Topic Modeling on Podcast Short-Text Metadata

44th European Conference on Information Retrieval

Francisco B. Valero, Marion Baranes, and Elena V. Epure

Deezer Research, 22-26, rue de Calais, 75009 Paris, France
research@deezer.com

April 12, 2022

Table of contents

1. Introduction
2. Related work
3. Methods
4. Datasets
5. Experiments
6. Results and Discussion
7. Conclusions

Introduction

Context

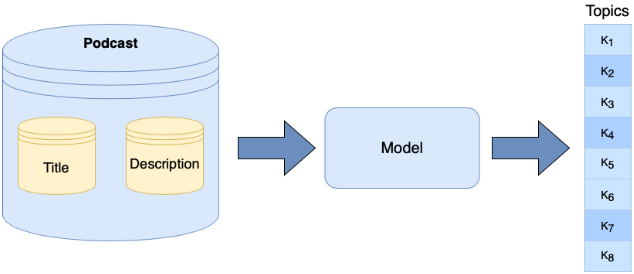
- **Podcasts**
  - "Spoken" version of the blog posts (audio content)
  - Massively popularised in the recent years
- **Topics** very useful for
  - Categorization
  - Retrieval
  - Recommendation

Challenges

- Topic annotation still quite problematic
- **Manual annotation** (curators, creators ...)
  - Broad, noisy, or unreliable topics as podcast genres
- **Automatic annotation** (data limitations):
  - Speech transcription is expensive and with high WER for NEs
  - Textual metadata (titles and descriptions) is a short text

Objective

- Revisit the **feasibility** of discovering **relevant** topics from podcast metadata, titles and descriptions.
  - Economic alternative
  - Categories at different granularity levels



Topic modeling on short text

- **Data sparsity** challenge
  - Topic-related words rarely co-occur in the same context
  - Ambiguity, noise, limited context
  - Conventional topic modeling techniques such as LDA unsuitable
- But there is recent advancement of topic modeling techniques on short text with good results

Contributions

1. The most extensive benchmark of short-text topic modeling techniques on podcast metadata
2. **NEiCE**
   - NE-informed Corpus Embedding for NMF-based topic modeling
   - Injecting NEs cues largely improves SOTA topic coherence results
3. A new podcast metadata corpus, the largest in terms of shows

Deezer's podcast example

Title: Shields Up! Podcast
Description: *Join Chris* & *Nev* as they talk about their favourite *Star Trek* episodes covering everything from *TOS* to *Lower Decks*.

Related work

Topic modeling on short text

- Pseudo-documents-based
  - Aggregate connected short texts in longer documents
  - Apply conventional topic modeling
- Probabilistic
- Neural
- NMF-based

Models overview

GPU-DMM (Li et al., 2016)

- Sampling process to promote topic-related words
- Word association estimated by exploiting pre-trained word embeddings

NQTM (Wu et al., 2020)

- Encoder generates peakier distributions (quantification)
- Decoder uses negative sampling for discovering non-repetitive topics

SeaNMF (Shi et al., 2018)

- Adjust NMF to integrate word-context semantic correlations

CluWords (Viegas et al., 2019)

- Enhance corpus representation before applying NMF
- Custom TF-IDF strategy exploiting pre-trained word embeddings

Methods

Intuition

- Leverage NEs in a NMF framework (CluWords)
  - High frequency of NEs in podcast metadata
  - NEs convey topic information

Example

"That **Peter Crouch** Podcast" is related to football or sport

- Why NMF-based topic modeling?
  - Better results on short text
  - NEs' integration more straightforward than in deep neural networks

Preliminaries: CluWords (Viegas et al., 2019)

- NMF-based topic modeling
- Novel document representation for term-document matrix (A)
  - Leverages pre-trained embeddings to overcome data sparsity
  - Inspired by TF-IDF (discriminant words » popular words)

$$\text{tf\_idf}(t, d) = \text{tf}(t, d) \cdot \log\left(\frac{|\mathcal{D}|}{n_t}\right) \tag{1}$$

  - where $\text{tf}(t, d)$ is the number of times t appears in document d and n_t is the number of documents in corpus $\mathcal{D}$ where t appears

Preliminaries: CluWords (Viegas et al., 2019)

1. Compute matrix C where $C_{t,t'}$ is the cosine similarity (cos) of the embeddings corresponding to the pair of terms $t, t' \in \mathcal{V}$.
   - α^{word} used to select the most similar term pairs

$$C_{t,t'} = \begin{cases} \cos(v_t, v_t') & \text{if } \cos(v_t, v_t') > \alpha^{word} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

2. Compute TF-IDF over vector-based term representations instead of individual frequencies.
   - t replaced by $C_{t,:}$ in order to expand the term's context

Preliminaries: CluWords (Viegas et al., 2019)

- CluWords term-document matrix:

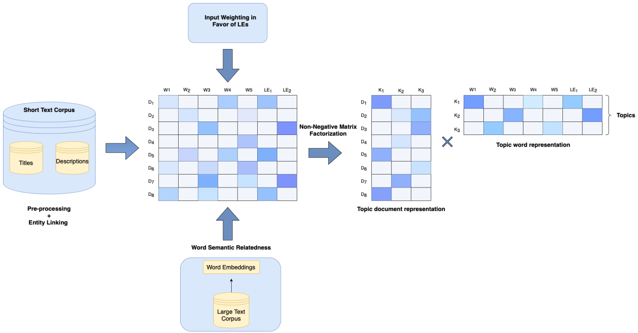
$$A^*_{d,t} = \text{tf}^*(d,t) \cdot \text{idf}^*(t) = (AC)_{d,t} \cdot \log\left(\frac{|\mathcal{D}|}{\sum_{d \in \mathcal{D}} \mu(t,d)}\right) \quad (3)$$

- $\mu(t,d)$ is the mean cosine similarity between the term t and its semantically related terms t' in document d denoted $\mathcal{V}^{d,t} = \{t' \in d | C_{t,t'} \neq 0\}$

$$\mu(t,d) = \begin{cases} \frac{1}{|\mathcal{V}^{d,t}|} \cdot \sum_{t' \in \mathcal{V}^{d,t}} C_{t,t'} & \text{if } |\mathcal{V}^{d,t}| > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

NE-informed Corpus Embedding (NEiCE)

- A new corpus representation matrix A^{NE} leveraging NEs
- Based on a **preprocessing** step and a **computation** step



Preprocessing step

- NE linking using REL (van Hulst et al., 2020)
  - Identify NE mentions in podcast textual metadata
  - Link NE mentions to Wikipedia entities
- Consider as single words NE mentions whose confidence is low
  - Exclude from these common names (e.g. *Steve*, *Anna*, *France* ...) (NameDataset[1])
- Leverage Wikipedia2Vec (Yamada et al., 2018) word and entities embeddings (for C)

[1] https://github.com/philipperemy/name-dataset

Computation step

- Consider NEs without including them in the vocabulary
- Boost NEs importance by re-weighting their semantically-related words

$$\mathrm{tf}^{NE}_{d,t} = \begin{cases} (AC)_{d,t} + \max_{t' \in \mathcal{V}^{d,t}} (AC)_{d,t'} & \text{, if } t \in \mathcal{E}^e \text{, } e \text{ in } d \text{ and } |\mathcal{V}^{d,t}| > 0 \\ (AC)_{d,t} & \text{otherwise} \end{cases} \tag{5}$$

$\mathcal{E}^e = \{t | \cos(v_e, v_t) \geq \alpha^{ent}, \forall t \in \mathcal{V} - \mathcal{E}\}$ is the set of non-NE words from $\mathcal{V}$ most similar to a NE e.

Datasets

Statistics

- Deezer is the largest podcast dataset in terms of number of shows
- Large number of podcasts with NE mentions in all datasets

Dataset	$\|\mathcal{D}\|$	$\|\mathcal{V}\|$	#NE mentions	#podc. with NE	#w/title	#w/descr.
Spotify	17 456	7 336	20 885	9 198	3.5	38.2
iTunes	9 859	7 331	24 973	6 994	4.9	56.4
Deezer	**29 539**	**14 322**	**67 083**	**19 969**	4.0	62.6

Table: Summary of the podcast datasets: the number of podcasts, the vocabulary size, the total number of NE mentions, the total number of podcasts with NEs in metadata, the mean number of words per title, and the mean number of words per description.

Experiments

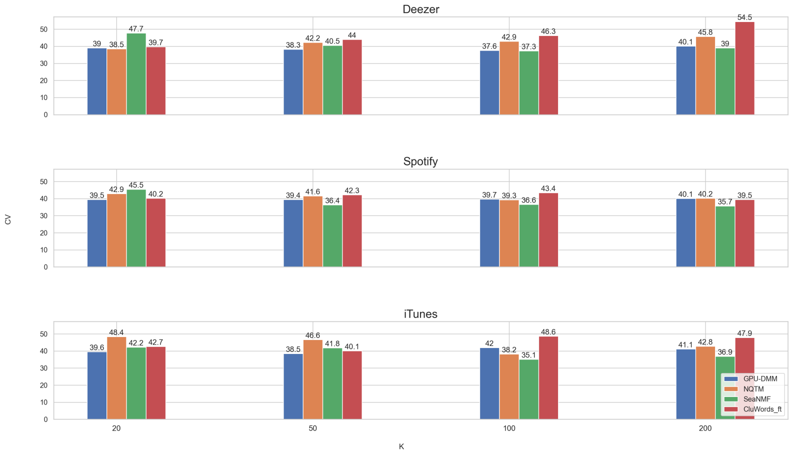
Experimental setup and environment

- Evaluation metric: topic coherence (C_V) Röder et al. (2015)
  - Correlates best with human judgement of topic ranking
- Number of top words T: 10
- Number of topics K: 20, 50, 100 and 200
- α^{word} and α^{ent} in NEiCE: 0.2, 0.3, 0.4, 0.5
- Default hyper-parameters for the baselines
- Environment
  - Intel Xeon Gold 6134 CPU @ 3.20GHz with 32 cores and 128GB RAM

Results and Discussion

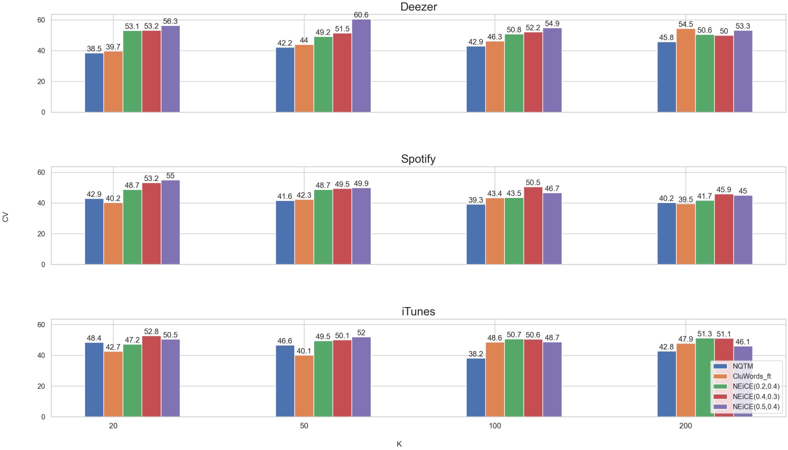
Topic coherence scores obtained by baselines

- NMF-based methods obtain the best scores
- CluWords ranking first in most cases (7/12)



Topic coherence scores obtained by NEiCE

- NEiCE obtains larger coherence scores than the baselines in most cases
- The introduction of NE cues has a positive impact, no matter the choice of α^{word} and α^{ent}



Examples

k	NEiCE	NQTM
1	mindfulness, yoga, meditation, psychotherapy, psychotherapist, hypnotherapy, psychoanalysis, hypnosis, therapist, psychology	psychotherapist, beirut, displays, remixes, weddings, adversity, namaste, kimberly, agenda, introducing
2	fiction, nonfiction, novel, author, book, novelist, horror, cyberpunk, anthology, fantasy	avenues, werewolf, criminal, pure, imaginative, strategies, demand, agree, oldies, hang
3	republican, senator, senate, libertarian, election, candidate, nonpartisan, conservative, caucus, liberal	hour, sudden, key, genres, keeps, round, neighbor, conservatives, realize, fulfillment

Table: Topics obtained with NEiCE or NQTM on Deezer and $K = 50$.

Conclusions

Conclusions

- Detailed study of topic modeling on podcast metadata
- Release the largest podcast metadata dataset[2]
- Propose NEiCE, a new NE-informed document representation exploited in a NMF framework
- Take into account NEs helps to be more effective in terms of topic coherence than the baselines in various evaluation scenarios
- Future work: conduct expert studies with editors to further validate mined topics in order to find best NEiCE configuration

[2] https://zenodo.org/record/5834061.YkBGuC8lO_z

Topic Modeling on Podcast Short-Text Metadata

44th European Conference on Information Retrieval

Francisco B. Valero, Marion Baranes, and Elena V. Epure

Deezer Research, 22-26, rue de Calais, 75009 Paris, France
research@deezer.com

April 12, 2022

References I

Li, C., Wang, H., Zhang, Z., Sun, A., and Ma, Z. (2016). Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Shi, T., Kang, K., Choo, J., and Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1105–1114.

van Hulst, J. M., Hasibi, F., Dercksen, K., Balog, K., and de Vries, A. P. (2020). Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM.

References II

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: exploiting semantic word clustering representation for enhanced topic modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 753–761.

Wu, X., Li, C., Zhu, Y., and Miao, Y. (2020). Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. (2018). Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. *arXiv preprint arXiv:1812.06280*.