroVoc thesaurus that fall under that label

Renewable Energy: bioenergy, biogas, geothermal energy, marine energy, renewable energy, soft energy, solar energy, wind energy
Prices: reduced price, price index, price reduction, farm prices, world market price, target price, producer price, price list, price increase
Environmental Policy: nature reserve, waste recycling, industrial hazard, environmental tax, emission allowance, environmental impact

Table 1: Selected EuroVoc labels (bold) and some of their associated keyphrases.

Our method consists of two steps:

1. thesaurus filtering, in order to retain only domain-relevant labels (see “Label filtering” step for

more details on the exact curation)

2. an algorithm to map a topic (represented as a weighted list of words) to one or more labels (each represented as an unweighted set of associated phrases).

Topic Labeling

Given a DTM topic k represented as kˆ, we want to assign the top N EuroVoc labels that best match the topics content. We approach the automatic labelling task in two ways.
The first is a match-based approach and the second uses word embeddings to label topics.

Importance-Based Topic Labeling

Intuitively, a label is relevant to a learnt topic if

- it contains the topic's most relevant terms; and
- these keyphrases are unique to the label, and do not occur widely across EuroVoc labels (" keyphrase uniqueness").

If a term occurs in many labels, it is often less informative as it loses the ability to distinguish labels. We quantify term-topic relevance as kˆ[w] the probability of w in the re-normalized topic representation; and keyphrase uniqueness as TFIDF[w, l], the TFIDF value of w under l, where the documents are all EuroVoc labels.
The final topic-label score is:

$$\sigma_{k,l}^{imp} = \sum_{w \in \hat{k} \cap l} \hat{k}[w] \times TFIDF[w, l].$$

We define the intersection in the summation based on either a full or a sub-token match between topic term and label keyphrase (e.g., topic term *solar* would match label keyphrase *solar energy*).
The proposed method is fast and simple to implement, and requires no resources beyond the trained topics and thesaurus labels.
A disadvantage is its string-matching approach, which is oblivious to synonyms, or thematically related words.

Embedding-based Topic Labeling

The second approach makes use of pre-trained word embeddings.
At a high level we produce an aggregated representation of our top word vector kˆ as well as each EuroVoc label l in word vector space. We obtain a similarity score as the cosine similarity between the topic and label embeddings.
We use 50-dimensional pre-trained GloVe em-beddings.

We convert our top word vector kˆ into an embedding-based vector emb_k, by taking a *weighted* average of the GloVe embedding representations of each word in it, where each word embedding is weighted by the words topic relevance kˆ[w].
An embedding for label l, emb_l, is computed as an unweighted average over its keyphrases. Multi-token topic terms (or keyphrases in EuroVoc) are represented as an unweighted average over their token embeddings.

The relevance score σ_k^l for DTM topick and EuroVoc label l is then defined as the cosine similarity between their representations,

$$\sigma_{k,l}^{emb} = \text{cosine\_sim}(emb_k, emb_l).$$

We finally associate each topic with its top l≥1 associated labels as measured by either $\sigma_{k,l}^{\bar{imp}}$ or $\sigma_{k,l}^{emb}$.

Corpus	Topic terms	Embedding	TFIDF
Journals	market; price; electricity; paper; competition; company; investment; risk; reform; industry	1 business ops & trade 2 production	1 prices 2 business ops & trade
AEO	resource; oil; production; natural_gas; tight; gas; shale_gas; drilling; estimate; technology	1 oil & gas industry 2 renewable energy	1 oil & gas industry 2 production
IEO	projection ; energy ; eia ; model ; international ; outlook ; include ; analysis ; world ; case	1 economic analysis 2 research & ip	1 renewable energy 2 world organisations

Table 4: One topic from each of our corpora, with its top-2 EuroVoc labels as assigned by the embedding and tfidf-strategy, respectively.

ID	Top 10 Topic terms	Embedding	TFIDF
0	power; system; heat; generation;electricity; chp; energy; electric; district_heating; electrical	renewable energy (0.88); mechanical engineering (0.86); electronics and electrical engineering (0.84); technology and technical regulations (0.83)	electrical and nuclear industries (9.47); business operations and trade (4.88); renewable energy (3.67); organisation of transport (2.0)
4	emission; carbon; reduction; cost; ghg; greenhouse_gas; reduce; policy; result; country	deterioration of the environment (0.79); environmental policy (0.78); renewable energy (0.77); production (0.74)	environmental policy (15.19); accounting (2.61); deterioration of the environment (1.4); economic conditions (1.15)
22	china; carbon; reduction; sector; reduce; intensity; result; energy; increase	environmental policy (0.84); renewable energy (0.82); production (0.81); deterioration of the environment (0.81)	environmental policy (13.92); asia and oceania (3.85); renewable energy (2.15); economic conditions (1.71)
28	energy; energy_efficiency; building; system; paper; analysis; indicator; measure; present; energy_consumption	renewable energy (0.96); environmental policy (0.86); production (0.85); technology and technical regulations (0.85)	renewable energy (22.08); world organisations (5.76); electrical and nuclear industries (3.9); building and public works (1.86)
29	engine; fuel; emission; injection; diesel; co; combustion; high; low; increase	mechanical engineering (0.84); renewable energy (0.8); electrical and nuclear industries (0.8); oil and gas industry (0.8)	environmental policy (4.61); oil and gas industry (3.32); mechanical engineering (2.67); electrical and nuclear industries (1.28)

Table 5: Five example topics induced from the **Journals corpus**, with their top-4 EuroVoc labels (scores) as assigned by the Embedding and TFIDF-strategy, respectively.

ID	Top 10 Topic terms	Embedding	TFIDF
1	coal; ton; production; cost; percent; productivity; export; u.s; increase; region	oil and gas industry (0.89); coal and mining industries (0.88); production (0.81); renewable energy (0.77)	coal and mining industries (16.94); production (4.35); regions and regional policy (2.82); accounting (2.19)
17	gasoline; ethanol; gallon; fuel; mtbe; sulfur; blend; motor; percent; requirement	oil and gas industry (0.85); renewable energy (0.72); food technology (0.7); deterioration of the environment (0.69)	oil and gas industry (2.69); electrical and nu clear industries (0.81); taxation (0.35); organisation of transport (0.19)
19	vehicle; fuel; sale; percent; economy; new; increase; hybrid; car; standard	organisation of transport (0.88); production (0.88); prices (0.86); marketing (0.83)	economic conditions (8.02); organisation of transport (5.56); marketing (5.1); land transport (3.21)
21	emission; carbon; co; ton; metric; ghg; carbon_dioxide; energy; relate; percent	renewable energy (0.78); oil and gas industry (0.76); deterioration of the environment (0.74); electrical and nuclear industries (0.73)	environmental policy (11.52); renewable energy (2.28); deterioration of the environment (1.24); technology and technical regulations (1.21)
29	cost; market; electricity; price; competitive; customer; state; utility; transmission; power	prices (0.91); business operations and trade (0.91); production (0.9); accounting (0.9)	prices (26.58); business operations and trade (14.09); accounting (5.78); environmental policy (3.57)

Table 6: Five example topics induced from the **AEO corpus**, with their top-4 EuroVoc labels (scores) as assigned by the Embedding and TFIDF-strategy, respectively.

ID	Top 10 Topic terms	Embedding	TFIDF
5	coal; import; ton; export; increase; percent; world; project; trade; coke_coal	oil and gas industry (0.9); coal and mining industries (0.88); production (0.82); renewable energy (0.77)	coal and mining industries (15.34); business operations and trade (8.08); world organisations (1.28); deterioration of the environment (1.18)
6	natural_gas; cubic; foot; gas; lng; reserve; increase; percent; year; production	oil and gas industry (0.88); renewable energy (0.85); production (0.84); deterioration of the environment (0.82)	oil and gas industry (2.01); production (2.01); environmental policy (1.01); agricultural activity (0.99)
9	coal; china; world; percent; use; increase; consumption; share; total; btu	production (0.87); business operations and trade (0.83); oil and gas industry (0.83); prices (0.83)	coal and mining industries (11.93); business operations and trade (4.49); asia and oceania (2.47); world organisations (2.1)
25	emission; sulfur; reduce; reduction; standard; fuel; new; require; target; dioxide	deterioration of the environment (0.82); renewable energy (0.8); environmental policy (0.8); electrical and nuclear industries (0.78)	environmental policy (10.89); technology and technical regulations (2.3); oil and gas industry (1.43); asia and oceania (0.94)
27	generation; natural_gas; renewable; nuclear; capacity; electricity; cost; increase; coal; power	renewable energy (0.91); electrical and nuclear industries (0.88); production (0.87); oil and gas industry (0.84)	electrical and nuclear industries (8.21); coal and mining industries (2.97); accounting (2.36); demography and population (1.75)

Table 7: Five example topics induced from the **IEO corpus**, with their top-4 EuroVoc labels (scores) as assigned by the Embedding and TFIDF-strategy, respectively.

Motivation

Studies are often difficult to reproduce, compare against or build upon due to a lack of public resources, as well as ad-hoc subjective choices of the researchers.
We address the latter problem by proposing a conceptually simple and theoretically sound method for automatic topic labelling, draw- ing from thesauri over the political domain.

We also leverage our automatic labelling to un- cover change within a topic over time (We create a normalised representation for a topic through its top 10 most probable words at each timestep).

Our automatic topic labels allow us to identify differences in discussion between publications (We combine topics from a model that have the same top-1 automatically assigned label by summing their proportions over time. We sum again over all topics that have the same top-1 label assignment to achieve an overall proportion for the top-1 label at timestep t).

Topic modeling

DTM

Topic modeling parameters

Nr of topics: 30

Nr. of topics

90 (30 per corpus)

Label

One or more (Top-1 or Top-4) thesaurus labels taken from the EuroVoc thesaurus which contains 127 general “topics”, each associated with a list of phrases.

Label selection

\

Label quality evaluation

We evaluated our thesaurus-based topic labeling approach through human judgements.

We presented annotators with DTM-induced topics, together with three label options: one based on the TFIDF mapping, one based on the Embedding-based mapping, and a randomly selected label. Annotators were asked to select the most appropriate label in a forced-choice paradigm.
Given that we want to compare what label best represents a topic when our strategies differ, we do not include topics in the task where the embedding and TFIDF labels are automatically assigned the same label. Over our three models, this occurs for 23 topics.
We evaluated two ver-sions of our strategy: one where we paired each topic k with the single most highly associated label l in terms of labeling score $\sigma^*_{l,k}$ (top-1); and a second where we associate topics with their four most associated labels, capturing a mixture of information (top-4).

Each annotation task consisted of a random sample of 30 out of a total of 90 induced topics (30 per corpus).

We collected 20 sets of annotations for the top-4 strategy, and 16 sets for the top-1.
Table 3 summarizes the human preferences.

Strategy	TFIDF	Embedding	Baseline
Top-1	0.46	0.47	0.07
Top-4	0.45	0.47	0.08

Table 3: Human preferences (%) of automatic topic labeling methods when considering the top-1 or top-4 predicted labels by our methods or a random baseline.

We can see that both our strategies significantly outperform the random baseline from filtered topics.
In both the top-1 and top-4 strategy we see no difference between annotator preference toward either the TFIDF or embedding labelling strategy.
The same pattern holds for each individual corpus.

The user study shows that non-experts can discern meaningful labels from our method.

Assessors

We obtained annotations from a group of 36 annotators who are proficient English speakers. All but one annotator were not domain experts.

Domain

Paper: Energy

Dataset: Energy

Problem statement

First, we release two diachronic corpora covering 23 years of energy discussions in the U.S. Energy Information Administration.
Secondly, we propose a simple method for automatic topic labelling drawing on domain knowledge via political thesauri.

We empirically evaluate the quality of our labels, and apply our labelling to topics induced by diachronic topic models on our energy corpora, and present a detailed analysis.

Corpus

EIA Corpus and AEO corpus

Origin: US Energy Information Administration (EIA)

Nr. of documents: 4320

Details:

- *Annual Energy Outlook* (AEO) and *International Energy Outlook* (IEO) are mandated to provide US citizens and lawmakers with future-oriented evaluations of, respectively, domestic and international energy trends
- all IEO and AEO releases between 1997–2020

The Journals corpus

Origin: *Applied Energy* and *Energy Policy* Journals

Nr. of documents: 24353

Details:

- article abstracts published in these two journals for the period 1997-2020

Document

- EIA report into header-paragraph pairs
- Journal abstract

Pre-processing

EIA Corpus

We automatically split each parsed EIA report into header-paragraph pairs, which are then used as documents to train our topic models.

The Journals corpus

The format is already machine-readable and contains metadata on publication date requiring only minimal data cleaning

- tokenisation

```
@inproceedings{scelsi_2021_principled_analysis_of_energy_discourse_across_domains_with_thesaurus_based_automatic_topic_labeling,
    title = "Principled Analysis of Energy Discourse across Domains with Thesaurus-based Automatic Topic Labeling",
    author = "Scelsi, Thomas and Arranz, Alfonso Martinez and Frermann, Lea",
    booktitle = "Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association",
    month = dec,
    year = "2021",
    address = "Online",
    publisher = "Australasian Language Technology Association",
    url = "https://aclanthology.org/2021.alta-1.11",
    pages = "107--118",
    abstract = "With the increasing impact of Natural Language Processing tools like topic models in social science research, the experimental rigor and comparability of models and datasets has come under scrutiny. Especially when contributing to research on topics with worldwide impacts like energy policy, objective analyses and reliable datasets are necessary. We contribute toward this goal in two ways: first, we release two diachronic corpora covering 23 years of energy discussions in the U.S. Energy Information Administration. Secondly, we propose a simple and theoretically sound method for automatic topic labelling drawing on political thesauri. We empirically evaluate the quality of our labels, and apply our labelling to topics induced by diachronic topic models on our energy corpora, and present a detailed analysis.",
}
```

#Thesis/Papers/FS