23_11_2022

A brief overview on the impact of the proximity constraint on the selected papers

The choice of introducing a proximity constraint between the root terms `topic*` and `label*` on the original query results has been justified by the desire of filtering out those papers containing the root terms `topic model*` and `label*` that were unlikely to carry information that would be useful in the context of this review.
The reasoning behind this choice is that if the root term `label*` appears in isolation (i.e. not in the vicinity of `topic*`) it is likely not being used to describe a topic labeling activity.
A few examples of such cases are highlighted here:

- Wu et al., 2019 proposes an approach to model opinion targets a their sentiment. In this context, the `label*` root term is often used to refer to the assignment of opinion targets to snippet of text:
  - "It's labor-intensive to manually **label** opinion targets in each domain"
  - "… we propose an unsupervised method to identify opinion targets automatically, which solves the difficulty of **labeling** opinion targets manually in different domains"
- Demszky et al., 2019 proposes an NLP framework to analyse political polarisation in social media. In this case, the root term `label*` can be seen used to describe the activity of attaching a political affiliation to a given user:
  - "We begin by quantifying polarization […] between the language of users **labeled** Democrats and Republicans"
  - "We **label** a user as a Democrat if they followed more Democratic than Republican politi- cians in November 2017"

Adding a proximity constraint generally allows to filter out these unwanted instances by only flagging sentences (within documents) where the two terms are used in near conjunction with one another. Some examples of such sentences are provided below:

- "…, **topics** can be more readily interpretable when they are assigned semantically meaningful **labels** ." - Marani et al., 2022
- "… various methods have been proposed to assign concise **labels** to **topics** to improve interpretability." - Zosa et al., 2022
- "An interpretable **topic** is one that can be easily **labeled**. How easily a **topic** could be **labeled**…" - Doogan & Buntine, 2021

The choice of using 20 terms as a (somewhat broader) constraint can be justified by the information found in Griffies et al., 2020 which states that:

"The average length of sentences in scientific writing is only about **12-17 words**.".
In this context, the chosen proximity constraint should generally be able to account for paper containing instances of the two root terms appearing in the same sentence.

Another potentially interesting insight can be provided by observing the influence of stricter proximity constraints over the set of selected papers.

Papers discarded by imposing stricter constraints

20 → 5

- KDD, "CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring"
- EMNLP, "Adapting Topic Models using Lexical Associations with Tree Priors"

5 → 3

- Expert Systems with Applications, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis"
- SIGIR, "Read what you need: Controllable Aspect-based Opinion Summarization of Tourist Reviews"

By moving the proximity constraint down to a maximum distance of three terms (from the original value of 20) it is possible to observe how a total of 4 out of 65 papers (~6%) would be excluded.

- KDD, "CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring"
  - Proposes a method to perform seed-guided topical taxonomy construction. Within the constructed taxonomy, each node (topic) is represented by a **concept name** and a cluster of coherent terms. Unlike the topically similar paper by Zhang et al., 2018 (also published in KDD), this paper does not directly refer to concept names as topic labels. The sentence allowing for the retrieval of this papers is as follows:
    - "Then we use majority votes to **label** the pairs and use all the true parent-children pairs from different methods to construct a gold standard taxonomy. Since each **topic** is represented by a cluster of words, …"
- EMNLP, "Adapting Topic Models using Lexical Associations with Tree Priors"
  - Proposes the use of tree priors (using tree LDA) to improve interpretability of topics. In this context, **concepts** appearing in multi levelled tree priors can be used to assign topic names. Even though the paper does not refer to single word topic identifiers as "labels", the paper is retrieved due to the following

sentence:

- "All the results are averaged across five-fold cross-validation using 20 **topics** with hyper-parameters α = β = 0.01. For 20NewsGroups classification, a post's newsgroup is its **label**."

- Expert Systems with Applications, "Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis"
  - Proposes an LSA model (based on Word2vec and Spherical k-means clustering) to perform a trend analysis on blockchain research. In this context: "the name of the cluster is defined by considering the characteristics of the words assigned to the cluster, and it is considered as a topic.". Notice that in this paper the **word cluster** (or topic) name is not referred to as a label. Instead, the word "topic" is used to refer to both the identifier and the terms contained in the cluster. The paper is retrieved due to the following sentence:
    - "For measuring the accuracy of allocated **topics** to the documents, existing studies have used the data for text classification, which was already categorized or assigned to the **topic**. Since there are no exact **labels** for our data, we propose a quantitative evaluation method"
- SIGIR, "Read what you need: Controllable Aspect-based Opinion Summarization of Tourist Reviews"
  - The manual labeling step is captured by the laxer proximity constraint thanks to the following sentence:
    - "each extracted **topic** is manually interpreted by looking at its representative words and assigned a genuine aspect **label**."

With the exception of one publication, a common characteristic of the papers that would be excluded by a stricter proximity constraint resides in the fact that they do not refer to topic identifier as "labels". In fact, the terms "concept name", "concept" or simply "topic" are used instead.
Despite this fact, the highlighted work is kept in the final selection due to its relevance with regards to the proposed survey.

Backward snowballing (Outgoing references), 2017 onwards

In order to further extend the selection of examined work starting from the set of the 65 selected papers, backward snowballing is performed.
Backward snowballing refers to the activity of extracting relevant work from the references appearing in papers that have already been deemed relevant. In other words, this means finding yet undiscovered (cited) work that should be included in the final review.
This activity of backward snowballing can be summarised in the following steps:

1. Go through the selected list of papers and, for each document, extract the list of references.
2. Filter the selected references with regards to the imposed time constraints.
3. Gather the initial set of publications resulting from this reference extraction process.
4. Apply the previously constructed search query to filter down the gathered publications.
5. Inspect the content of each remaining document and determine its relevance by applying the same set of inclusion/exclusion criteria utilised for the main selection.

Initial results

For the purpose of this activity, it has been decided to apply the same time-related constraints that have been used for the main selection (i.e. work released from 2017 onwards). On the other hand, no limitation on the origin venue (journal / conference) of the extracted work has been applied. Using this selection criteria, an initial corpus of **735 items** is extracted.

No filtering is applied with regards to the venue to which the extracted work belongs to. This means that forward snowballing allows to extend the scope of this review beyond the set of initially selected journals and conferences.

On this broad selection, the previously described query containing the 20 terms proximity operator (`"topic label*" OR ("topic model*" AND ("label*" NEAR "topic*"))`) is applied.

Side note: This query is applied on the set of locally stored papers using the FoxTrot Professional Search tool.

The corpus resulting from this filtering steps returns a total of **157 items** (228 if the proximity constraint is removed).

B.S. selection (ongoing, currently analysed 76 out of 157, selected 17 out of 76)

All papers contained in the corpus resulting from the initial reference extraction (and filtering) phase are individually inspected and evaluated against the established selection criteria.

From the initial set of 157 papers, a final selection of […] was found to meet the imposed requirements and therefore added to the systematic review.

The set of publications, together with their host venue, is presented below:

- Evaluating Visual Representations for Topic Understanding and Their Effects on Manually Generated Topic Labels, Transactions of the Association for Computational Linguistics
- Evaluating Topic Representations for Exploring Document Collections, Journal of the Association for Information Science and Technology
- United We Stand: Using Multiple Strategies for Topic Labeling, Natural Language

Processing and Information Systems
- Transfer Topic Labeling with Domain-Specific Knowledge Base: An Analysis of UK House of Commons Speeches 1935–2014, Research & Politics
- In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics, The Journal of Machine Learning Research
- Using structural topic modeling to identify latent topics and trends in aviation incident reports, Transportation Research Part C
- An Ontology-Based Labeling of Influential Topics Using Topic Network Analysis, Journal of Information Processing Systems
- Global Surveillance of COVID-19 by mining news media using a multi-source dynamic embedded topic model, BCB
- Free associations of citizens and scientists with economic and green growth: A computational-linguistics analysis, Ecological Economics
- Recommendation System for Knowledge Acquisition in MOOCs Ecosystems, SBSI
- Managing the Boundaries of Taste: Culture, Valuation, and Computational Social Science, Social Forces
- Scientific Evolutionary Pathways: Identifying and Visualizing Relationships for Scientific Topics, Journal of the Association for Information Science and Technology
- Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation, DSAA
- Topic modeling and sentiment analysis of global climate change tweets, Social Network Analysis and Mining
- Applying LDA topic modeling in communication research: Toward a valid and reliable methodology, Communication Methods and Measures
- 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
- Narratives of the Refugee Crisis: A Comparative Study of Mainstream-Media and Twitter, Media and Communication

Additionally, the following 10 papers that were already part of the initial selection of publications were retrieved once again following the backward snowballing procedure:
- Multimodal Topic Labelling, EACL
- Topic Model or Topic Twaddle? Re-evaluating Semantic Interpretability Measures, NAACL
- Automatic Generation of Topic Labels, SIGIR
- BART-TL: Weakly-Supervised Topic Label Generation, EACL
- Labeling Topics with Images using a Neural Network, ECIR
- Neural Models for Documents with Metadata, ACL

- W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis, Expert Systems With Applications
- Document-based topic coherence measures for news media text, Expert Systems with Applications
- Detecting and predicting the topic change of Knowledge-based Systems: A topic-based bibliometric analysis from 1991 to 2016, Knowledge-Based Systems
- TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering, KDD

B.S. venues (ongoing)

The following is the list of previously unexplored venues to which the work selected by the backward snowballing task belongs to:

Conferences

- ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB)
- IEEE International Conference on Data Science and Advanced Analytics (DSAA)
- Natural Language Processing and Information Systems (NLDB)
- Sustainable Business and Social Impact Conference (SBSI)

Journals

- Communication Methods and Measures
- Ecological Economics
- Journal of Information Processing Systems
- Journal of the Association for Information Science and Technology
- Media and Communication
- Political Science Research and Methods
- Research & Politics
- Social Forces
- Social Network Analysis and Mining
- The Journal of Machine Learning Research
- Transactions of the Association for Computational Linguistics
- Transportation Research Part C

Next steps

- Finish the selection for backward snowballing
- Perform forward snowballing

- Depending on the size of the final selection (after snowballing)
  - Increase/Decrease the time-frame / selected venues
- Build paper graphs using Pajek
- Start to rewrite the notes into the “Methods” section (i.e. transcribe the work done so far on Overleaf)
- After December 11th, gather papers from emnlp 2022
- Establish details of data collection process

#Thesis/Temporary notes#