Exploiting Contextual Embeddings to Extract Topic Genealogy from Scientific Literature

Exploiting Contextual Embeddings to Extract Topic Genealogy from Scientific Literature

Alfio Ferrara[1], Stefano Montanelli[1], Sergio Picascia[1,*] and Davide Riva[1,*]

[1]Università degli Studi di Milano
Department of Computer Science
Via Celoria, 18 - 20133 Milano, Italy

Abstract

Modeling the evolution of topics and forecast future trends is a crucial task when analyzing scientific papers. In this work we propose tASKE (temporal Automated System for Knowledge Extraction), a dynamic topic modeling approach which exploits zero-shot classification and contextual embeddings in order to track topic evolution through time. The approach is evaluated against a corpus of data science papers, assessing the ability of tASKE to correctly classify documents and retrieving relevant derivation relationships between older and new topics in time.

Keywords

Natural Language Processing, Scientometrics, Topic Genealogy

1. Introduction

With the amount of published scientific literature increasing each year, keeping track of newly formulated topics and their derivation process becomes a challenge for researchers, scholars, and publishers. The problem lies in the fact that the total amount of definitions, theorems, properties, tasks, and subdomains tends to grow exponentially, since several of them may be conceived starting from a single one or the interaction of a few ones. For instance, in the domain of Machine Learning, the idea of *neural networks* gave rise to that of *deep learning*, which has then been applied to problems such as *image reconstruction* and *partial differential equations*, and was further deepened with topics such as *attention*, which in turn provided the intuition behind *transformers* and a basis for *explainability*.

Referring to definitions, theorems, properties, tasks, subdomains and the like with the generic label of "topics", abstract objects a text refers to, it is possible to study "topic genealogy" in a diachronic corpus, i.e. the descent of topics from older ones over time. The task of extracting topic genealogy falls within the scope of Knowledge Extraction (KE), and it consists of two main sub-tasks: i) *topic extraction*, by which we aim to retrieve topics that are important in a written document, possibly in a timely manner in order to discover topics when they actually appear, and ii) *genealogy reconstruction*, in which extracted topics are placed in a tree structure representing their lineage in the history of the discipline.

In this paper, we present tASKE, a method to extract topics from a diachronic corpus of scientific papers and reconstruct their genealogy in a completely unsupervised way. Our method is developed upon our Automated System for Knowledge Extraction (ASKE) framework [1], which relies on pre-trained contextual embedding models to represent documents and topics in the same vector space and on a cyclical term extraction and clustering phase to extract new topics. Besides presenting tASKE as a time-aware extension of ASKE, we introduce an evaluation framework and a case study on a corpus of abstracts of scientific papers related to the Data Science domain, with the goal of demonstrating the effectiveness of tASKE both for topic extraction and for extracting topic-to-topic derivation relationships.

The work is organized as follows: Section 2 *Related Work* reports on the literature about topic modeling as well as the technology underlying our method. Section 3 *Methodology* presents the methodology and techniques enforced in tASKE. Section 4 *Case Study and Evaluation* presents the case study on a Data Science Literature corpus, on which the evaluation was conducted. Section 5 *Concluding Remarks* draws some conclusions and sketches some future work.

2. Related Work

The task of classifying large amounts of textual documents without relying on labeled data and presenting latent features of texts, such as hidden topics, is commonly addressed employing topic modeling techniques. Latent

SDU2023: The Third AAAI Workshop on Scientific Document Understanding, February 14, 2023, Washington, DC
*Corresponding author.
alfio.ferrara@unimi.it (A. Ferrara); stefano.montanelli@unimi.it (S. Montanelli); sergio.picascia@unimi.it (S. Picascia); davide.riva1@unimi.it (D. Riva)
0000-0002-4991-4984 (A. Ferrara); 0000-0002-6594-6644 (S. Montanelli); 0000-0001-6863-0082 (S. Picascia); 0009-0003-9681-9423 (D. Riva)

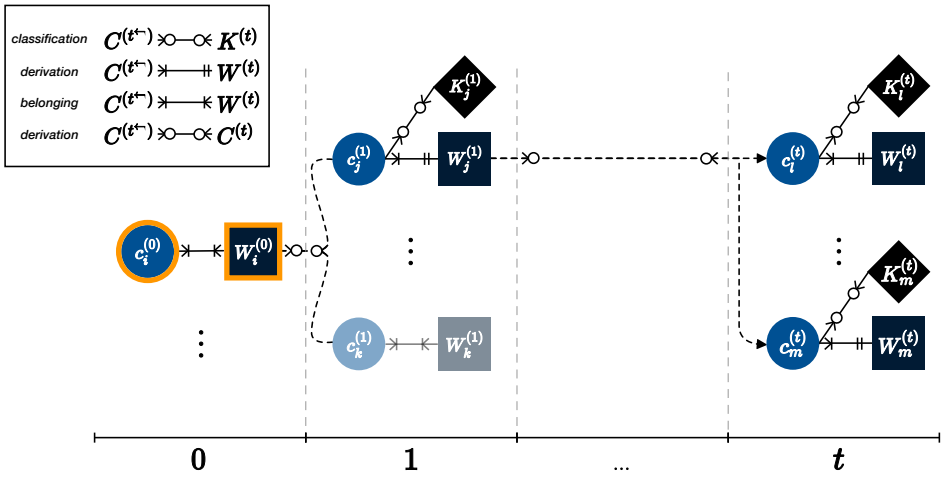

Figure 1: The tASKE Conceptual Graph.

Semantic Analysis (LSA) [2] was one of the first proposed approaches, exploiting Singular Value Decomposition (SVD) in order to reduce the number of dimensions of a document-term matrix and to easily compute similarity between document vector representations. LSA was soon followed by Latent Dirichlet Allocation (LDA) [3], which employs Bayesian analysis in order to optimize the distributions of documents belonging to topics, and of words defining these topics. The majority of recent works in topic modeling takes its inspiration from the original LDA with several variations proposed, such as Correlated Topic Modeling [4] and Hierarchical Topic Modeling [5].

Common topic modeling methods are not able to capture the changes of topics over time. For this reason, techniques of Dynamic Topic Modeling (DTM) are employed when dealing with diachronic corpora. Since the first approach (Dynamic LDA [6]) was proposed, the field has been attracting attention among researchers. Among the possible applications of the designed methods, the study of scientific papers, also known as "Scientometrics", was addressed with the aim to assess past and present trends in a specific discipline [7] or to forecast possible future subareas of research interest [8].

Later studies have been taking into consideration the integration between DTM and word embeddings [9] so to further capture the semantic aspect of the analyzed documents [10]. Embedding techniques are vastly employed in the field of Natural Language Processing (NLP), in order to represent textual data in a vector space. Several models capable of computing contextual token embeddings have been released since the presentation of BERT [11], each of them being tailored to specific tasks, such as semantic similarity [12] and zero-shot learning [13].

Zero-Shot Learning (ZSL) is a problem setup in the field of machine learning, where a classifier is required to predict labels of examples extracted from classes that were never observed in the training phase. It was firstly referred to as *dataless classification* in 2008 [14] and has quickly become a subject of interest, particularly in the field of NLP. The great advantage of this approach consists in the resulting classifier being able to operate efficiently in a partially or totally unlabeled environment.

tASKE aims at dynamically modeling the presence and evolution of latent topics in a diachronic corpus of documents. It exploits zero-shot learning and contextual embeddings not only to perform the classification task, but also to extract relevant knowledge from textual data.

3. Methodology

The objective of tASKE is to extract a genealogy of topics from a diachronic corpus of documents. Every piece of information is stored in a graph-based data structure called tASKE Conceptual Graph (ACG), whose architecture is illustrated in Figure 1.

The nodes in the ACG model belong to three different categories:

- **document chunks** K**:** the object of the analysis, they are small portions of the original documents extracted through the application of tokenization techniques. They are tuples of the form $(k, \mathbf{k})$, where k is the text of the document chunk and $\mathbf{k}$ is its vector representation;

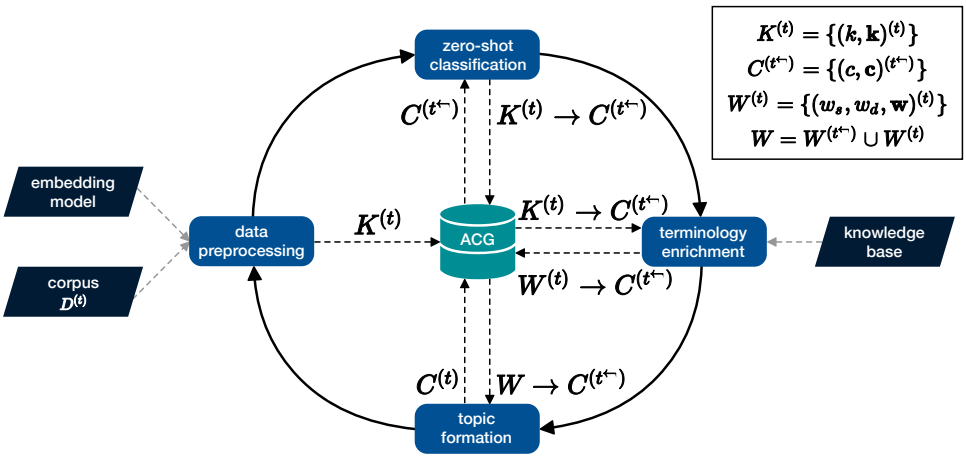

Figure 2: The tASKE cycle at time t.

- **topics** C**:** they represent the abstract objects to which documents chunks are assigned and, in practice, they are clusters of related terms. They are tuples of the form $(c, \mathbf{c})$, where c is the label given to the topic and $\mathbf{c}$ is its vector representation;
- **terms** W**:** they are extracted from document chunks and clustered together in order to form topics. They are triplets of the form $(w_s, w_d, \mathbf{w})$, where w_s is the label of the term, w_d is a short sentence giving the term definition, and $\mathbf{w}$ is the term vector representation.

The vector representations $\mathbf{k}$ of document chunks and $\mathbf{w}$ of terms are computed by an embedding model which maps a text into a vector space: for $\mathbf{k}$, the embedding model is applied over the document chunk text k, while for $\mathbf{w}$, it is applied over the term definition w_d. The vector representation $\mathbf{c}$ of topics is computed as the mean of the vectors $\mathbf{w}_i$ of all the terms w_i belonging to c. The label of each topic corresponds to the label w_s of the term w that is the closest to $\mathbf{c}$.

At the beginning of the analysis (i.e., time 0) the user is required to define a set of initial topics $C^{(0)}$ of interest. Each topic $c_i^{(0)} \in C^{(0)}$ is associated with a set of corresponding terms $W_i^{(0)}$, whose definitions are also provided by the user. At each subsequent time t, tASKE performs one or more iterations of the cycle depicted in Figure 2. As a first step, tASKE extracts the set of document chunks $K^{(t)}$ from the subset $D^{(t)} \in D$ belonging to that period. Such document chunks are classified with respect to the topics discovered up to the previous time period $t^{\leftarrow}$, $C^{(t^{\leftarrow})}$. Moreover, tASKE extracts new terms $W^{(t)}$ from the document chunks and assigns them to the topics $C^{(t^{\leftarrow})}$, finally updating the set of current topics $C^{(t)}$.

As a consequence of this process, in ACG a topic $c_j^{(t^{\leftarrow})}$ can have multiple relations with the other components of ACG. In particular, for $c_j^{(t^{\leftarrow})}$, we have i) relation *classification* with document chunks $K^{(t)}$; ii) a relation *derivation* with terms $W_j^{(t)}$ discovered from document chunks associated with $c_j^{(t^{\leftarrow})}$; iii) a relation *belonging* with terms W_j in its cluster; iv) a relation *derivation* with a new topic $c_l^{(t)}$ formed by some of the terms in W_j. It can happen that topic $c_k^{(t^{\leftarrow})}$ is not associated with any document chunk at time t. This means $c_k^{(t^{\leftarrow})}$ is no longer a useful topic with respect to the documents of time t. In this case, the topic $c_k^{(t^{\leftarrow})}$ becomes *inactive*, together with the set of terms W_k belonging to it, and it will not be able to form new topics. This can be interpreted as the disappearance of interest towards a certain topic, which emerged in past periods $t^{\leftarrow}$ but has lost its relevance in the current corpus, $K^{(t)}$.

In the remaining part of this section, we will discuss each phase in Figure 2, explaining in deeper details how each of the aforementioned relations is discovered.

3.1. Data Preprocessing

Preprocessing is the starting point of the tASKE cycle. At each time period $0, \ldots, t$, the model retrieves documents from the period-specific subcorpus $D^{(t)}$.

Documents are first split into document chunks $K^{(t)}$, each of which can fit into the maximum input length of a contextual embedding model. In this case, we employed

Sentence-BERT [12], a modification of the original BERT model, which exploits siamese and triplets networks, being able to derive semantically meaningful sentence embeddings in form of numeric vectors. Such a model is employed in order to extract the semantic features of term definitions and document chunks and map them into the same vector space.

3.2. Zero-Shot Classification

In the zero-shot classification phase, document-topic *classification* relationships are defined. Given the coexistence of topics and document chunk embeddings in the same vector space, it is possible to perform a zero-shot classification, $f : K^{(t)} \rightarrow C^{(t^{\leftarrow})}$, without having the model exposed to training examples. A similarity measure σ (e.g., cosine similarity) between the embedding vector $\mathbf{k}_j^{(t)}$ of each document chunk $k_j^{(t)}$ in $K^{(t)}$ and the embedding vector $\mathbf{c}_i^{(t^{\leftarrow})}$ of each topic $c_i^{(t^{\leftarrow})}$ in $C^{(t^{\leftarrow})}$ is computed and, eventually, the two are associated if their similarity is higher than a predefined threshold α:

$$f_{C^{(t^{\leftarrow})}}(k_j^{(t)}) = \{c_i^{(t^{\leftarrow})} \in C^{(t^{\leftarrow})} : \sigma(\mathbf{k}_j^{(t)}, \mathbf{c}_i^{(t^{\leftarrow})}) \geq \alpha\}$$

Tuning hyperparameter α is crucial since it may remarkably affect the classification output: for example, choosing a high value of α could result in a highly precise classification, despite potentially finding only a small set of document chunks for each topic (low recall).

Finally, *classification* relationships are stored in ACG by considering documents as the simple concatenation of their chunks, so that a document d_j is labelled with all topics its chunks are labelled with.

For example, the document chunk

> *[...] graphical representations of causation have been used for at least seventy years, and the modern development of directed acyclic graphs to portray causal systems continues the trend. It is sometimes difficult to understand, however, what it is about these diagrams that is causal [...]*

is classified by tASKE with the topic *'causality'* with a similarity score of 0.652.

3.3. Terminology Enrichment

For each topic $c_i^{(t^{\leftarrow})}$ in the ACG, tASKE retrieves the set of lemmatized terms $W_i^{(t)}$ appearing in the subset of document chunks $K_i^{(t)}$ associated with $c_i^{(t^{\leftarrow})}$ by a *classification* relation. These terms vectors are placed in the same semantic space, together with $\mathbf{K}$ and $\mathbf{c}$, retrieving their definition w_d from an external knowledge base, such as WordNet [15], and computing its vector representation $\mathbf{w}$ by the aforementioned embedding model. This approach addresses the problem of sense disambiguation, since it maps distinct senses of polysemic words to different embedding vectors.

For each retrieved term sense, the same similarity measure σ used for classification is exploited in order to compute the similarity between $\mathbf{w}$ and the vectors representing topics and document chunks. The terms whose sum of similarities is greater than the hyperparameter β become candidates for enriching the terminology of the topic $c_i^{(t^{\leftarrow})}$:

$$g(c_i^{(t^{\leftarrow})}, W_i^{(t)}, K_i^{(t)}) = \{w^{(t)} \in W_i^{(t)} : \sigma(\mathbf{w}^{(t)}, \mathbf{c}_i^{(t^{\leftarrow})}) + \sigma(\mathbf{w}^{(t)}, \overline{\mathbf{k}_i^{(t)}}) \geq \beta\}$$

where $\overline{\mathbf{k}_i^{(t)}}$ is the centre of the embeddings of chunks in $K_i^{(t)}$.

The set of candidate terms is sorted in descending order according to the similarity score. In addition, one can also define a learning rate γ, which represents the maximum number of terms that can be associated to a certain topic at each iteration. Applying the bounds β and γ ensures that, at each iteration, the process of terminology enrichment will include only a small set of terms that are supposed to be meaningful with respect to the topic at hand.

Taking as example the topic mentioned in the previous section, *'causality'*, it has been associated, among others, with the following terms and similarity scores: *'causality'* (0.773), *'etiologic'* (0.741), *'noncausal'* (0.737).

3.4. Topic Formation

Finally, tASKE may generate new topics in a topic formation phase. In this phase a clustering algorithm, such as Affinity Propagation [16], is applied over the embedding vectors $\mathbf{w}$ of the terms $W_i^{(t)}$ related to each topic $c_i^{(t^{\leftarrow})}$. According to the results, a different operation is enforced:

- *derivation*: if new clusters, different from $c_i^{(t^{\leftarrow})}$ are formed, each of them becomes a new topic, derived from $c_i^{(t^{\leftarrow})}$, whose label is set equal to the term w closer to the cluster center;
- *conservation*: if no new cluster is formed, the original topic $c_i^{(t^{\leftarrow})}$ is preserved, represented by the cluster in which the term w corresponding to the concept label of $c_i^{(t^{\leftarrow})}$ is present;
- *pruning*: if a new cluster $c_j^{(t)}$ is formed but all its member terms belong also to $c_i^{(t^{\leftarrow})}$, the newer topic is absorbed by the older one.

In the end, term-topic *belonging* relationships and topic-topic *derivation* relationships are stored in the ACG

together with document-topic *classification* relationships defined in the zero-shot classification phase, building up the topic genealogy. Topics $C^{(t)}$ defined in this phase will serve as input for the next iteration.

Considering the topic *'causality'*, consisting of the following set of terms {*'causality'*, *'etiologic'*, *'noncausal'*, *'event'*, *'issue'*, *'circumstance'*, *'interpretation'*, *'explanandum'*}, the tASKE model has formed three topics with the corrisponding sets of terms: *'causality'* = {*'causality'*, *'etiologic'*, *'noncausal'*}, *'event'* = {*'event'*, *'issue'*, *'circumstance'*}, *'interpretation'* = {*'interpretation'*, *'explanandum'*}.

4. Case Study and Evaluation

tASKE is here evaluated on a case study on Data Science literature. The evaluation framework has to account for three targets:

1. correctness of extracted topics,
2. correctness of the time of extraction,
3. correctness of topic-topic derivation relationships.

First, a "Data Science in Scopus" corpus (hereon ScopusDS Corpus), made of abstract of journal papers ranging from January 2000 to December 2021, is constructed. Then keywords defined by authors of each paper are exploited to generate a ground truth for all three targets, and our method is evaluated against the ground truth. Finally we perform a brief qualitative analysis of results, which is complementary to quantitative evaluation.

4.1. Corpus Construction

The ScopusDS corpus has been retrieved from Elsevier Scopus by downloading publications in the time interval from January 2000 to December 2021 according to selected subject areas that are concerned with the "data science" subject. For each publication, eid, year, title, abstract, document type, and author-assigned keywords have been downloaded. Furthermore, additional metadata are retrieved (e.g., author name and affiliation, journal/conference name, ISSN, publication type). The corpus content is described in Table 1 in terms of considered subject areas and corresponding number of retrieved publications.

Besides the paper abstract, two pieces of metadata were taken into account in the analysis: the publication date and the list of keywords provided by the author(s). We selected only documents of type "article" that are accompanied by at least 3 keywords and are at least 30 words long, finally amounting to 766,867 documents. Figure 3 shows the number of documents and keywords per year.

ID	Scopus Subject area	# of pub.
1702	Artificial Intelligence	1,024,703
1800	General Decision Sciences	65,254
1801	Decision Sciences (miscellaneous)	39,058
1802	Information Systems and Management	377,259
1803	Management Science and Operations Research	258,898
1804	Statistics, Probability and Uncertainty	168,219
2613	Statistics and Probability	426,341
Total		2,359,732

Table 1
Composition of the ScopusDS corpus used for evaluation

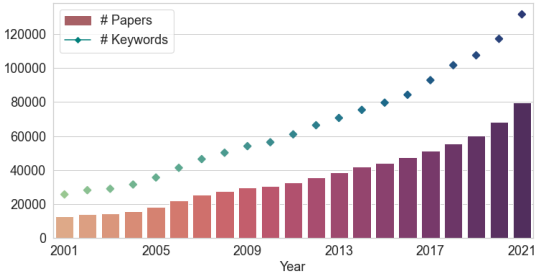

Figure 3: Number of documents and keywords per year in ScopusDS corpus.

4.2. Definition of a Ground Truth

Keywords provided by the authors of each paper are natural candidates to form a ground truth for topic modelling of scientific papers. Exact matching between keywords and extracted topics, however, would yield no significant result, because topics are defined as sets of terms whereas keywords are strings, and author-assigned keywords may not be linked to terms in the external knowledge base employed in tASKE. Hence we define an alternative evaluation methodology which makes use of a non-contextual word embedding model to compute the similarity between keywords and extracted topics.

For target (1), we compare clusters extracted by tASKE with the set of keywords at each time t.

For target (2), we are interested in knowing whether the topics were extracted at the correct time, so we compare clusters extracted at each time with the entire set of keywords. A comparison of the resulting metrics with the ones obtained for target (1) provides an indicator of the timeliness of tASKE extraction: if a topic c, extracted by tASKE at time t is more similar to keywords from time $t' \neq t$ than to the ones from t, then c can be deemed more appropriate to describe the subcorpus at time t' and was extracted either "too soon" or "too late".

Defining target (3) is more complicated, since no genealogical structure is inherently defined on paper keywords. We must first define a set of heuristics to derive a ground truth from the keyword lists assigned to documents. Specifically, we say that a subsequent keyword w' is derived from an antecedent keyword w at time t if:

- w was associated to any document at any time $t^{\leftarrow} < t$;
- w' has never been associated to any document at any $t^{\leftarrow} < t$;
- the number of keyword co-occurrences at t, F_t, is such that $F_t(w, w') \geq 1$.

4.3. Quantitative Evaluation

We run tASKE on the ScopusDS corpus by selecting years as time units in which the corpus is split. Since tASKE requires to be initialized with a set of input topics $C^{(0)} = \{c_1^{(0)}, \dots, c_n^{(0)}\}$, we exclude papers of year 2000 from the evaluation and use the set of keywords assigned to them to derive $C^{(0)}$. This set of terms $W^{(0)}$ is first filtered to retain only terms that appear in WordNet, i.e. the knowledge base used for this evaluation. To avoid the injection of spurious topics into the system, $C^{(0)}$ is further filtered in order to keep only monosemic terms, i.e. terms that are linked to a single WordNet synset, and the 100 with the highest frequency are sampled. In order to retrieve initial topics from this set of terms, we apply Affinity Propagation [16], eventually obtaining $n = 20$ topic clusters, mostly related to mathematics (e.g. *regressions analysis* = {*regression analysis, linear regression, multiple regression*}) and computer science (*internet* = {*internet, information system, bandwidth, world wide web, electronic mail*}), but also to domain of application (*air pollution* = {*air pollution, air transport*}).

As for hyperparameters, we set thresholds α and β equal to one another so to have a single learning rate, and since we found the system to be effective for $\beta \leq 0.35$, the experiments were conducted with $\alpha = \beta = 0.35$ to achieve efficiency in terms of computation time.

To assess the closeness of topics retrieved by tASKE to the ground truth, we train a Word2Vec model on a pseudo-corpus whose documents are a concatenation of document chunks and their ground truth keywords. By exploiting this model, as was done for instance in [17], it is possible to embed keywords and extracted terms in the same vector space. For each year, we define topic embeddings again as the centroids of the embeddings of topic-related terms, which may change from year to year even for the same topic, and we compute cosine similarity between the resulting vectors and the set of keywords. This is done by single linkage, i.e. finding the closest keyword for each topic embedding. Figure 4 reports the mean and the standard deviation of the results for each year.

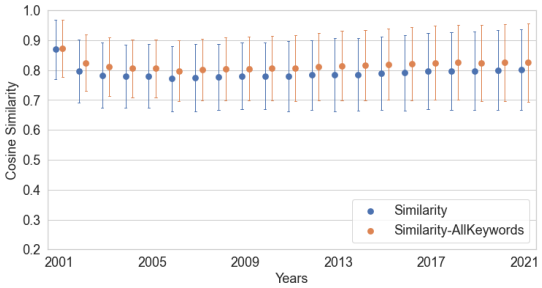

Figure 4: Distributions of similarities between topics extracted in each year and the closest keyword from the same year (blue bars), or the closest keyword from all years (orange bars).

Outcomes displayed in Figure 4 are promising, with a mean similarity going from 86.99% in 2001 ($sd = 9.98\%$) to 80.20% in 2021 ($sd = 13.47\%$), touching a minimum equal to 77.12% ($sd = 10.98\%$) for year 2006. The figure does not prove only the effectiveness of tASKE for target (1), i.e. to discover topics in a corpus, but also for target (2), i.e. to discover them at the proper time. Indeed, at each year, matching with keywords from other years yields better similarities only for few topics per year, as is proven by the overlapping of the similarity distributions.

In the same way as we did for each topic, we can measure the maximal similarity between each keyword and the set of topics in each year, which may be considered a proxy for recall. Resulting similarity distributions, going from 34.77% mean ($sd = 11.28\%$) in 2001 to 65.81% mean ($sd = 12.76\%$) in 2021, are displayed in Figure 5. Although maximising recall was not our main interest, we found that the system gets closer and closer to finding at least a topic for each keyword.

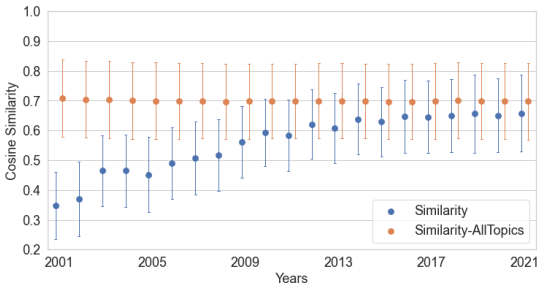

Figure 5: Distributions of similarities between keywords from each year and the closest topic from the same year (blue bars), or the closest topic from all years (orange bars).

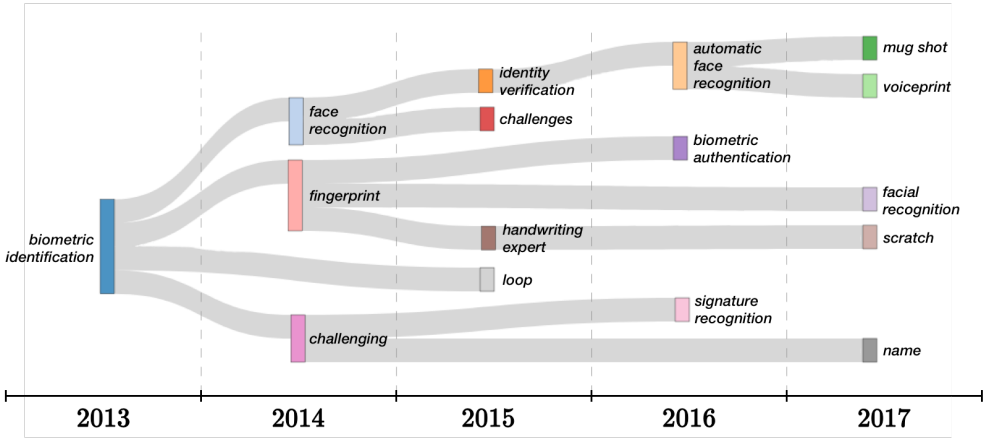

Figure 6: A sample of the final topic genealogy produced by tASKE.

As for target (3), we experimented with the same evaluation method, taking into account the derivation pairs defined in the ground truth, of the type *(antecedent topic, subsequent topic)*, together with the year of derivation. Topic and keyword embeddings are concatenated, forming derivation pair embeddings; similarities are then computed by finding the keyword pair closest to each topic pair.

Results for target (3) are shown in Table 2, both in the case that accounts only for direct derivation relationships $c_i^{(t^{\leftarrow})} \rightarrow c_j^{(t)}$ and for the one in which indirect derivations were considered as well, i.e. $c_i^{(t^{\leftarrow})} \rightarrow c_j^{(t)}$ if $\exists c^{(\tau_1)}, \dots, c^{(\tau_L)}$ with $\tau_1, \dots, \tau_L \in (t^{\leftarrow}, t)$ such that $c_i^{(t^{\leftarrow})} \rightarrow c^{(\tau_1)}$, $c_i^{(\tau_l)} \rightarrow c^{(\tau_{l+1})}$ $\forall l = 1, \dots, L$ and $c^{(\tau_L)} \rightarrow c_j^{(t)}$.

	Mean	Std
Only direct derivations	67.24%	14.51%
Including indirect derivations	69.79%	14.43%

Table 2
Mean and standard deviation of similarities between topic derivation pairs and keyword derivation pairs.

Results are naturally better when indirect derivation relationships are included, but the difference between these and the ones that accounted only for direct relationships is small enough to assume tASKE can find short-term derivations, but has more difficulty in managing long-term ones, likely due to cumulative errors.

4.4. Qualitative Analysis

To grasp the potential as well as the current limitations of tASKE in a broader perspective, we looked at the genealogy it produces and at the topics having low similarity with keywords from the same year, as well as derivation relationships they are involved in. An example of the topic genealogy produced by tASKE is shown in Figure 6.

We noticed that the number of extracted topics tends to grow quadratically in the first iterations, going from 57 (containing 178 terms) in year 2001 to 3039 (with 8950 terms) in 2009, while slowing down at later iterations, reaching 6135 topics and 18189 terms in 2021. This behavior is indicative of the fact that the system accelerates until most common knowledge is retrieved. A surplus of generic topics is produced. Such topics contain few terms and also contribute to lower the similarity with keywords, as most of these belong to domain lexicon. For instance, topics *'diagram'*, *'cast'*, *'fill'*, *'known'*, *'let'*, *'lie'*, *'play'* all have similarity lower than 0.5 with keywords from the same year, and give rise to relationships that further diverge from the domain of interest: from *'play'* to *'toy'* and *'fun'*, from *'diagram'* to *'display'* and *'drafting'*. These are topics that do appear in the form of terms in the ScopusDS corpus, but attention has to be put on the system misinterpreting their meaning or their importance.

tASKE has proved to be capable of capturing some of the topics that marked recent developments or applications in the Data Science domain, such as: *'face recognition'* (2014, from *'biometric identification'*) (as shown in Figure 6), *'speech production'* (2004, from *'wavelet'*),

'search engine' (2004, from *'internet'*), *'ontology'* (2006, from *'knowledge'*), *'clustering'* (2006, from *'class'*), *'natural language processor'* (2008, from *'internet'*), *'graphical user interface'* (2008, from *'internet'*), *'cryptanalytic'* (2010, from *'cryptography'*), *'flight control'* (2012, from *'flight simulator'*), *'machine readable'* (2017, from *'internet'*), *'automatic face recognition'* (2016, from *'face recognition'* through *'identity verification'*). Another category of topics is the one that includes topics of interest but provides a spurious derivation, e.g. *'neural network'* (2006), here derived from *'internet'*, or *'cryptography'* (2008), that descends from *'air pollution'*. An even clearer example of the boundaries the external knowledge base imposes on tASKE is given by the topic *'percolation'*, which may refer to *'clique percolation technique'* in the documents but is here linked to *'air pollution'* due to the absence of any non-physical sense of term *'percolation'* from WordNet. We acknowledged also that most extracted topics are related to domains of application, e.g. medicine, physics, chemistry, social sciences, etc. Including these topics in a hierarchical class structure may prove beneficial to simplify visualization of the topic genealogy.

5. Concluding Remarks

Starting from the increasingly current need to understand the evolution of ideas and research themes in scientific literature, in this work we have presented tASKE, a method for identifying topics in a diachronic corpus of scientific articles. Time in tASKE is a crucial aspect, as the goal is not only to identify the topics in their right temporal collocation, but also to understand how a topic can derive from previous topics, in order to reconstruct the genealogy of the topics in time. tASKE makes it possible to achieve these objectives with an unsupervised approach, i.e., without the need to resort to large and complex pre-annotated datasets. The experimental results, conducted on a corpus of real scientific publications covering a period of 21 years, show how tASKE is able to identify the topics deemed relevant by the authors of the papers and expressed by means of thematic keywords. In particular, the topics identified by tASKE are not only adequate, but also placed in the correct time period and related to each other in a genealogy that described their evolution. Our current and future work on tASKE is aimed at three main goals: i) introduce an adaptive learning rate, with the aim of controlling the number of new topics discovered by tASKE for each time period according not only to the topic relevance but also to the capability of each topic to potentially induce the discovery of new topics in future iterations; ii) make tASKE independent from external knowledge bases, exploiting contextual embeddings, so to avoid restricting a-priori the vocabulary of terms that can be extracted; iii) perform further evaluations both by comparing tASKE with other temporal topic modeling methods and by assessing the quality of topics and their genealogy through the evaluation of domain experts.

References

[1] A. Ferrara, S. Picascia, D. Riva, Context-aware knowledge extraction from legal documents through zero-shot classification, in: R. Guizzardi, B. Neumayr (Eds.), Advances in Conceptual Modeling, Springer International Publishing, Cham, 2022, pp. 81–90.

[2] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Processes 25 (1998) 259–284. doi:10.1080/01638539809545028.

[3] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[4] M. Rabinovich, D. Blei, The inverse regression topic model, in: E. P. Xing, T. Jebara (Eds.), Proceedings of the 31st International Conference on Machine Learning, volume 32 of *Proceedings of Machine Learning Research*, PMLR, Bejing, China, 2014, pp. 199–207.

[5] A. Gruber, Y. Weiss, M. Rosen-Zvi, Hidden topic markov models, in: M. Meila, X. Shen (Eds.), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, volume 2 of *Proceedings of Machine Learning Research*, PMLR, San Juan, Puerto Rico, 2007, pp. 163–170.

[6] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 113–120. doi:10.1145/1143844.1143859.

[7] L. Sun, Y. Yin, Discovering themes and trends in transportation research using topic modeling, Transportation Research Part C: Emerging Technologies 77 (2017) 49–66. doi:https://doi.org/10.1016/j.trc.2017.01.013.

[8] T. M. Abuhay, Y. G. Nigatie, S. V. Kovalchuk, Towards predicting trend of scientific research topics using topic modeling, Procedia Computer Science 136 (2018) 304–310. doi:https://doi.org/10.1016/j.procs.2018.08.284, 7th International Young Scientists Conference on Computational Science, YSC2018, 02-06 July2018, Heraklion, Greece.

[9] A. B. Dieng, F. J. R. Ruiz, D. M. Blei, The dynamic embedded topic model, CoRR abs/1907.05545 (2019). URL: http://arxiv.org/abs/1907.05545. arXiv:1907.05545.

[10] Q. Gao, X. Huang, K. Dong, Z. Liang, J. Wu,

Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec, Scientometrics 127 (2022) 1543–1563. URL: https://doi.org/10.1007/s11192-022-04275-z. doi:10.1007/s11192-022-04275-z.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[12] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, CoRR abs/1908.10084 (2019). URL: http://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[14] M.-W. Chang, L.-A. Ratinov, D. Roth, V. Srikumar, Importance of semantic representation: Dataless classification., in: Aaai, volume 2, 2008, pp. 830–835.

[15] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, Language, Speech, and Communication, MIT Press, Cambridge, MA, 1998.

[16] D. Dueck, Affinity propagation: clustering data by passing messages, University of Toronto Toronto, ON, Canada, 2009.

[17] F. Role, S. Morbieu, M. Nadif, Unsupervised evaluation of text co-clustering algorithms using neural word embeddings, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1827–1830. URL: https://doi.org/10.1145/3269206.3269282. doi:10.1145/3269206.3269282.