Leveraging Natural Language Processing to Analyze Scientific Content: Proposal of an NLP pipeline for the field of Computer Vision

Henrik Kortum
German Research Center for Artificial Intelligence

Max Leimkühler
German Research Center for Artificial Intelligence

Oliver Thomas
Universität Osnabrück, German Research Center for Artificial Intelligence

Leveraging Natural Language Processing to Analyze Scientific Content: Proposal of an NLP pipeline for the field of Computer Vision

Henrik Kortum[1], Max Leimkühler[1] and Oliver Thomas[1,2]

[1] German Research Center for Artificial Intelligence, Smart Enterprise Engineering, Osnabrück, Germany
{henrik.kortum,max.leimkuehler}@dfki.de
[2] Universität Osnabrück, IMWI, Osnabrück, Germany
{oliver.thomas}@universität-osnabrueck.de

Abstract. In this paper we elaborate the opportunity of using natural language processing to analyze scientific content both, from a practical as well as a theoretical point of view. Firstly, we conducted a literature review to summarize the status quo of using natural language processing for analyzing scientific content. We could identify different approaches, e.g., with the aim of clustering and tagging publications or to summarize scientific papers. Secondly, we conducted a case study where we used our proposed natural language processing pipeline to analyze scientific content about computer vision available at the database IEEE. Our method helped us to identify emerging trends in the recent years and give an overview of the field of research.

Keywords: Natural Language Processing, Machine Learning, Emerging Trends, Computer Vision, W2V

1 Introduction

The number of scientific publications is developing rapidly and has been growing in recent years [1–4]. Due to the large number of publications, it is increasingly difficult to gain an overview of complex scientific topics and to derive trends for researchers [5–7]. For example, 23,777 publications on the topic of computer vision exist on the IEEE platform alone. Of these, 2,887 papers were published in 2019[1]. A manual review of the publications is connected with a very high effort and is almost impossible to handle. Nevertheless, researchers have a legitimate interest in gaining an overview of a research topic, e.g. computer vision. This problem has been addressed in various publications. A possible solution scenario for the aggregation of information is the use of natural language processing (NLP) to evaluate scientific content. E.g. NLP is used to summarize scientific papers or to extract

[1] see chapter 3.2 for the derivation of the numbers

16th International Conference on Wirtschaftsinformatik,
March 2021, Essen, Germany

key phrases. Based on this motivation and the resulting problems, the following research questions (RQ) are addressed in this paper:

RQ1: What is the status quo of using NLP for analyzing scientific content?

RQ2: Can NLP be utilized to structure keywords of a scientific text corpus and to identify trends?

To answer the RQ, this paper is structured as follows. First, in Section 2 foundations about NLP are presented. This is followed in section 3 by the concretization of the research approach. Sections 4 and 5 present the results, which are critically discussed in section 6. Finally, the paper is completed by the conclusion in section 7.

2 Foundations about Natural Language Processing

Tokenization is used to make a text processable by algorithms. Therefore, the string representing the text itself, should first be broken down into smaller elements, so called tokens. These can be, sentences, words, word pairs (n-grams) or single characters. The process of token generation is not trivial and ranges from a simple separation on the basis of "spaces" between words, over the use of lexicons, to the use of more complex procedures, such as conditional random fields or deep neural networks [8].

Besides tokenization *normalization* is an essential part of the preprocessing of texts and can be carried out by various methods e.g. stemming or lemmatizing. During stemming, words are traced to their word stem by using heuristics. This is often done by removing certain word endings [9]. It should be noted that stemming also inevitably leads to the loss of information and certain errors can occur.

In general, the preprocessing of a corpus also includes a *cleanup* process. Certain words can have a negative influence on NLP tasks, because they do not provide any semantic or contextual value [10]. These words are called stop words. They increase the dimensionality of the data set, which in turn has a negative influence on performance. Stop words can be divided into two categories: general and domain-specific stop words. General stop words occur in all texts and are independent of the subject of a text. Typical examples are articles or prepositions. Domain-specific stop words, on the other hand, have no explanatory value for a specific domain or a concrete analysis objective [11]. To achieve better results with NLP tasks, both general and domain-specific stop words should be removed during preprocessing [11].

In order to make texts or tokens processable by neural networks or other NLP algorithms a conversion into numerical representation is necessary. A widespread problem of many NLP techniques is the lack of the ability to map similarities and relationships between words and to consider contextual information [12]. Word embeddings are a popular and effective way to transform words into a machine-processable format [13]. They are capable of mapping both syntactic and semantic relationships between words by taking into account the context in which a word is mentioned [14]. Word embeddings represent words as vectors of real numbers. The entire vocabulary occurring in the training data set is transferred into a multi-dimensional vector space whose dimensions function as latent, continuous features. The transformation takes place via a flat neural network, which is trained on the basis of a

very large text corpus. The words used in the training vocabulary in a similar or identical context are arranged close to each other in the generated vector space [15]. Using similarity measures for vectors – e.g., cosine similarity – the similarity between words can be determined. Word vectors can be used to map semantic and contextual relationships between words [12]. A widely used method for clustering and comparing entire documents of a corpus is topic modeling [16]. In this context, latent dirichlet allocation (LDA) [17] is the most widespread approach. It is based on the assumption that each document can be represented as a probabilistic distribution over latent topics, where a topic in turn is characterized by a distribution over words [16]. Another, comparatively recent method that can be used for different NLP tasks are Bidirectional Encoder Representations from Transformer (BERT). To use BERT for NLP tasks pretraining and finetuning are required. During pretraining on unlabeled texts BERT learns deep bidirectional representations. In the finetuning step an additional layer can be added and BERT can be trained to solve specific tasks, like language inference or question answering. With BERT state of the art results have been archived on several natural language processing tasks [18]. Another approach called ELMo, short for embeddings from language models, can also be used for a variety of natural language processing tasks and is state of the art. In ELMo a deep bidirectional language model pretrained on large text corpus is used. These representations can be added to existing models to improve the performance on different NLP tasks [19].

3 Research Approach

To answer RQ1, first a literature review as described in section 3.2 was performed. The results of the literature review are also included to answer RQ2. Furthermore, a case study was conducted to investigate RQ2. A proposed method based on a NLP-pipeline was tested to structure keywords within a research area and identify emerging trends. The proposed method is described in section 3.3 In this specific case study the research area of computer vision was investigated by the automated processing of *author keywords, abstracts and publication years* of scientific publications. The data collection is described in detail in section 3.2.

3.1 Literature Review

In order to answer RQ1 and to get first insights for RQ2 an structured literature review was conducted in consideration of [20] and [21]. With RQ1 and RQ2 the focus and thus also step 1 of the literature search according to [21], definition of review scope, was concretized firstly. Since the main focus is on the analysis of scientific content using NLP, these two expressions were integrated into the search term. The search string was formulated as followed secondly: “natural language processing” AND “scientific content”. According to [21] the third step of the literature review is the literature search. For the literature search the databases AISeL, Ebsco, IEEE, ISI Web of Knowledge, JSTOR, ScienceDirect, SpringerLink and Wiley were considered. The table below

gives an overview of the results of the literature search, which was conducted in august 2020.

Table 1. Findings of the initial literature search

database	total result	sorted by title	Sorted by content	without duplicates
AISeL	2,245	15	3	3
Ebsco	684	11	0	0
IEEE	83	8	4	3
Web of Knowledge	2,028	24	15	14
JSTOR	4	0	0	0
ScienceDirect	28	4	2	2
SpringerLink	78	16	4	4
Wiley	25	9	0	0
Sum	**5,175**	**87**	**28**	**26**

In addition to the initial search, a backward search was conducted to identify further relevant literature. During backward search [7], [22–32] were identified. A total of 38 sources were thus included in the literature analysis and synthesis. In order to ensure the actuality of the review, it was examined when the publications were released. The oldest publication to be considered in the further analysis is from the year 2006. The majority of the selected publications are from 2013 to 2020, which underlines the up-to-datedness of the topic. The fourth step of the literature search is the literature analysis and synthesis. A concept matrix according to [20] was used for the synthesis and content analysis of the literature. As concepts the goals of the NLP workflow of the respective paper were abstracted.

3.2 Data Collection

Computer vision is an established field of research, which has been in existence for many years, but has gained in relevance especially in recent times. For this reason, we have decided to investigate the research field in more detail. Therefore, all publications about computer vision were extracted from IEEE to test our proposed method. The search was limited to the keyword "computer vision" in abstract or title in order to obtain only relevant hits on the topic. The following information was extracted for each publication: *author keywords, abstracts and publication years*. The search was conducted on 13.08.2020. A total of 23,777 publications were identified and the above-mentioned information were extracted. Figure 1 shows the distribution of the publications over the time period.

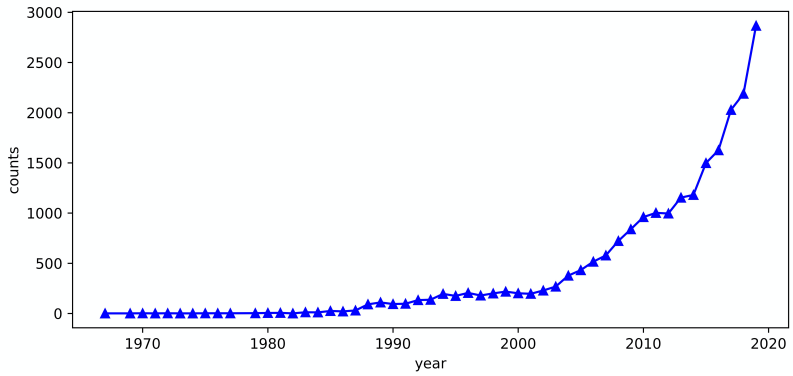

Figure 1. Distribution of publications about computer vision over time period

3.3 Proposed Method: NLP-Pipeline

Our proposed method consists of the five steps: preprocessing, training, prediction, evaluation and trend analysis. Each step of the NLP pipeline is divided into input, action and output as shown in figure 2. The key element of our pipeline is the generation of word embeddings to represent terms in a vector space. This word-level representation allows us to identify keywords used in similar contexts in order to expand the literature search and identify trends.

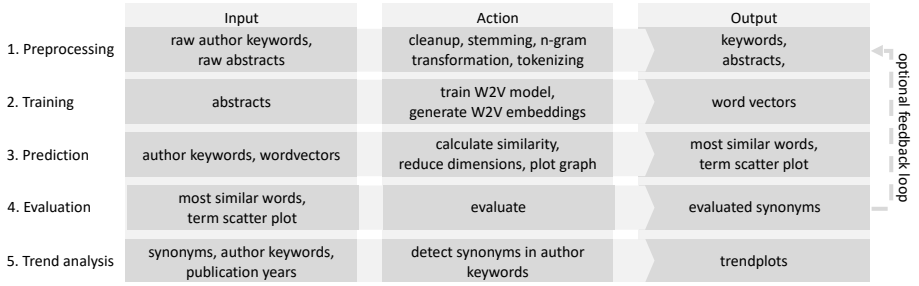

Figure 2. Proposed NLP pipeline for identifying emerging trends

1. The preprocessing of a raw text corpus is an essential step of every NLP pipeline [33]. Nevertheless, it is important to note that the implementation of preprocessing steps affects the resulting word embeddings and the performance of their calculation [9], [34]. In case of our pipeline the preprocessing consists of the steps, stop word removal stemming, n-gram transformation and tokenizing, which are applied to the data fields *author keywords* and *abstract*. These are normalized by a stemming process using the porter stemmer [35]. The purpose is to merge keywords with identical content (e.g., "*network*" and "*networks*"), so that they are considered the same in the following analysis. Additionally, stop words are removed from title and abstract that do not provide semantic or contextual meaning. The stemming and the cleaning have the purpose to optimize the subsequent training of the word vectors and to minimize the

number of data points in the resulting vector space. To include contextually relevant n-grams, we have merged all author keywords that are n-grams with a "_" character (e.g., *deep learning* → *deep_learning*), created a mapping and replaced matching n-grams in the abstract with these tokens. Finally, the full abstract string was split into lists of tokens by separating between spaces.

2. In the following step, a word vector model is trained, generating a 300-dimensional vector space based on the *preprocessed abstract*, whereby terms used in a similar context are placed close together. In order to learn the relevant contexts for a considered use case, the word vector model must be fitted to the corresponding scientific texts. We use the Python library *gensim* and a Word2Vec (W2V) model to generate the word vectors. The model was trained with the continuous bag of words (CBOW) method, over 500 epochs, with a window size of five. We decided to apply a word vector model because our case study is about identifying synonyms for keywords and therefore requires an approach that allows a calculation of similarities on term level. The strength of word vector models - also compared to more recent approaches, such as BERT - is based on the possibility to simply analyze terms in the spanned vector space using vector geometry. Specifically, W2V was chosen as the underlying calculation method, since it tends to perform well on stemmed corpus [34].

3. Calculating cosine similarity, the most similar terms for given keywords can now be retrieved to find synonyms in vector space. The idea is that through the trained W2V model, the user gets suggestions for synonyms which he might not have found on its own. For further exploration, a visualization of the learned word representations is useful. Since we are particularly interested in maintaining local similarity structures for synonym recognition, we choose UMAP [36] to reduce the dimensionality of our embeddings. The resulting 2D vectors are displayed in a scatterplot to visualize the subject area and provide a starting point for identifying additional keywords.

4. Evaluation: Similar terms identified should be treated as suggestions and carefully evaluated by the researchers, since not all terms discovered are necessarily contextual synonyms. The identification of unsuitable terms leads to a feedback into the preprocessing phase, where they can be added as stop words. If necessary, preprocessing steps can be adapted, e.g. to adjust the degree of stemming if terms cannot be interpreted by the researcher or are over/under stemmed [9].

5. Finally, matching synonyms can be included in the search by looking for the corresponding substrings in the *author keywords* data. For each publication year, all papers are selected that contain the keyword to be analyzed or its synonyms as *author keyword* in order to show a trend of the chosen topic. Section 5 shows an instantiation of our proposed method for a scientific text corpus from the field of computer vision.

4 Status Quo of NLP for Analyzing Scientific Content

The structured literature review has shown that NLP is used to analyze scientific content mainly for summarization, clustering and tagging of publications and to optimize as well as simplify a literature search. NLP is also used to create bibliometric networks, to analyze citations and to predict future research trends.

Table 2. Concept matrix for goals of using NLP for analyzing scientific content (S=Summarization, CT=Clustering and tagging, BN=Bibliometric networks, CS=Citation semantics, SL=Simplify literature search, OF=Overview and future trends)

#	author and year	S	CT	BN	CS	SL	OF
[37]	Abuhay et al., 2018						X
[24]	Abu-Jbara et al., 2013				X		
[26]	Achakulvisut et al, 2016					X	
[38]	Almeida et al., 2016					X	
[39]	Almugbel et al., 2019	X	X			X	
[40]	Avram et al., 2014			X	X		
[27]	Beltagy et al., 2019		X				
[31]	Chen and Zhuge, 2014	X					
[5]	Cohan and Goharian, 2018	X			X		
[28]	Collins et al., 2017	X					
[41]	Ghosh and Shah, 2020			X	X		
[42]	Giannakopoulos et al., 2013		X				
[43]	Hassan et al., 2018		X				
[44]	Janssens et al., 2006		X				
[32]	Joorabchi and Mahdi, 2013		X				
[3]	Kerzendorf, 2019					X	
[45]	Khan et al., 2016		X			X	
[46]	Koukal et al., 2014					X	
[47]	Krapivin et al., 2008		X			X	
[48]	Krasnov et al., 2019		X	X			X
[1]	La Quatra et al., 2020	X					
[29]	Li et al., 2019	X					
[49]	Li et al., 2018	X	X		X		
[50]	Łopuszyński and Bolikowski, 2015	X	X				
[51]	Łopuszyński and Bolikowski, 2014	X	X				
[6]	Ma et al., 2018	X					
[2]	Mueller and Huettemann, 2018	X	X				
[23]	Nam et al., 2016		X			X	
[52]	Nédey et al., 2018		X				
[53]	Petrus et al., 2019		X				
[22]	Prabhakaran et al., 2016		X				X
[25]	Qazvinian et al., 2013	X					
[7]	Qazvinian and Radev, 2008	X	X	X			
[54]	Sateli and Witte, 2014					X	
[55]	Schafer and Spurk, 2010		X	X	X	X	
[30]	Schäfer et al., 2008					X	
[4]	Sergio et al., 2019		X			X	
[56]	Szczuka et al., 2012		X			X	
Sum	38	13	21	5	6	13	3

The aim of the summarization is to provide the essential core statements of a scientific publication in a short and succinct manner. One approach of summarization is the processing of citations [5]. This approach is chosen because in citations a high aggregation of the contents has already been done [6]. In the table 2 the results of the literature analysis and synthesis are summarized.

Related to this is clustering and tagging. In clustering, an attempt is made to combine publications that deals with the same topic. Tagging is close to clustering. In tagging with NLP keywords were automatically assigned to publications by analyzing e.g., title and abstracts. Tagging is often used for organizing digital content [43].

Bibliometric networks are useful for visualizing connections between publications. Indicator for the networks can be e.g., authors, affiliations or keywords as well. In connection with summarization, bibliometric networks can help to give an overview of an entire topic [7]. Another aim is citation semantics. The aim is to predict in what context a citation is used. E.g., a citation can be used to criticize the scientific results of the cited paper, but it could also be used in a neutral and descriptive context. Possible approaches to predict the purpose and polarity of citations are supervised methods [24] as well as unsupervised ones [5].

The identification of relevant literature is important for the researchers [26]. Therefore, researchers are trying to improve the literature search with NLP. All above mentioned concepts are utilized to simplify the literature search. E.g. summarization [5], [39], clustering and tagging [4], [39], [56], bibliometric networks [55] and citation semantics [55] are used to optimize literature search and help researchers to identify relevant literature. NLP is also used to get an overview of a scientific area and to predict future trends. E.g. in [37] a non-negative matrix factorization topic modeling method is used to identify relevant research topics from scientific papers. The results are stored in time series data which is the basis for predicting future research trends with the help of auto-regressive integrated moving averages. A differentiated approach is described in [22]. Relevant topics were identified by using topic modelling. In addition to the pure terms, a classifier is used to examine in which context the extracted terms are used, e.g., as a method or as an objective. According to the authors' argumentation this has an influence on how a topic will develop in the future.

5 Emerging Trends in Computer Vision

In this chapter we operationalized our proposed method. Therefore, we conducted a case study to find synonyms for keywords and identify emerging trends in the field of computer vision. The identification of the trends is not to be understood as a forecast, but serves as an overview for the development of the different topics within the research area computer vision.

5.1 Preprocessing

In our case study the raw abstract contains 115,987 different terms. After a first cleaning process 55,195 terms remain. Table 3 shows the most common keywords for our computer vision corpus.

Table 3. Most relevant keywords (before stemming)

year	keyword (counts, relative counts)
overall	computer vision (1,679, 0.07), deep learning (970, 0.04), image processing (552, 0.02), machine learning (380, 0.02), object detection (342, 0.01), convolutional neural network (325, 0.01), feature extraction (272, 0.01), convolutional neural networks (272, 0.01), image segmentation (247, 0.01), segmentation (228, 0.01), face recognition (221, 0.01)
2019	deep learning (370, 0.13), computer vision (302, 0.11), convolutional neural network (118, 0.04), image processing (99, 0.03), machine learning (95, 0.03), object detection (94, 0.03), convolutional neural networks (92, 0.03), cnn (88, 0.03), segmentation (47, 0.02), image classification (45, 0.02), feature extraction (44, 0.02)
2018	deep learning (219, 0.10), computer vision (204, 0.09), machine learning (72, 0.03), convolutional neural network (69, 0.03), image processing (69, 0.03), convolutional neural networks (64, 0.03), cnn (43, 0.02), object detection (40, 0.02), feature extraction (30, 0.01), recognition (30, 0.01), image classification (27, 0.01)

The table can be explained using the example of "*deep learning*". In 2019 a total of 370 publications were tagged with the keyword "*deep learning*", corresponding to about 10 % of the publications in 2019. A look at the previous year 2018 shows a distinct trend.

5.2 Training and Prediction

As you can see in table 3, there are many synonyms or close related terms in the 10 top words per year, like "*convolutional neural network*" and "*deep neural network*". Therefor a word vector model is used to find similar words (based on cosine similarity) and to aggregate them for further analysis. The example of the keyword "*object detection*" demonstrates how adding synonyms to the keywords can help to provide a more reliable overview of the research area. Here the search for the substrings "*detect*" and "*recognit*" reveals specific use cases of object detection which otherwise would not have been considered (e.g., *"mango species detection", "recognition of cars", "makeup detection", "malaria parasite detection").* The trained NLP model provides the researcher with knowledge in form of close related terms that he himself might not have known. For further investigation of the word vectors, we have visualized them in a scatterplot in which each term represents a data point as well as implemented a function to display the *n* most similar words to a given keyword. The overall scatterplot and the 10 most similar terms for the keywords "*object detection*", "*deep learning*" and "*classification*" are shown in Figure 3.

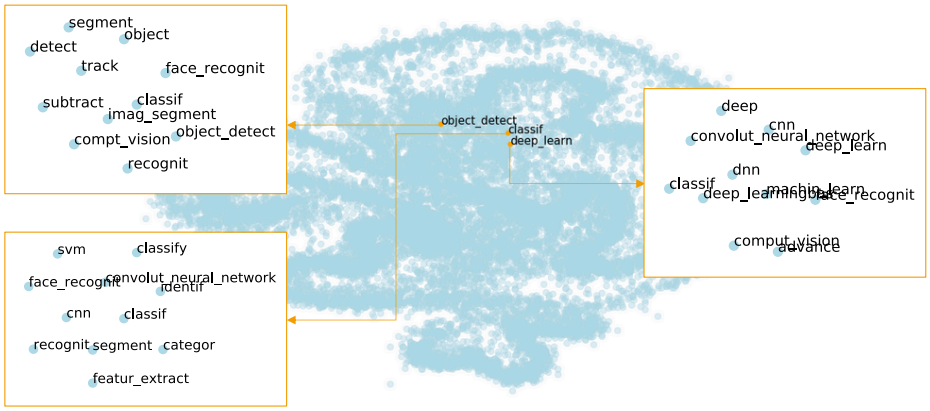

Figure 3 Visualization of trained word vectors reduced to 2 dimensions using UMAP

5.3 Evaluation

We treat the most similar words as suggestions and manually remove terms that we do not want to consider for the following trend analysis and therefore not to be added to the corresponding keywords. Similar terms can then be summarized, included in the analysis and help to obtain a better understanding of the subject area. Table 4 shows the 10 most similar terms for "*object detection*", "*deep learning*" and "*classification*" the removed words for the respective keyword are crossed out.

Table 4. evaluation of synonym suggestions for relevant keywords

keyword	synonyms (similarity score)
object detection	detect (0.41), object (0,4), subtract (0.37), recognit (0.36), classif (0.34), segment (0.33), ~~compt_vision~~ (0.33), imag_segment (0.32), track (0.32), face_recognit (0.32)
deep learning	machin_learn (0.53), deep (0.51), cnn (0.49), convolute_neural network (0.47), advance (0.38), dnn (0.37), ~~comput_vision~~ (0.35), ~~face_recognit~~ (0.34), deep_learningbas (0.34), ~~classif~~ (0.33)
classification	classify (0.59), recognit (0.58), categor (0.44), face_recognit (0.39), feature_extract (0.38), segment (0.36), ~~convolute_neural_network~~ (0.36), identif (0.36), svm (0.35), ~~cnn~~ (0.35)

5.4 Trend Analysis

The synonyms are now used for an investigation of all *author keywords* by searching for matching substrings. If a substring is contained in a keyword, the corresponding paper is considered relevant for our analysis. The following graphs in figure 4 show the development over time for three selected computer vision topics (including their synonyms): *"object_detection", "classification", "deep_learning"*. The results of the trend analysis can be confirmed by adding expert knowledge, e.g., for the keyword

"*deep learning*". Since 2000, deep learning has been successfully used for object detection, classification and segmentation. However, the breakthrough did come in 2012, when Krizhevsky et al. won the imagenet classification challenge [57]. They trained large, deep convolutional neural networks to classify images. This was the breakthrough of deep neural networks in the computer vision scene and deep learning has been one of the predominant methods for the detection and classification of objects [58].

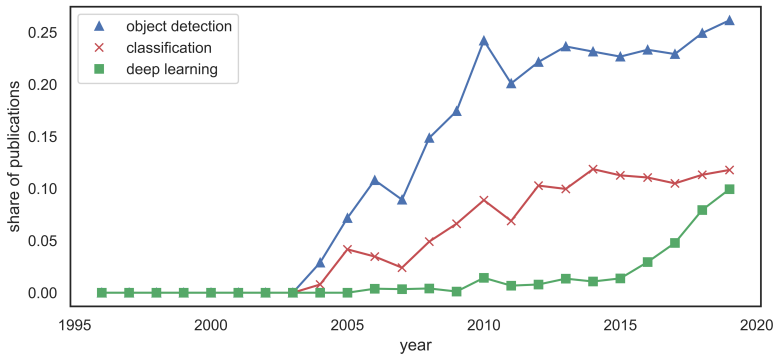

Figure 4. Emerging trends in computer vision

6 Discussion

In the following, implications as well as limitations will be discussed. Probably the most important implication of NLP for the analysis of scientific content arises for scientists themselves. Because of the large number of publications, it is difficult to get an overview of research areas [5]. NLP can help to solve exactly this problem with the concepts identified in the literature review. By clustering and summarizing publications, information is made available in an aggregated form. Bibliometric networks as well as the identification of emerging trends help to monitor the development of research. Sentiment analysis of citations provide an indication of the quality of a publication. The method presented in this paper also can be classified into the concept matrix of the conducted literature review: overview and future trends as well as simplify literature search.

There are also implications for practitioners. The identification of emerging trends plays an important role in open innovation. In open innovation, enterprises broaden their perspective and use external sources of information to identify innovations in order to improve their technologies [59]. Science is an established source for innovation in open innovation [60]. Our proposed method can help to optimize the open innovation process and to identify emerging trends early. Furthermore, the defined concepts during literature review can have impact on this. Due to the large number of scientific publications NLP can help in summarizing, clustering and tagging these documents. Thus, methods are made available to open innovation in order to handle the information overload. Related to this another implication can arise for economic planners and

training providers. The forecast of manpower requirements and the required skills is of particular importance for this target group [61]. Using the example of design science research, the connection can be illustrated. The goal of design science is the development, improvement and evaluation of powerful IT artifacts to support organizations in achieving their objectives [62]. At least when managers are convinced of the usefulness of an IT artifact, it is necessary to build up know-how in this area. Our method can help to identify these needs in advance. Furthermore, fast response times are a central component of a company's success and require the processing of large amounts of data [63]. As our NLP pipeline is not restricted to scientific texts and can also be transferred to corporate documents, it might be of assistance here. In this sense, our pipeline represents an approach to gain a better overview of large unstructured text sets and is thus a tool for text-dominated data ecosystems.

Our research has some limitations, which we present in the following. In relation to our proposed method, the question of generalizability arises, e.g., for fields with less frequency, because for the training of word-vector-models large data sets are required. Using the example of "Computer Vision", which provides a large data set, we were able to show that our proposed NLP pipeline is capable of structuring key terms of a scientific field and to identify emerging trends. Nevertheless, a case study cannot provide comprehensive evidence [64]. We want to encourage researchers to use our method to investigate other fields to identify emerging trends and to provide expert knowledge to support further evidence. In addition to expert knowledge for the evaluation of the results, other data sources can be used in further research projects, such as google search trends. From a data analysis point of view, it can be assumed that an extension of the text corpus on which the training is based would further improve the quality of the word vectors and learned connections. We therefore suggest connecting additional data sources for further work. The used abstracts provide a good basis, as they summarize the essential statements of a paper. However, an abstract does not reflect the full level of detail of a scientific paper or may even contain non-existent contributions [5], [65]. Due to this fact, further research has to be conducted to extend our method to full text analysis. Our presented NLP pipeline is to be understood as a support system, but not as an approach for a full automation. In addition, it should be verified if transfer learning approaches lead to better results by re-training pre-trained embeddings with the domain texts, instead of learning the word vectors from scratch. Further potential exists with regard to the model for generating word representations. In principle, the W2V model proposed in our pipeline can be substituted by other models as long as they support a vector representation at word level. LDA2Vec [66], for example, enables the joint training of word, topic and document vectors in a common representation space and thus offers a promising approach to combine the strengths of LDA with W2V like vector representations [66].

7 Conclusion

With regard to RQ1, the literature review showed that NLP is used to examine scientific literature. The main focus is the optimization of a literature search. Summarization as

well as clustering and tagging are common concepts that are used for this. With respect to RQ2, concepts have been identified during the literature search that address the problem of structuring and deriving research trends. In addition, the case study showed that our proposed NLP pipeline can be used to get a better overview of relevant terms within a research area. Therefore, we trained word-vector-models based on abstracts to find and aggregate most similar words. In the next step, emerging trends could be identified by using the synonyms for a given sets of keywords to search for the corresponding substrings in the Authors keywords. For the present use case we could show that our proposed NLP-pipeline helps to identify trends and to gain a more holistic picture of relevant terms within the topic area. The extent to which these findings can be applied to other fields and text corpus within and beyond the scientific field will have to be examined in further research.

References

1. La Quatra, M., Cagliero, L., Baralis, E.: Exploiting pivot words to classify and summarize discourse facets of scientific papers. Scientometrics. (2020).
2. Mueller, R.M., Huettemann, S.: Extracting Causal Claims from Information Systems Papers with Natural Language Processing for Theory Ontology Learning. Presented at the Hawaii International Conference on System Sciences (2018).
3. Kerzendorf, W.E.: Knowledge discovery through text-based similarity searches for astronomy literature. J Astrophys Astron. 40, 23 (2019).
4. Sergio, M.P., Costa, T. de S., Pessoa, M.S. de P., Pedro, P.S.M.: A Semantic Approach to Support the Analysis of Abstracts in a Bibliographical Review. In: 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). pp. 259–264. IEEE, Napoli, Italy (2019).
5. Cohan, A., Goharian, N.: Scientific document summarization via citation contextualization and scientific discourse. Int J Digit Libr. 19, 287–303 (2018).
6. Ma, S., Xu, J., Zhang, C.: Automatic identification of cited text spans: a multi-classifier approach over imbalanced dataset. Scientometrics. 116, 1303–1330 (2018).
7. Qazvinian, V., Radev, D.R.: Scientific Paper Summarization Using Citation Summary Networks. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. pp. 689–696. Association for Computational Linguistics, USA (2008).
8. Tomanek, K., Wermter, J., Hahn, U.: Sentence and Token Splitting Based On Conditional Random Fields. Presented at the (2007).
9. Jivani, A.: A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. Appl. 2, 1930–1938 (2011).
10. Mohan, V.: Preprocessing Techniques for Text Mining - An Overview. (2015).
11. Makrehchi, M., Kamel, M.S.: Automatic Extraction of Domain-Specific Stopwords from Labeled Documents. In: Macdonald, C., Ounis, I., Plachouras, V.,

Ruthven, I., and White, R.W. (eds.) Advances in Information Retrieval. pp. 222–233. Springer Berlin Heidelberg, Berlin, Heidelberg (2008).
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. (2013).
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. (2013).
14. Turney, P.D., Pantel, P.: From Frequency to Meaning: Vector Space Models of Semantics. jair. 37, 141–188 (2010).
15. Levy, O., Goldberg, Y.: Linguistic Regularities in Sparse and Explicit Word Representations. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning. pp. 171–180. Association for Computational Linguistics, Ann Arbor, Michigan (2014).
16. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications. 78, 15169–15211 (2019).
17. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research. 3, 993–1022 (2003).
18. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs]. (2019).
19. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. (2018).
20. Webster, J., Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Q. 26, (2002).
21. Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A.: Reconstructing the giant: On the importance of rigour in documenting the literature search process. In: ECIS (2009).
22. Prabhakaran, V., Hamilton, W.L., McFarland, D., Jurafsky, D.: Predicting the Rise and Fall of Scientific Topics from Trends in their Rhetorical Framing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1170–1180. Association for Computational Linguistics, Berlin, Germany (2016).
23. Nam, S., Jeong, S., Kim, S.-K., Kim, H.-G., Ngo, V., Zong, N.: Structuralizing biomedical abstracts with discriminative linguistic features. Computers in Biology and Medicine. 79, 276–285 (2016).
24. Abu-Jbara, A., Ezra, J., Radev, D.R.: Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In: HLT-NAACL (2013).
25. Qazvinian, V., Radev, D.R., Mohammad, S.M., Dorr, B., Zajic, D., Whidby, M., Moon, T.: Generating Extractive Summaries of Scientific Paradigms. jair. 46, 165–201 (2013).
26. Achakulvisut, T., Acuna, D.E., Ruangrong, T., Kording, K.: Science Concierge: A Fast Content-Based Recommendation System for Scientific Publications. PLoS ONE. 11, e0158423 (2016).
27. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A Pretrained Language Model for Scientific Text. (2019).

28. Collins, E., Augenstein, I., Riedel, S.: A Supervised Approach to Extractive Summarisation of Scientific Papers. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). pp. 195–205. Association for Computational Linguistics, Vancouver, Canada (2017).
29. Li, L., Zhu, Y., Xie, Y., Huang, Z., Liu, W., Li, X., Liu, Y.: CIST@CLSciSumm-19: Automatic Scientific Paper Summarization with Citances and Facets. In: BIRNDL@SIGIR (2019).
30. Schäfer, U., Uszkoreit, H., Federmann, C., Marek, T., Zhang, Y.: Extracting and Querying Relations in Scientific Papers. In: Dengel, A.R., Berns, K., Breuel, T.M., Bomarius, F., and Roth-Berghofer, T.R. (eds.) KI 2008: Advances in Artificial Intelligence. pp. 127–134. Springer Berlin Heidelberg, Berlin, Heidelberg (2008).
31. Chen, J., Zhuge, H.: Summarization of scientific documents by detecting common facts in citations. Future Generation Computer Systems. 32, 246–252 (2014).
32. Joorabchi, A., Mahdi, A.E.: Automatic keyphrase annotation of scientific documents using Wikipedia and genetic algorithms. Journal of Information Science. 39, 410–426 (2013).
33. Aklouche, B., Bounhas, I., Slimani, Y.: Query Expansion Based on NLP and Word Embeddings. In: TREC (2018).
34. Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., Mitra, M.: Using Word Embeddings for Information Retrieval: How Collection and Term Normalization Choices Affect Performance. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 1835–1838. ACM, Torino Italy (2018).
35. Porter, M.F.: An algorithm for suffix stripping. Program. 40, 211–218 (2006).
36. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. (2018).
37. Abuhay, T.M., Nigatie, Y.G., Kovalchuk, S.V.: Towards Predicting Trend of Scientific Research Topics using Topic Modeling. Procedia Computer Science. 136, 304–310 (2018).
38. Almeida, H., Jean-Louis, L., Meurs, M.-J.: Mining Biomedical Literature: An Open Source and Modular Approach. In: Khoury, R. and Drummond, C. (eds.) Advances in Artificial Intelligence. pp. 168–179. Springer International Publishing, Cham (2016).
39. Almugbel, Z., El, N., Bugshan, N.: Automatic Structured Abstract for Research Papers Supported by Tabular Format using NLP. ijacsa. 10, (2019).
40. Avram, S., Velter, V., Dumitrache, I.: Semantic Analysis Applications in Computational Bibliometrics. Control Engineering and Applied Informatics. 16, 62–69 (2014).
41. Ghosh, S., Shah, C.: Identifying Citation Sentiment and its Influence while Indexing Scientific Papers. Presented at the Hawaii International Conference on System Sciences (2020).
42. Giannakopoulos, T., Dimitropoulos, H., Metaxas, O., Manola, N., Ioannidis, Y.: Supervised Content Visualization of Scientific Publications: A Case Study on the ArXiv Dataset. In: Kłopotek, M.A., Koronacki, J., Marciniak, M., Mykowiecka,

A., and Wierzchoń, S.T. (eds.) Language Processing and Intelligent Information Systems. pp. 206–211. Springer Berlin Heidelberg, Berlin, Heidelberg (2013).
43. Hassan, H.A.M., Sansonetti, G., Gasparetti, F., Micarelli, A.: Semantic-based tag recommendation in scientific bookmarking systems. In: Proceedings of the 12th ACM Conference on Recommender Systems. pp. 465–469. ACM, Vancouver British Columbia Canada (2018).
44. Janssens, F., Leta, J., Glänzel, W., De Moor, B.: Towards mapping library and information science. Information Processing & Management. 42, 1614–1642 (2006).
45. Khan, A., Tiropanis, T., Martin, D.: Exploiting Semantic Annotation of Content with Linked Open Data (LoD) to Improve Searching Performance in Web Repositories of Multi-disciplinary Research Data. In: Braslavski, P., Markov, I., Pardalos, P., Volkovich, Y., Ignatov, D.I., Koltsov, S., and Koltsova, O. (eds.) Information Retrieval. pp. 130–145. Springer International Publishing, Cham (2016).
46. Koukal, A., Gleue, C., Breitner, M.H.: Enhancing literature Review Methods - towards More Efficient literature Research with Latent Semantic Indexing. In: ECIS (2014).
47. Krapivin, M., Marchese, M., Yadrantsau, A., Liang, Y.: Unsupervised key-phrases extraction from scientific papers using domain and linguistic knowledge. In: 2008 Third International Conference on Digital Information Management. pp. 105–112. IEEE, London, United Kingdom (2008).
48. Krasnov, F., Dimentov, A., Shvartsman, M.: Comparative Analysis of Scientific Papers Collections via Topic Modeling and Co-authorship Networks. In: Ustalov, D., Filchenkov, A., and Pivovarova, L. (eds.) Artificial Intelligence and Natural Language. pp. 77–98. Springer International Publishing, Cham (2019).
49. Li, L., Mao, L., Zhang, Y., Chi, J., Huang, T., Cong, X., Peng, H.: Computational linguistics literature and citations oriented citation linkage, classification and summarization. Int J Digit Libr. 19, 173–190 (2018).
50. Łopuszyński, M., Bolikowski, Ł.: Towards robust tags for scientific publications from natural language processing tools and Wikipedia. Int J Digit Libr. 16, 25–36 (2015).
51. Łopuszyński, M., Bolikowski, Ł.: Tagging Scientific Publications Using Wikipedia and Natural Language Processing Tools. In: Bolikowski, Ł., Casarosa, V., Goodale, P., Houssos, N., Manghi, P., and Schirrwagen, J. (eds.) Theory and Practice of Digital Libraries -- TPDL 2013 Selected Workshops. pp. 16–27. Springer International Publishing, Cham (2014).
52. Nédey, O., Souili, A., Cavallucci, D.: Automatic Extraction of IDM-Related Information in Scientific Articles and Online Science News Websites. In: Cavallucci, D., De Guio, R., and Koziołek, S. (eds.) Automated Invention for Smart Industries. pp. 213–224. Springer International Publishing, Cham (2018).
53. Petrus, J., Ermatita, Sukemi: Soft and Hard Clustering for Abstract Scientific Paper in Indonesian. In: 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). pp. 131–136. IEEE, Jakarta, Indonesia (2019).

54. Sateli, B., Witte, R.: Collaborative Semantic Management and Automated Analysis of Scientific Literature. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., and Tordai, A. (eds.) The Semantic Web: ESWC 2014 Satellite Events. pp. 494–498. Springer International Publishing, Cham (2014).
55. Schafer, U., Spurk, C.: TAKE Scientist's Workbench: Semantic Search and Citation-Based Visual Navigation in Scholar Papers. In: 2010 IEEE Fourth International Conference on Semantic Computing. pp. 317–324. IEEE, Pittsburgh, PA, USA (2010).
56. Szczuka, M., Janusz, A., Herba, K.: Semantic Clustering of Scientific Articles with Use of DBpedia Knowledge Base. In: Bembenik, R., Skonieczny, L., Rybiński, H., and Niezgodka, M. (eds.) Intelligent Tools for Building a Scientific Information Platform. pp. 61–76. Springer Berlin Heidelberg, Berlin, Heidelberg (2012).
57. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM. 60, 84–90 (2017).
58. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 521, 436–444 (2015).
59. Galbraith, B., McAdam, R.: The promise and problem with open innovation. Technology Analysis & Strategic Management. 23, 1–6 (2011).
60. Cassiman, B., Di Guardo, M.C., Valentini, G.: Organizing links with science: Cooperate or contract? Research Policy. 39, 882–892 (2010).
61. Wong, J., Chan, A., Chiang, Y.H.: A Critical Review of Forecasting Models to Predict Manpower Demand. CEB. 4, 43–56 (2012).
62. Hevner, March, Park, Ram: Design Science in Information Systems Research. MIS Quarterly. 28, 75 (2004).
63. Thomas, O., Varwig, A., Kammler, F., Zobel, B., Fuchs, A.: DevOps: IT-Entwicklung im Industrie 4.0-Zeitalter: Flexibles Reagieren in einem dynamischen Umfeld. HMD. 54, 178–188 (2017).
64. Abercrombie, N., Hill, S., Turner, B.S.: The Penguin dictionary of sociology. Penguin Books, London (1986).
65. Atanassova, I., Bertin, M., Larivière, V.: On the composition of scientific abstracts. Journal of Documentation. 72, 636–647 (2016).
66. Moody, C.E.: Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. (2016).