Vorabfassung des Artikels

Hirschmeier, Stefan, and Johannes Melsbach. "Automated Keyword Generation in the Public Radio Sector Using Word Embeddings." In Proceedings of the European Conference on Information Systems (ECIS), 2019. https://aisel.aisnet.org/ecis2019_rip/74.

AUTOMATED KEYWORD GENERATION IN THE PUBLIC RADIO SECTOR USING WORD EMBEDDINGS

1 Introduction

Public broadcasters find themselves in a difficult situation when it comes to digital transformation. In more and more use cases, metadata is needed, e.g. to allow radio editors to search for content pieces, to provide short summaries of items in the media library, to set up content-based recommendation services for listeners, to allow users to browse by categories or tags, or to optimize content for search engines on the internet. Public broadcasters are in need of proper metadata to survive the digital transformation and to offer new and timely services.

Unfortunately, current technical infrastructures of radio broadcasters are primarily optimized for linear distribution of the content. These systems that exist for decades already have not been designed to bring rich metadata along with the content. Typically, at the time when content goes on air for linear distribution, only few to none metadata about radio shows is available. Even if metadata is generated afterwards, e.g., for enriching the digital representation of content on the website or for archiving purposes, the dominance of linear distribution structures complicates the provisioning of digital content on websites, media centers, and for recommender systems. Whereas few broadcasters have already overthought their metadata generating processes, the situation of lacking metadata still holds true for many public broadcasting agencies.

Without metadata, however, public broadcasters risk being left behind from competitors such as private audio streaming services that become more and more popular and offer a rich and individualized user experience. Already today, public broadcasters have lost certain customer segments and fear a generation tear-off. For public broadcasters, it is no longer sufficient to product high-quality content and to distribute it over linear channels, but to care for accessibility of their content, and to ensure the content finds its way to the audience. If public broadcasters fail to innovate their offerings, they also fail to comply with their public-service remit, which is one of the most valuable services of the public sector with respect to democracy and the forming of opinions.

Metadata is key to offer new services and to get listeners engaged with content and media platforms. In the medium term, if not already short term, public broadcasters have to find a solution how to generate metadata, either manually or automatically. Both have their challenges.

In manual metadata generation–especially in data rich applications–the human factor becomes a bottleneck. The result is that only few keywords are assigned, not all items are covered, or keyword generation is subjective and not repeatable. As an alternative, automatic keyword generation techniques can provide quantity, but have other drawbacks, as the keyword generation techniques are often either not accurate enough, produce too many irrelevant keywords, are too specific or too general, or simply do not match a standard vocabulary of the organization. In other words, the information quality of automatically generated keywords in the comprehension of fitness-for-use (Wang and Strong, 1996) is hard to generate and often very limited.

Public broadcasters in Germany have a standard vocabulary which is used as a defined reservoir for assigning keywords. This vocabulary is a predefined and finite set of keywords in form of a taxono-

my. Considering supervised machine learning (i.e. document classification) for this task, a classifier would need to be trained to automatically assign items from the taxonomy (i.e. keywords) to documents. However, next to the challenge of a huge training set (several hundred instances per class) and classification performance (i.e. precision, recall and F-score), both the vocabulary of the documents and the vocabulary of the taxonomies change over time. For new upcoming words (like "Brexit"), a classifier might perform poor on accuracy, and each time taxonomy vocabulary needs to be adapted, classifiers have to be trained again.

Our research we document in this paper can be best described by asking the question: **In the realm of public radio broadcasting, how can keywords be derived from a collection of documents while only using a predefined taxonomy?**

The remainder of this paper is structured as follows: in section 2, we give an overview of approaches for keyword generation. In section 3, our approach is outlined. Section 4 presents our premilinary evaluation and section 5 concludes with a discussion.

2 Related Work

2.1 Existing approaches for keyword identification in text documents

One of the most fundamental approaches in keyword identification is the term-frequency/inverse document frequency method (TF/IDF) (Baeza-Yates and Ribeiro-Neto, 2010). TF/IDF identifies keywords that quantitatively best differentiate documents within a document collection, is unsupervised and easy to compute. Unfortunately, TF/IDF does not allow to match keywords to items of an existing taxonomy. Furthermore, it is not applicable to single documents, as the measure requires a document collection to evaluate the keywords' descriptiveness. Next to TF/IDF, topic modeling methods such as PLSI (Hofmann, 1999) or LDA (M. Blei, Y. Ng and Jordan, 2001) have been proposed to identify word collections from documents that aim to describe topics the documents deal with. Similar to TF/IDF, topic modeling methods are unsupervised which makes them promising for application on a large scale. Nevertheless, extracted keyword collections have been reported as being hard to interpret (Debortoli, Müller, Junglas and Brocke, 2016) which limits applicability in end-user systems. Furthermore, additional supervised processing is necessary to match extracted keywords to existing taxonomies. In this case, the results of TF/IDF-based or topic modelling methods have to be used as features in supervised document classification, which leads to challenges in precision and recall when the predefined taxonomy is large, or taxonomy items are not perfectly distinctive.

Finally, previously proposed keyword generation approaches all rely on a so-called bag-of-words (BoW) perspective. This means, that the quantitative presence or absence of words is the foundation of any representation vector (e.g. document or word). In keyword relevance computations this is challenging, as spelling errors, synonyms or word flections lead to large but sparse matrices. However, applying methods like stemming, lemmatizing, spelling correction, synonym or ontology mapping does only address the symptoms (of sparse matrices), but not the root cause as BoW approaches are limited in capturing semantics like context. Often, low accuracy is the result of mapping an arbitrary set of keywords to a predefined taxonomy1.

A promising approach to address BoW context loss was proposed by Mikolov et al. (Mikolov, Chen, Corrado and Dean, 2013). Word2Vec creates a distributed representation of words2 (or larger entities such as phrases or documents) that is not dependent on the presence or absence of the target word,

[1] The terms mapping, classification, annotation or tagging are used interchangeably throughout this paper,

[2] Distributed representation is often connected to the term "word embeddings" and both denote the results of the word2vec approach by Mikolov et al. (Mikolov, Chen, Corrado and Dean, 2013).

but creates a vector representation of a word's context. One important characteristic of these word embeddings is that semantic similarity corresponds to arithmetic distance. The paper at hand considers this representation as a foundation for novel keyword generation approaches. Therefore, related literature will be detailed in the following.

2.2 Classification with Distributed Representations of Documents

The most prominent approach of distributed representation of text documents is the paragraph2vec approach of Le and Mikolov (Le and Mikolov, 2014). It proposes two different methods to train local document vectors along with global word vectors. Before Le and Mikolov, other researchers have proposed extensions of the word2vec model to obtain distributed representations of sentences, phrases or documents (Mitchell and Lapata, 2010; Zanzotto, Korkontzelos, Fallucchi and Manandhar, 2010; Yessenalina and Cardie, 2011; Grefenstette et al., 2013; Mikolov, Sutskever, et al., 2013). Approaches reach from simple ones that calculate an average of the words in a sentence, phrase or document, to more complex ones, e.g. that combine the word vectors in an order given by a parse tree (Socher, Lin, Ng and Manning, 2011).

Distributed representations of documents on the basis of word2vec approaches allow for a classification of documents with a subsequent classifier, typically a neural network. However, for all subsequent classification tasks on top of word2vec, manual effort is required. Experts need to link documents to classes to form a training set for the classifier.

For taxonomies, we typically face the challenge of having a high number of keywords (in magnitude of 10.000, sometimes even 100.000) and therefore as many classes for the classification task. With the growing number of classes for a classification task, also a significantly large manually annotated training set is required. Furthermore, this is not a one-time effort, but ongoing. If vocabulary changes over time (as new word like "Brexit" come up), both word2vec model and classifier have to be trained periodically to reflect newest words.

To sum up, the usage of word embeddings allow to perform a more meaningful feature engineering than solely relying on BoW based approaches. But considering the drawbacks of supervised classification for text documents on top of word embedding approaches, we find that a) a large number of manual tags have to be assigned by experts due to a high number of classes, b) manual tags have to be assigned not only initially, but continuously, and c) not only word2vec, but also the subsequent classifier has to be trained periodically. These drawbacks make classifiers on top of word2vec suitable in theory, but less suitable in practice.

There are few approaches that operate on the distributed representations of words and documents directly to find relations between documents or documents and keywords. One is the Word Mover Distance Approach of Kusner et al. (Kusner, Sun, Kolkin and Weinberger, 2015). It provides a distance function between documents based on word vectors without computing document vectors. Another approach was presented by Shperber (Shperber, 2017) by assigning 17 tags to documents without supervised training.

3 Approach

3.1 Approach Requirements

From public broadcasters' growing need for metadata, we elicited three requirements for keyword annotation, which we depict in short:

R1. Keyword annotation should take place automatically with a minimum of human effort (as unsupervised as possible),

R2. Identified keywords should be annotatable to an organization's specific taxonomy,

R3. New upcoming words (like “Brexit”) in documents should be matched as well.

In the following, we depict an approach to match the elicited requirements.

3.2 Approach Details

Assigning a definite set of taxonomy keywords to documents is a classification task. Given a document space and a taxonomy without any relation between them that could be used for supervised learning, we present an approach that does not require manually annotated training data and is robust against changes in vocabulary. Our approach consists of three steps:

First step: Train word vectors and document vectors on a company specific data set which we denote as world corpus vectors (WCV). Here we use the approach of Le and Mikolov as described in (Le and Mikolov, 2014).

Second step: Under the prerequisite, that the keywords of the predefined taxonomy are also found within the world corpus (see first step), we collect the word vectors generated for the keywords in the taxonomy. As a result, we receive a common vector representation of both documents and taxonomy keywords.

Third step: We compute dot products as distances between documents vectors and word vectors to find best matches. For each document, we compute the dot products to all taxonomy word vectors and choose the ones with the highest cosine similarity. Multiple taxonomy keywords can be assigned by choosing the top n keywords or keywords over a certain threshold of similarity. Emerging keywords (e.g. due to changing vocabulary or emerging topics) might be discovered when word vectors from the WCV are found similar to a document vector, but do not exist in the predefined taxonomy.

This way, we can simplify the problem of keyword classification and discovery to algebraic vector operations (just as a modulo n function classifies all numbers into n buckets). The approach is not truly unsupervised, as some authors emphasize that word2vec is not unsupervised, but self-supervised, as some error backpropagation takes place through correct and incorrect predictions (Lilleberg, Zhu and Zhang, 2015). But in the sense that no annotation of human experts is required for training, the method can be considered as unsupervised.

Figure 1 depicts the approach in contrast to simple classification approaches (Figure 1 top), where the classification is done on single words or n-gram representations only. Figure 1 (middle) shows advanced classification approaches, that use distributed representations as a better input for the classification task. Figure 1 (bottom) depicts the approach presented here that classifies via cosine similarity after transforming the taxonomy into vector representation as well. Considering the outlined requirements, R1 and R3 distinguish the approach from classifiers as depicted in Figure 1 (top and middle), and R2 distinguishes the approach from TF/IDF and topic modelling approaches.

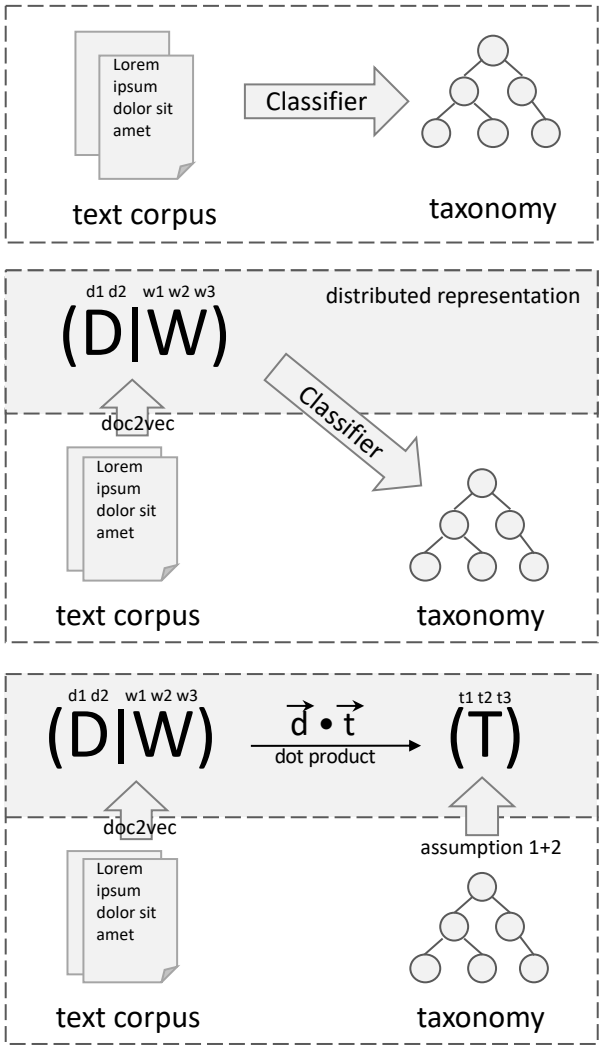

Figure 1: *(top) simple classification without distributed representation, (middle) advanced classification with distributed representation of words and documents, (bottom) our proposed approach.*

3.3 Approach Assumptions

The feasibility and simplicity of the approach bases on two assumptions we made. We discuss both in the following.

Assumption 1: We assume that it is valid to make a lookup for keyword word vectors.

The keywords in the taxonomy are just words without any given context except from their position in the taxonomy, which is a rather weak information compared to the richness of word vectors computed by word2vec. As we have no possibility to compute a rich context of taxonomy keywords, and the keywords in the taxonomy should have the same context as the words in the document corpus, we assume it is valid to make a lookup on the word vectors of the WCV and to use these vectors for the keywords in the taxonomy as well.

Assumption 2: The keywords in the taxonomy are part of the vocabulary of the document corpus.

When new words come up in the document corpus, the taxonomy must either be updated as well, or the taxonomy is required to consist of a more general, slower changing vocabulary that does not adopt fashion words. In general, the assumption that the vocabulary of the taxonomy is a subset of the vocabulary of the document corpus seems valid, as the taxonomy is designed to describe the documents. In other

words, no case should exist where a taxonomy keyword is not reflected in the document corpus, otherwise the taxonomy would fail to reflect the document corpus.

4 Evaluation

We evaluate our approach with a large text corpus of a German nation-wide public radio broadcaster that covers 400.000 manuscripts and about 6 million words. The broadcaster has an archive process where archivist manually assign keywords to the manuscripts. For all 400.000 documents, these manually annotated keywords are provided. The keywords are embedded in a company specific taxonomy that already lasts for decades and slowly changes over time.

We evaluate our approach from two perspectives:

First, we compare how many keywords, that have been manually assigned by archivists (which we consider as a reference set), have also been identified by our approach. Within this evaluation, we vary the threshold and top n words that the approach delivers to find an optimal threshold. As a result, we gain insights about the best accuracy that our approach can deliver. We also compare our accuracy results to other methods of keyword identification.

Second, for a subset of those documents where our approach identified more keywords than have been manually assigned by archivists, we manually assess whether these extra keywords fit, and if our approach outperforms manual tagging. Here, we carefully consider that our window of keywords is variable depending on the top n or threshold we defined.

Variations within the evaluation might include the use of the Continuous Bag-of-word (CBOW) vs. the Skip-gram approach (Mikolov, Chen, et al., 2013) resp. the Distributed Bag of Words version of Paragraph Vector (PV-DBOW) vs. the Distributed Memory version of Paragraph Vector (PV-DM) (Le and Mikolov, 2014), or the use of a general knowledge corpus for word vectors (e.g. derived from Wikipedia) instead of a company specific document corpus.

Preliminary results

5 Discussion

We proposed an approach to identify taxonomy keywords for document corpora with the help of word embeddings. In the same way as Le and Mikolov stated, that "an important advantage of paragraph vectors is that they are learned from unlabeled data and thus can work well for tasks that do not have enough labeled data" (Le and Mikolov, 2014), we pursued this advantage and extended it to keyword identification.

Our approach makes use of two assumptions that allow the use of word vectors for taxonomy keywords, which, regarded in themselves, have only few contextual information. Following a critical stance, one might say that the lookup for word vectors we perform is not valid (assumption 1) and nothing more than a simple keyword match between the document corpus and the taxonomy. From this point of view, our approach would suffer from the same problems that simplest keyword matching algorithms do. In a contrary view, exactly this lookup for word vectors, which have global validity over the document corpus, is the simple, yet value-adding step in the process.

From a traditional understanding, both the distributed vector representation and the subsequent classification contribute to the quality of a classification. We however question that the classification step adds value to the whole process and argue that the sophistication lies in the distributed representation, not in the classification. The classification is just a necessary task that needs to be done, but it does not necessarily need to be a machine learning classifier. We therefore design our approach without a separate classifier and obtain a much simpler and robust approach.

References

Baeza-Yates, R. and B. Ribeiro-Neto. (2010). *Modern Information Retrieval* (2ed edition). New York: Addison Wesley.

Debortoli, S., O. Müller, I. Junglas and J. vom Brocke. (2016). "Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial." *Communications of the Association for Information Systems*, *39*(1).

Grefenstette, E., G. Dinu, Y.-Z. Zhang, M. Sadrzadeh and M. Baroni. (2013). "Multi-Step Regression Learning for Compositional Distributional Semantics." Presented at the International Conference on Computational Semantics (IWCS).

Hofmann, T. (1999). "Probabilistic Latent Semantic Analysis." In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 289–296). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Kusner, M. J., Y. Sun, N. I. Kolkin and K. Q. Weinberger. (2015). "From Word Embeddings to Document Distances." In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37* (pp. 957–966). Lille, France: JMLR.org.

Le, Q. and T. Mikolov. (2014). "Distributed Representations of Sentences and Documents." In: *International Conference on Machine Learning* (pp. 1188–1196).

Lilleberg, J., Y. Zhu and Y. Zhang. (2015). "Support vector machines and Word2vec for text classification with semantic features." In: *2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)* (pp. 136–140).

M. Blei, D., A. Y. Ng and M. Jordan. (2001). "Latent Dirichlet Allocation." In: *The Journal of Machine Learning Research* (Vol. 3, pp. 993–1022).

Mikolov, T., K. Chen, G. Corrado and J. Dean. (2013). "Efficient Estimation of Word Representations in Vector Space." In: *ICLR Workshop*.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean. (2013). "Distributed Representations of Words and Phrases and their Compositionality." In: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc.

Mitchell, J. and M. Lapata. (2010). "Composition in Distributional Models of Semantics." *Cognitive Science*, *34*(8), 1388–1429.

Shperber, G. (2017, July 26). "A gentle introduction to Doc2Vec." Retrieved from https://medium.com/scaleabout/a-gentle-introduction-to-doc2vec-db3e8c0cce5e

Socher, R., C. C.-Y. Lin, A. Y. Ng and C. D. Manning. (2011). "Parsing Natural Scenes and Natural Language with Recursive Neural Networks." In: *Proceedings of the 28th International Conference on International Conference on Machine Learning* (pp. 129–136). USA: Omnipress.

Wang, R. Y. and D. M. Strong. (1996). "Beyond Accuracy: What Data Quality Means to data Consumers." *Journal of Management Information Systems*, *12*(4), 5–33.

Yessenalina, A. and C. Cardie. (2011). "Compositional Matrix-space Models for Sentiment Analysis." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 172–182). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zanzotto, F. M., I. Korkontzelos, F. Fallucchi and S. Manandhar. (2010). "Estimating Linear Models for Compositional Distributional Semantics." In: *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1263–1271). Stroudsburg, PA, USA: Association for Computational Linguistics.