IDIAP RESEARCH REPORT

BROADCAST MEDIA CONTENT CATEGORIZATION USING LOW-RESOLUTION CONCEPTS

Esaú VILLATORO-TELLO[a] Shantipriya Parida[b]
Petr Motlicek Subhadeep Dey Qingran Zhan

Idiap-RR-06-2021

Version of JULY 07, 2021

[a]Idiap
[b]Idiap Research Institute

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Broadcast Media Content Categorization Using Low-Resolution Concepts*

Esa Villatoro-Tello[1,3], **Shantipriya Parida**[1], **Petr Motlicek**[1], **Subhadeep Dey**[1], and **Qingran Zhan**[1]

[1]Idiap Research Institute, Martigny, Switzerland.
{firstname.lastname}@idiap.ch
[2]Ecole Polytechnique Fdrale de Lausanne, Switzerland.
mael.fabien@epfl.ch
[3]Universidad Autnoma Metropolitana, Unidad Cuajimalpa, Mexico City, Mexico.
evillatoro@correo.cua.uam.mx

Abstract

Short text clustering is a challenging problem due to its sparseness of text representation. This paper proposes to employ a low-resolution representation for accurately categorizing German broadcast media content. Our proposed approach guarantees document clusters using a highly dense representation, denominated low-resolution concepts. We first identify the fundamental semantic elements in the document collection, subsequently used to build the low-resolution texts representation, which serves as input to a k-means clustering process. We performed experiments using a dataset from a German TV channel. Results demonstrate that using low-resolution concepts for representing the broadcast media content allows obtaining a relative improvement of 70.4% in terms of the Silhouette coefficient compared to deep neural architectures.

1 Introduction

Current broadcast platforms utilize the Internet as a cross-promotion source, thus, their produced materials tend to be very short and thematically diverse. Besides, modern Web technologies allow the rapid distribution of these informative content through several platforms. As a result, the broadcast media content monitoring represents a challenging scenario for current Natural Language Understanding (NLU) approaches to efficiently exploit this type of data due to a lack of structuring and reliable information associated with these contents (Morchid and Linarès, 2013; Doulaty et al., 2016; Staykovski et al., 2019). Furthermore, if we consider that documents are very short (a few sentences long) and that they come from a very narrow domain, the task of clustering becomes harder.

Traditionally, the Bag-of-Words (BoW) has recently been the most widely used text representation technique for solving many text-related tasks, including document clustering, due to its simplicity and efficiency (Ribeiro-Neto and Baeza-Yates, 1999). However, the BoW has two major drawbacks: *i)* document representation is generated in a very high-dimensional space, *ii)* it is not feasible to determine the semantic similarity between words. As widely known, previous problems increase when documents are short texts (Li et al., 2016). It becomes more difficult to statistically evaluate the relevance of words given that most of the words have low-frequency occurrences, the BoW representation from short-texts results in a higher sparse vector, and the distance between similar documents is not very different than the distance between more dissimilar documents.

To overcome some of the BoW deficiencies, semantic analysis (SA) techniques attempt to interpret the meaning of the words and text fragments by calculating their relationship with a set of predefined concepts or topics (Li et al., 2011). Examples of SA techniques are LDA (Blei et al., 2003), LSA (Deerwester et al., 1990), and word embeddings (Le and Mikolov, 2014). Accordingly, these strategies learn word or document representations based on the combination of the underlying semantics in a dataset. Similarly, more recent approaches, with the help of word embeddings, learn text representations using deep neural network architectures for document classification (De Boom et al., 2016; Adhikari et al., 2019; Ostendorff et al., 2019; Sheri et al., 2019). However, most of these approaches focus either on solving supervised classification tasks or clustering formal-written short documents.

In this paper, we propose an efficient technological solution for the unsupervised categorization of broadcast media content. Our proposed approach generates document clusters using a highly dense

*The final and extended version of this report was published at https://ufal.mff.cuni.cz/pbml/115/art-villatoro-tello-et-al.pdf

representation, called low-resolution concepts. We first identify the fundamental semantic elements (i.e., concepts) in the document collection, then, these are used to build the low-resolution representation, which is later used in a clustering process.

The main contributions of this paper are summarized as follows: *(i)* to the best of our knowledge, this is the first attempt to explore the feasibility and effectiveness of the low-resolution bag-of-concepts in solving one unsupervised task, broadcast media content categorization; and, *(ii)* we conducted our experiments on a real-life dataset of German spoken documents, results demonstrate that the proposed methodology achieves good performance in terms of the Silhouette score, to be considered for practical deployment.

The remainder of the paper is organized as follows: in Section 2 we describe the proposed methodology. In Section 3 we provide some details regarding the employed dataset. Experimental results and analyses are presented in Sections 4 and 5. Finally, in Section 6 we draw our main conclusions and future work directions.

2 Proposed Method

Inspired by the work of (López-Monroy et al., 2018), we propose using a highly dense representation, denominated low-resolution concepts, for solving the task of clustering short transcript-texts, i.e., broadcast media documents. The intuition behind this approach is that highly abstract semantic elements (concepts) are good discriminators for clustering very short transcript texts that come from a narrow domain. The proposed methodology is depicted in Figure 1. Generally speaking, we first identify the underlying concepts contained in the dataset. For this, we can employ any semantic analysis (SA) approach for learning words representation; thus, learned representation allows us to generate sets of semantically associated words. After obtaining the main concepts, documents are represented by a condensed vector, which counts for the occurrences of the concepts, i.e., a concept distribution vector. Finally, the build texts representation serves as the input to a clustering process, in this case, the K-means algorithm.

More formally, let $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ denote the set of short transcript texts, and let $\mathcal{V} = \{w_1, w_2, \ldots, w_m\}$ represent the vocabulary of the document collection $\mathcal{D}$. As first step, we aim at inferring the underlying set of concepts $\mathcal{C} = \{c_1, c_2, \ldots, c_p\}$ contained in $\mathcal{D}$, where every $c_l \in \mathcal{C}$ is a set formed by semantically related words. Notice that in order to obtain the concepts $\mathcal{C}$ we can apply any SA technique for learning the vector representation $\mathbf{v_i}$ of each word $w_i \in \mathcal{V}$, for example LDA, LSA, or word embeddings. Next, for obtaining the document $\mathbf{d_j}$ representation, we account for the occurrence of each c_l within d_i, in other words, the document vector $\mathbf{d_j}$ is a vector that contains concepts distribution. Finally, the generated document-concepts matrix $\mathbf{M}_{\mathcal{D}\times\mathcal{C}}$ serves as the input to a clustering process aiming at finding the more suitable documents groups according to the concept-based representation.

The proposed method has two main parameters, the resolution parameter (p) and the group parameter (k). The former, p, represents the number of concepts that will be generated from the SA step. The lower the number of concepts, the more abstract the resolution. The second parameter, k, indicates the number of categories to be generated from the clustering process. Given the nature of the dataset, i.e., very short texts from a narrow domain, we hypothesize that the clustering algorithm will be able to find groups of documents that share the same amount of information about the same sub-set of concepts, resulting in a more coherent categorization of the documents. Thus, using low-resolution concepts will generate groups of documents referring to the same general topics, while using higher resolution values will result in a more fine-grained topic categorization of the documents.

3 Dataset Description

The dataset used in our paper is from n-tv[1], a German free-to-air television news channel. There are mainly two different sets of files in the proprietary data. One part of the dataset is represented by the speech segments (audio data) with an average duration of 1.5 minutes where each recording has multiple speakers recorded in a relatively noisy environment. The other part of the dataset is the textual transcripts (German) associated with the speech segments. The dataset contains both labeled (topic) and unlabelled data. Each of the transcript files contains articles (text documents) spread across different topics. To perform our experiments, we have used the unlabeled set of texts, i.e., a total of 697 articles.

Table 1 shows some statistics from the employed

[1] https://www.n-tv.de/

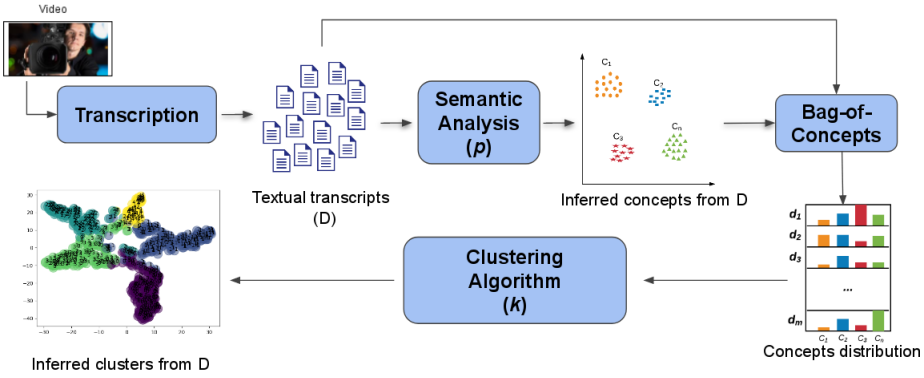

Figure 1: General framework to categorize short transcript texts using low-resolution concepts as representation.

	W/O Pre-processing Average (σ)	Total
Tokens	234.68 ($\pm$ 124.45)	163,572
Vocabulary	161.79 ($\pm$ 51.92)	22078
LR	0.717 ($\pm$ 0.073)	0.134
	W/ Pre-processing Average (σ)	Total
Tokens	63.02 ($\pm$ 31.52)	43,928
Vocabulary	47.86 ($\pm$ 16.30)	11,948
LR	0.785 ($\pm$ 0.092)	0.272

Table 1: Statistics of the German News Channel text data in terms of number of tokens, vocabulary and lexical richness.

dataset; before applying any pre-processing operation and after pre-processing. As pre-processing operations, we removed stop-words, numbers, special symbols, all the words are converted to lowercase, and we preserve only German nouns[2]. We compute the average number of tokens, vocabulary, and lexical richness (LR) in the dataset. A couple of main observations can be done at this point. On the one hand, we notice that individual texts are very short, on average 63.02 tokens with an average vocabulary of 47.86 words, resulting in a very high LR (0.785). This suggests that very few words are repeated within one article, very few redundancies, making the categorization task more challenging. On the other hand, globally speaking, the complete dataset has an LR=0.272, which indicates, to some extent, that the information across texts is highly overlapped (narrow domains).

4 Experimental framework

For all the performed experiments we ran the k-means algorithm for a range of $k = 2 \ldots 15$, and as the evaluation metric, we employed the Silhouette coefficient (Rousseeuw, 1987).

4.1 Obtaining word vectors

One crucial step of our approach is learning word representations. For this, an important parameter is the resolution value (p), which indicates the number of concepts that will be employed for building the document representation. Accordingly, we evaluate three different methods for inferring the set $\mathcal{C}$ ($|\mathcal{C}| = p$):

BoC: As described in (López-Monroy et al., 2018), concepts are inferred from applying a clustering process over $\mathcal{V}$, using as word representation pre-trained word embeddings. For our experiments we used word embeddings trained with FastText[3] on 2 million German Wikipedia articles.

LDA: For this, we used the Mallets LDA implementation from within Gensim[4]. After obtaining the concepts, we compute the concepts distribution over each d_j for generating the $\mathbf{d_j}$ representation.

LSA: For this we employed the SVD (singular value decomposition) algorithm as implemented in the sklearn toolkit[5].

4.2 Comparisons

We compare the proposed methodology against three different approaches:

BoW(*tf-idf*): For this experiment short texts are represented using a traditional BoW considering a *tf-idf* weighing scheme.

Avg-Emb: Every short text is represented using

[2]We employed the German POS tagger from https://spacy.io/

[3]https://www.spinningbytes.com/resources/wordembeddings/

[4]https://radimrehurek.com/gensim/models/wrappers/ldamallet.html

[5]https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

k	**BoW** (*tf-idf*)	**Avg-Emb** (*fasttext*)	**BoC** (*p=5*)	**LDA** (*p=5*)	**LSA** (*p=5*)	**CNNs**
2	0.0047	0.0612	**0.2720**	0.2769	0.4738	**0.1514**
3	0.0066	**0.0670**	0.2504	0.2943	**0.5123**	0.1412
4	0.0071	0.0582	0.2533	0.3653	0.4981	0.1453
5	0.0079	0.0622	0.2311	**0.4167**	0.4809	0.1242
6	0.0095	0.0608	0.2223	0.384	0.4892	0.1227
7	0.0088	0.0619	0.2273	0.3693	0.4645	0.1228
8	0.0090	0.0635	0.2152	0.3556	0.4603	0.1101
9	0.0123	0.0632	0.2189	0.3383	0.3204	0.1007
10	0.0118	0.0594	0.2173	0.3199	0.3046	0.0953
11	0.0110	0.0597	0.2143	0.2935	0.3087	0.0956
12	0.0104	0.0573	0.216	0.3015	0.3106	0.0908
13	0.0128	0.0600	0.2081	0.2935	0.3087	0.1065
14	**0.0142**	0.0539	0.2122	0.2864	0.3101	0.0992
15	0.0124	0.0592	0.2096	0.2761	0.3153	0.0941

Table 2: Clustering performance, in terms of Silhouette score, considering $k = 2 \ldots 15$.

the average of the word embeddings which are respectively weighted with their *tf-idf* score. This strategy has been considered in previous research as a common baseline (Huang et al., 2012; Lai et al., 2015; Xu et al., 2015).

CNNs: A convolutional neural network for clustering short texts[6] designed to learn deep features representations without using any external knowledge (Xu et al., 2015).

5 Results

First, we determine the impact of the resolution parameter (p) in the clustering task. Then, we compare the proposed method using the best value of p against methods described in section 4.2.

5.1 Impact of the resolution

In Figure 2 we visually show the performance of the considered concepts-inferring approaches in the clustering task, i.e., BoC, LDA, and LSA. Each map depicts the performance of the different methods under several resolution values ($p = 5, 10, 20, 50, 100, 500, 1000$), and several required clusters ($k = 2, \ldots, 15$). The brighter the color in the heat-map, the higher the silhouette score. From these experiments we observe that: *i)* using low-resolution values ($p = 5, 10$) allows to obtain better performance; *ii)* inferring concepts with LSA obtains the highest performance across different values of k.

[6] As implemented in https://github.com/zqhZY/short_text_cnn_cluster

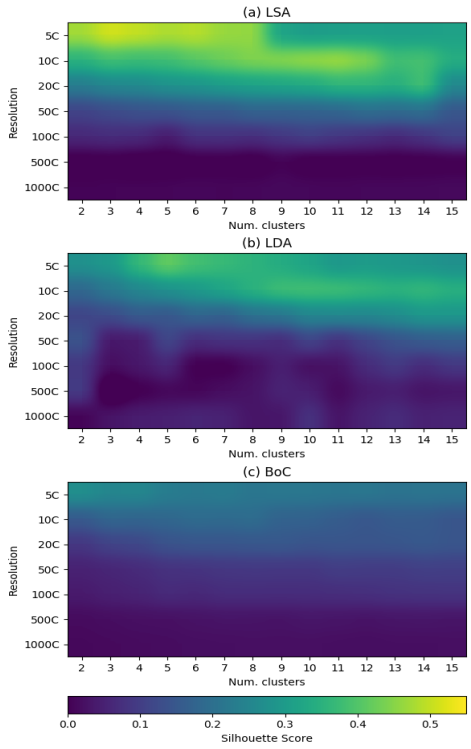

Figure 2: Heatmaps showing the impact of the resolution parameter in the clustering task.

5.2 Overall performance

Experiments from the previous section indicate that the optimal value for the resolution parameter is $p = 5$, independently of the approach for inferring concepts. Table 2 shows the performance of the proposed approach in comparison with the methods described in section 4.2. As can be observed, traditional BoW and Avg-Emb techniques obtain the worst performance. Although the method based on CNNs (Xu et al., 2015) improves the performance, its obtained results are far from reaching those obtained with the different configurations of the proposed approach (i.e., BoC, LDA, and LSA), specifically, our best configuration obtains a relative improvement of 70.4% against the CNN approach. Another interesting observation is that inferring low-resolution concepts with LSA allows the clustering algorithm to obtain an acceptable performance as more fine-grained categories are required, i.e., as k increases.

6 Conclusions

In this paper, we proposed using highly dense representations, denominated low-resolution concepts, for clustering German broadcast media contents. Performed experiments demonstrate that using small resolution values provides a better cluster-

ing performance, particularly, when concepts are inferred using the LSA approach, the clustering performance, over several values of k, overcome traditional approaches, as well as some recent CNN based methods for unsupervised categorization.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Docbert: BERT for document classification. *CoRR*, abs/1904.08398.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

M Doulaty, O Saz, RWM Ng, and T Hain. 2016. Automatic genre and show identification of broadcast media. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 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. ACM.

Zhixing Li, Zhongyang Xiong, Yufang Zhang, Chunyong Liu, and Kuan Li. 2011. Fast text categorization using concise semantic analysis. *Pattern Recognition Letters*, 32(3):441–448.

Adrian Pastor López-Monroy, Fabio A González, Manuel Montes, Hugo Jair Escalante, and Thamar Solorio. 2018. Early text classification using multi-resolution concept representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1216–1225.

Mohamed Morchid and Georges Linarès. 2013. A lda-based method for automatic tagging of youtube videos. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching bert with knowledge graph embeddings for document classification.

Berthier Ribeiro-Neto and Ricardo Baeza-Yates. 1999. Modern information retrieval. *Addison-Wesley*, 4:107–109.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

Ahmad Muqeem Sheri, Muhammad Aasim Rafique, Malik Tahir Hassan, Khurum Nazir Junejo, and Moongu Jeon. 2019. Boosting discrimination information based document clustering using consensus and classification. *IEEE Access*, 7:78954–78962.

Todor Staykovski, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. 2019. Dense vs. sparse representations for news stream clustering. In *Text2Story@ ECIR*, pages 47–52.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, Denver, Colorado. Association for Computational Linguistics.