

# Identifying Emerging Topics in Specific Domains via Novelty Analysis of Entities in Future Work Sentences from Academic Articles<sup>\*</sup>

Yang Yang, Yi Xiang and Chengzhi Zhang

*Department of Information Management, Nanjing University of Science and Technology 210094 Nanjing, China*

## Abstract

With the exponential growth of academic articles, identifying emerging topics from vast amounts of literature has become a critical task. Currently, researchers employ methods such as bibliometrics and natural language processing to accomplish the task. This study attempts to identify emerging topics by focusing on novel future work sentences. These sentences describe authors' prospects for subsequent research directions and provide a reference for grasping the latest research trends. This study focuses on the field of Natural Language Processing (NLP) and constructs a corpus of future work sentences. We then demonstrate the effectiveness of future work sentences in the identification of emerging topics. Finally, we apply the life-index novelty measurement method to assess the novelty of entities in future work sentences and filter emerging entities based on their novelty and influence. Building on this, we identify emerging research topics in conjunction with the corresponding research tasks of the papers. The results indicate that optimizations and applications of pre-trained language models represent a significant emerging research topic in this domain.

## Keywords

Future work sentence, Emerging research topics, Entity extraction, Academic articles

## 1. Introduction

The identification of emerging topics can provide researchers with valuable insights into future research directions and help funding agencies to optimize the allocation of research funds [1]. Currently, most researchers focus on historical data such as citations to identify emerging topics. However, identifying emerging topics through past topics has a time lag and does not meet the predictive needs of policy makers and researchers [2]. Predicting future research topics is often uncertain, making this task even more challenging.

## 2. Introduction

The identification of emerging topics can provide researchers with valuable insights into future research directions and help funding agencies to optimize the allocation of research funds [1]. Currently, most researchers focus on historical data such as citations to identify emerging topics. However, identifying emerging topics through past topics has a time lag and does not meet the predictive needs of policy makers and researchers [2]. Predicting future research topics is often uncertain, making this task even more challenging.

In the conclusion of academic articles, authors outline their perspectives on future research directions, termed future work sentences. Future work sentences emphasize potential directions for improvement and are more forward-looking compared to titles or abstracts, making them valuable clues for identifying emerging research topics. However, researchers' expressions of future work sentences are either explicit or ambiguous. Some sentences with low reference significance will be

<sup>\*</sup> Joint Workshop of the 2th Innovation Measurement for Scientific Communication (IMSC) in the Era of Big Data (IMSC2024), Dec 20th, 2024, Hong Kong, China and Online

<sup>\*</sup> Corresponding author.

yangyang1221@njust.edu.cn (Y. Yang); xiangyi@njust.edu.cn (Y. Xiang); zhangcz@njust.edu.cn (C. Zhang)

0009-0000-4685-9361 (Y. Yang); 0009-0006-7627-6603 (Y. Xiang); 0000-0001-9522-2914 (C. Zhang)



© 2024 Copyright 2024 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

found when reading them [3]. As illustrated in Figure 1, the paragraph includes four future work sentences. The first sentence highlights the need to improve word prediction accuracy but does not specify methodologies, while the following sentences provide detailed directions for enhancement.

Thus, future work sentences must be screened to enhance understanding of subsequent research. Current research effectively identified future work in papers and summarizes its characteristics by extracting keywords. However, most studies primarily analyzed word frequency in sentences, neglecting to explore whether annual future work sentences can reflect emerging topics of the time [4]. Therefore, this paper takes the field of Natural Language Processing (NLP) as an example to further analyze the specific content of future work sentences, combining emerging entities with research topics to discover emerging research topics.

As future work, first, we plan to improve the precision of word prediction preserving the recall at high. Second, we plan to improve our rewarding model to effectively incorporate translation probabilities and extend the model to reward not only words but also phrases. We will also consider a global constraint by predicting not only target words but their frequencies, and adjust rewards when a word has been used in translation. Finally, more experiments on datasets of various domains and language pairs will be conducted to investigate the generality of our approach.

**Figure 1:** Example sentences for future work sentences<sup>1</sup>

Specifically, we first construct a corpus of future work sentences in the field of NLP, then analyze their effectiveness in identifying emerging research topics through semantic similarity analysis. We also assess how many years' worth of future work sentences can contribute to this identification. Finally, we extract emerging entities from each year's future work sentences using the entity novelty measurement method, enabling us to infer emerging research topics in NLP.

### 3. Related Work

This section provides an overview of related work from two perspectives: the identification and analysis of future work sentences, and the prediction of emerging research topics.

#### 3.1. Extraction and Content Analysis of Future Work Sentences

Current research on future work sentences primarily falls into identification and content analysis. In identification, researchers mainly employ rule-based matching and machine learning or deep learning methods. While rule-based matching achieves high accuracy, its reliance on numerous rules makes it impractical for large academic texts. Consequently, recent studies increasingly utilize machine learning and deep learning models. Hao et al. provided an annotated dataset of future work sentences from the Association for Computational Linguistics (ACL) conference [5]. Zhang et al. expanded this dataset and used it as a training set to train a machine learning classification model for identifying future work sentences [4]. Zhu et al. utilized the BERT model for the automatic extraction of future work sentences [6].

In terms of content mining of future work sentences, Hu and Wan first categorized future work sentences into four types and analyzed the distribution of keywords in different research areas within computational linguistics [7]. Li et al. matched keywords in future work sentences with those in titles and abstracts to explore the conceptual connections between scientific papers and their future work sentences [8]. Hao et al. further categorized future work sentences in the NLP domain into six major categories and seventeen subcategories, analyzing the specific distribution of each type [5]. Qian et al. specifically analyzed the distribution characteristics of six future work

---

<sup>1</sup><https://aclanthology.org/2018.iwslt-1.3/>

sentence types in the field of natural language processing, as well as the focus of future work on different tasks in this field [9]. Zhang et al. analyzed research tasks and hot topics in future work sentences within NLP, and demonstrated the feasibility of using future work sentences to predict future research priorities [4]. Song et al. utilized future work sentences to generate academic innovation topics, offering new insights for technological innovation [10]. Xie et al. examined future work in integrated publishing to explore future research focuses and the evolution of frontier topics, providing valuable references for subsequent studies [11]. Suray et al. analyzed future work sentences from 29 papers presented at the SOUPS symposium, finding most sentences to be broad and vague, with limited impact on subsequent citations [12].

In summary, the identification of future work sentences has become a relatively straightforward task, with machine learning algorithms enabling accurate and efficient recognition. However, in terms of content analysis, current research mainly focuses on the frequency of keywords. Further studies are needed to conduct more fine-grained analyses of future work sentences.

### 3.2. Identifying of Emerging Research Topics

In recent years, identifying emerging topics has become a significant focus in academia, yet researchers lack a unified consensus on the definition and related attributes of this concept. [13]. Rotolo et al. identified several attributes of emerging technologies: radical novelty, rapid growth, coherence, significant impact, and uncertainty [14]. On this basis, Wang provided a comprehensive definition of emerging research topics as those that are novelty, rapidly growing, coherent, and influential [1]. This definition serves as the primary detection criteria for most researchers identifying emerging topics.

The study of emerging research topics dates back to 1965 [15]. Identifying these topics primarily relies on bibliometric and NLP methods. Some researchers applied bibliometric methods based on citation networks, including direct citation, co-citation, and citation coupling networks [16][17][18]. Shibata et al. argued that direct citation network can better discover emerging topics from the perspectives of visibility, speed, and relevance [19]. Kwon et al. demonstrated through direct citations that the emergence of emerging concepts is directly proportional to their future influence [20]. Meanwhile, Boyack et al. verified that using citation coupling for identification the best results from the perspectives of text coupling and network centrality [21].

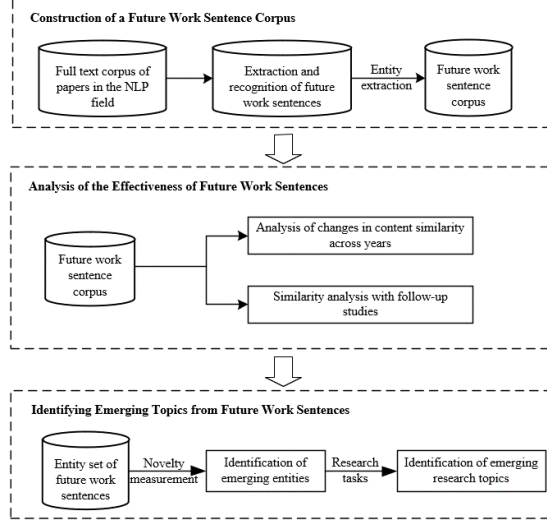
Additionally, researchers utilized NLP techniques for identification. Ohniwa et al. formed emerging themes through co-word analysis of keywords [22]. Liu et al. combined keyword co-occurrence networks, co-citation networks to study emerging research trends [23]. Xu et al. considered four attributes of emerging topics and employed various machine learning methods for emerging topic identification [24]. Ma et al. combined LDA, SAO analysis, machine learning and expert judgement to identify potential development opportunities for emerging technologies [25]. Alattar and Shaalan applied a filtered-LDA model to discover emerging themes [26]. Yang et al. used ecological theories to assess the emergence potential of keywords and identify emerging topics [27]. Wei et al. framed the detection of emerging topics as a cover article prediction problem, using various machine learning methods to predict cover papers [28]. Song et al. proposed a method combining the BERT model with semantic analysis to identify the proportion of emerging technologies [29].

In addition to identifying emerging research topics, researchers have also made efforts to predict future emerging topics. Jung et al. constructed a thematic network to analyze the evolution of themes, enabling them to prospectively predict subsequent research topics [30]. Yang et al. employed an LSTM model to predict the future emerging index of entity features, thereby identifying emerging research topics [31].

In summary, the development of NLP techniques has transcended the limitations of citation-based analysis, enabling researchers to explore more textual features for deeper and more comprehensive insights. Additionally, researchers are making efforts to further predict emerging

research topics in the future. This study broadens the analytical perspective on this task and conducts the analysis by referencing existing emerging indicators.

## 4. Methodology



**Figure 2:** Research framework

This research aims to analyze and summarize emerging entities found in future work sentences to identify emerging research topics in specific fields. The overall research process is illustrated in Figure 2, which is divided into three main steps. The first step involves constructing a future work sentence corpus. The second step is to validate the effectiveness of these sentences in identifying emerging topics. The third part involves calculating entity novelty to identify emerging entities and recognizing emerging research topics in conjunction with specific research tasks.

### 4.1. Construction of future work sentences corpus in the field of NLP

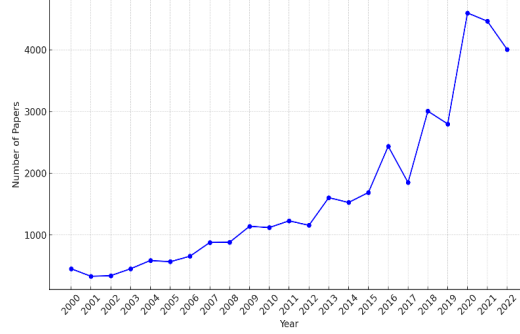
We choose the field of NLP due to its significant role in handling vast amounts of data and advancing artificial intelligence, especially with the emergence of large models like GPT, making it a widely recognized area of research in recent years. Currently, the future work sentence corpus in the field of NLP primarily consists of collections from three highly regarded conferences: the Association for Computational Linguistics (ACL), Empirical Methods in Natural Language Processing (EMNLP), and the North American Chapter of the Association for Computational Linguistics (NAACL) [4]. Identifying emerging research topics in a field requires analysis based on a substantial body of existing research. Therefore, to provide reliable data support for subsequent analyses of future work sentences, we construct a more comprehensive corpus of future work sentences in the NLP field.

#### 4.1.1. Data filtering and acquisition

The ACL Anthology<sup>2</sup> includes major conference and journal papers in Computational Linguistics (CL) and Natural Language Processing (NLP). Some researchers have already built NLP databases based on this site [32]. However, most datasets were constructed earlier and have not been updated in a timely manner. We selected the newly released ACL OCL dataset, which includes the structured full texts of 74,000 academic papers as of September 2022 [33]. This corpus significantly reduces the time cost of data acquisition. Considering factors such as paper quality, type, and language, we select all long and short papers in English from 2000 to 2022 across 46 major conferences, excluding demo and workshop papers. These conferences are authoritative in the NLP

<sup>2</sup> <https://www.aclweb.org/anthology/>

field and provide a comprehensive overview of various research topics and the latest advancements.



**Figure 3:** Changes in the number of academic articles in the NLP field

Additionally, to ensure data completeness, we supplement the corpus with papers published after September 2022 that were initially missing, resulting in a total of 37,791 academic articles. The number of papers published each year is shown in Figure 3. The lower counts in some years compared to the previous year can be attributed to certain conferences not being held that year. Overall, the number of research papers has shown an upward trend, particularly after 2019, when the volume nearly doubled. This indicates the accelerating pace of development in the NLP field in recent years, further highlighting the growing need to quickly grasp the latest research trends.

#### 4.1.2. Identification of future work sentences in the field of NLP

First, we apply a set of rules to initially extract paragraphs likely to contain future work. If a paper has a dedicated Future Work section, all sentences within that section are included as candidates. Otherwise, we only select paragraphs containing any of 42 relevant phrases, such as "In the future", "Future research", or "Future direction". Finally, we segment the selected paragraphs into individual sentences.

To further identify which sentences are future work sentences, we employ the future work sentence identification model developed by Zhang et al. [4]. They trained a future work sentence identification model based on Naive Bayes using over 9,000 labeled samples, achieving the  $F_1$  score of 90.73%. Thus, using this model allows for accurate identification of future work sentences in our dataset. We use the training set they provided and apply the future work sentence recognition model to automatically identify all the selected target sentences. Additionally, to ensure greater accuracy and completeness of the extraction, we manually perform further filtering by removing sentences in the past tense. We also identify sentences starting with pronouns and merge them with the preceding ones.

#### 4.1.3. Fine-grained entity extraction in future work sentences

To thoroughly analyze the content value of future work sentences, we extract fine-grained knowledge entities from these sentences for further research. We utilize the entity extraction model developed by Zhang et al. [34], which focuses on the NLP domain. They randomly selected 50 NLP papers for entity annotation, categorizing entities into four types: methods, tools, metrics, and datasets, as summarized in Table 1. These four types of entities effectively encompass the research content in the NLP field, particularly the method entities, which serve as the main driving force behind advancements in NLP research [35].

Moreover, the entity recognition model they developed is based on SciBERT and employs a cascading binary tagging framework. To enhance the model's performance and robustness, they designed a semi-supervised approach to expand the dataset. This was achieved by matching sentences containing annotated entities from unannotated abstracts, with only those samples

where all matched entities were already annotated being included. An equal number of sample data points were added to the training set, effectively addressing data scarcity.

**Table 1**

A brief description of the four types of entities

| Type    | Description                                   | Example          |
|---------|---|------------------|
| Method  | Algorithms or models to tackle NLP tasks      | SVM, LSTM, BERT, |
| Dataset | Relevant data resources                       | Twitter, WordNet |
| Metric  | Evaluation metrics tailored to specific tasks | Accuracy, BLEU   |
| Tool    | Open-source tools etc. used in the experiment | Python, SQL      |

The entity extraction model constructed using this method outperforms existing baseline models, achieving an  $F_1$  score of 87%, and demonstrates strong performance on the SciERC and TDM open datasets. This indicates that the model is well-suited for entity extraction tasks in the NLP field, fulfilling the needs of our research for extracting entities from future work sentences.

Finally, due to the existence of different expressions for the same entity, we normalize them by constructing an entity normalization dictionary to eliminate their impact on the experimental results. Specifically, we first match high-frequency abbreviated entities with their full forms, while low-frequency abbreviations are identified by searching for their full forms within the corresponding papers. Additionally, we perform lemmatization and remove plural forms. By calculating the similarity between entities, we standardize different expressions of the same entity, determining whether entities with high similarity share the same meaning, and replace them with the most formal full form. Lastly, we replace all entities requiring conversion with their standardized representations.

## 4.2. Analysis of content differences and effectiveness of future work sentences

Researchers generally believe that future work sentences can, to some extent, reflect the development trends of a field. However, there is a lack of empirical studies confirming their role in this regard. Therefore, we aim to analyze whether future work sentences can capture the evolution of research fields by examining the content differences across years and the similarity between future work sentences and subsequent research. This analysis will help determine which years' future work sentences should be used to identify emerging research topics.

### 4.2.1. Analysis of changes in the content of future work sentences

We calculate the similarity differences among future work statements by year , to analyze whether the annual variations in future work content can reflect shifts in research directions. Specifically, we extract the set of knowledge entities contained in all future work sentences, denoted as  $W = \{w_1, ..., w_n\}$ , where  $n$  is the total number of entities. For each year's future work sentences, we construct an  $n$ -dimensional vector, with each dimension corresponding to a knowledge entity. The value of each dimension is determined by the frequency of the corresponding entity appearing in the sentences of that year. After constructing the feature vectors for each year's future work, we calculate the cosine similarity between the future work vector of a specific year and the future work vectors of subsequent years to reflect the differences among the collections of future work sentence across different years. The formula for cosine similarity is as follows:

$$\cos(v_i, v_j) = \frac{v_i \cdot v_j}{|v_i| \cdot |v_j|}, \quad (1)$$

Where,  $v_i$  represents the collection of future work sentences for year  $i$ , and  $v_j$  represents the collection for year  $j$ , where  $i > j$ .

#### 4.2.2. Effectiveness analysis of using future work sentences to identify emerging topics

Furthermore, we analyze the relevance between the content of future work sentences from specific years and the content of abstracts from subsequent research papers. Based on this, we can determine how many years of past future work content are needed to predict subsequent research directions by analyzing the trend of similarity changes. Specifically, we again construct entity-based vector representations for the future work vector  $f_k$  of year  $k$ , and the abstract vector  $a_{k+n}$  of year  $k+n$  following the same steps as before. We then calculate the cosine similarity between the future work sentence vector from year  $k$  and the abstract vector from year  $k+n$ :

$$\cos(f_k, a_{k+n}) = \frac{f_k \cdot a_{k+n}}{|f_k| \cdot |a_{k+n}|} (n > 0) \quad (2)$$

If the similarity is high, it indicates that the future work proposed by researchers in that year is reflected in subsequent research, demonstrating its reference value.

#### 4.3. Identifying emerging research topics through novelty analysis of entities in future work sentences

After confirming that future work sentences can reflect changes in research content, we extract fine-grained entities from the future work sentences of each year for analysis. The rationale for selecting knowledge entities lies in their ability to clearly represent the improvement directions researchers focus on, while also minimizing the noise often introduced by keyword extraction.

Novelty is a critical metric for evaluating emerging research topics. Researchers commonly assess the novelty of academic papers through combinatorial innovation [36]. However, at a fine-grained semantic level, greater emphasis is placed on identifying new content within existing work [37]. We use the life-index novelty measurement to characterize the novelty of each entity [41]. This method introduces the concept of a term life index to characterize the recency of terms. In this paper, the life index of a single entity  $e$  in a future work sentence  $S$  is calculated as follows:

$$Lifeindex(e) = N(e) \times \ln(T_s - T_e + 1) \quad (3)$$

Where,  $T_s$  represents the time when the future work sentence  $S$  was proposed;  $T_e$  denotes the time when the entity  $e$  first appeared in the dataset; and  $N(e)$  indicates the number of times the entity  $e$  appeared in the future work sentence dataset during the period  $[T_e, T_s]$ . The smaller the value of  $Lifeindex(e)$ , the shorter the lifecycle of the entity.

To better represent the novelty of each entity, we define the novelty score for each entity as follows:

$$Lif e_{n(e)} = 1 - \frac{\ln(x_i + 1)}{\ln \max(x_i + 1)} \quad (4)$$

Where,  $x_i$  represents the  $Lifeindex(e)$  value of entity  $e$  in a specific year, and  $\max(x_i + 1)$  is the maximum value of all  $Lifeindex(e)$  values for that entity. We do not consider the case where  $\max(x_i)$  is 0, as this indicates that the entity was mentioned only once across all years, potentially compromising the reliability of the results due to noise.

Next, we filter out the emerging entities for each year based on their novelty. These entities must appear for the first time in a given year and be mentioned in subsequent research to ensure

their continuity and impact. Additionally, we annotate the research tasks in abstracts that contain these emerging entities. Although many research papers have multiple tasks, and some tasks may overlap, our annotation focuses only on the primary and most specific task of each paper. Finally, we analyze the co-occurrence of these entities with research tasks to identify emerging research topics in the NLP field.

## 5. Result

In this section, we present the results of entity extraction from future work sentences, analyze the changes in the content of these sentences and their similarity to subsequent research. Finally, by incorporating the novelty of the entities, we identify emerging entities within the future work sentences and summarize the emerging topics in the field of NLP.

### 5.1. Entity extraction results for future working sentences in NLP

Using the above method, we identify 19,730 academic articles containing future work sentences, accounting for 42.4% of all papers. A total of 45,799 future work sentences are collected, mainly appearing in sections such as Conclusion, Conclusion and Future Work, and Future Work, with most found in the Conclusion section.

**Table 2**

Distribution of entity counts in future work sentences

| Entity Type | Frequency | Ratio |
|-------------|-----------|-------|
| Methods     | 19246     | 67.5% |
| Datasets    | 5547      | 19.4% |
| Metrics     | 2740      | 9.6%  |
| Tools       | 1000      | 3.5%  |

Subsequently, we conduct entity extraction and normalization on the future work sentences. A total of 22,358 future work sentences yield 28,533 extracted knowledge entities. The entity extraction model categorizes these entities into four main types: methods, dataset, metrics, and tools, with the specific distribution detailed in Table 2. Among these, method-related entities are the most prevalent, with a total of 19,246, followed by data-related entities.

**Table 3**

High-frequency entities in four categories

| Methods                      | Datasets  | Metrics    | Tools  |
|------------------------------|-----------|------------|--------|
| BERT                         | WordNet   | Accuracy   | Moses  |
| Language model               | Wikipedia | Precision  | Python |
| Transformer                  | Twitter   | Recall     | OpenIE |
| Neural MT                    | Treebank  | Confidence | SQL    |
| Machine learning-based model | FrameNet  | Robustness | GIZA   |

Finally, we examine the high-frequency entities among the four types identified in the sentences pertaining to future work, with Table 3 illustrating the top five entities by frequency. It is evident that BERT is a prominent topic within the method entities, being referenced with greater frequency than other entities. Language models also receive considerable mentions, as they serve



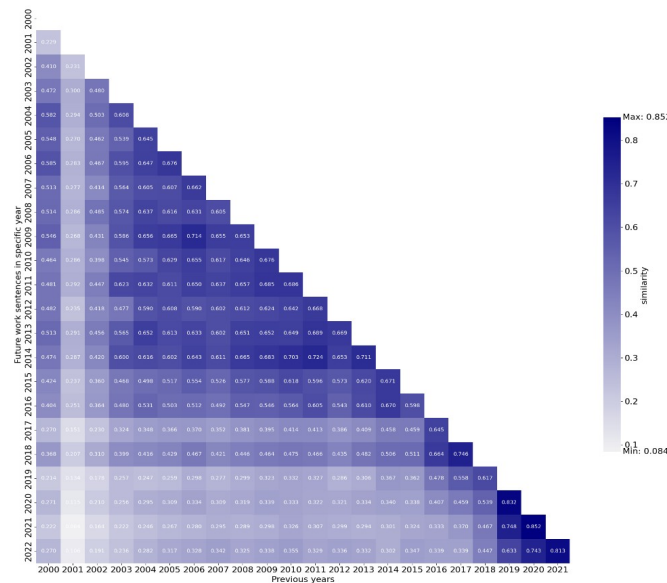
as foundational elements in the NLP field and remain crucial across various developmental phases. They have evolved from early statistical models like N-grams to contemporary pre-trained models like BERT. Regarding dataset entities, researchers primarily concentrate on two categories: one includes datasets related to various domains of knowledge and information, such as Wikipedia and Twitter, which are suitable for practical NLP tasks. The other category consists of datasets primarily containing semantic information and syntactic data, like WordNet, Treebank, and FrameNet, which serve to enhance the performance of language models. Overall, these two categories of data underpin all research tasks in the NLP field, underscoring the importance of data resources in this area.

## 5.2. Results of the analysis of the changes in the content and validity of future work sentences

To analyze whether future work sentences can be used to discover emerging research topics, we consider two aspects: the evolution of future work content and its similarity to subsequent research. On one hand, we can assess whether the content of future work changes each year in alignment with shifts in research trends. On the other hand, we can examine whether future work is reflected in later research, helping us identify which years' future work sentences are valuable for detecting emerging topics. The detailed experimental results are presented below.

### 5.2.1. Results of changes in the content of future work sentences in different years

The results of the content similarity calculations for future work sentences across different years are shown in Figure 4. Overall, there are noticeable differences in future work sentences from different years, with greater disparities observed as the year gap increases. This trend is particularly evident in recent years, especially after 2016, when the future work sentences from that year are primarily similar to those from the previous two years. Notably, the similarity of the future work in 2019 to the previous two years is relatively low. This may be attributed to disruptive developments in the NLP field during 2018-2019, prompting researchers to make significant adjustments in their research outlooks. After 2019, each year's future work exhibits substantial differences from the content before 2019, while the similarity to the previous year has noticeably increased. This indicates a major shift in research directions in the field after 2019, with some research outcomes receiving widespread attention and consensus among researchers.

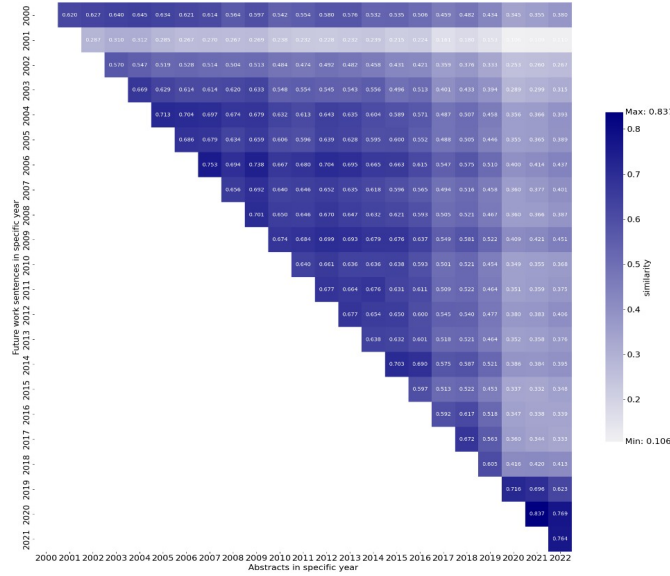


**Figure 4:** Similarity calculation results of future work sentences based on entities

In summary, the NLP field has witnessed some disruptive research outcomes in recent years, capturing the attention of researchers. This shift is clearly evident in the changes in future work, suggesting that using future work sentences to infer emerging research topics is a feasible approach.

### 5.2.2. Analysis of the results of the validity measure of future work sentences

The cosine similarity calculations between the future work sentences of each year and the subsequent yearly abstract collections are shown in Figure 5. There is a certain degree of similarity between the content of each year's future work sentences and the content of subsequent research papers, with the similarity tending to decrease as the year gap increases. This indicates that the field is developing rapidly, making it challenging to predict recent research topics based on earlier future work.



**Figure 5:** Similarity calculation results between future work of specific years and subsequent research content

It is noteworthy that before 2019, the similarity values between each year's future work and the research content of subsequent years do not vary much. However, there is a significant decline in similarity after 2019, indicating that influential new research outcomes emerge in 2018, resulting in substantial differences from previous future work. After 2019, the future work becomes more similar to the subsequent research content. Therefore, to grasp the current emerging research topics in the field of NLP, it is crucial to pay particular attention to the future work produced after 2018.

In conclusion, future work sentences effectively reflect the research trends in the field of NLP. Significant changes in research content occur between 2018 and 2019, resulting in substantial shifts in future work. Therefore, the period from 2018 to 2021 can be used as the time window for identifying emerging research topics in subsequent studies.

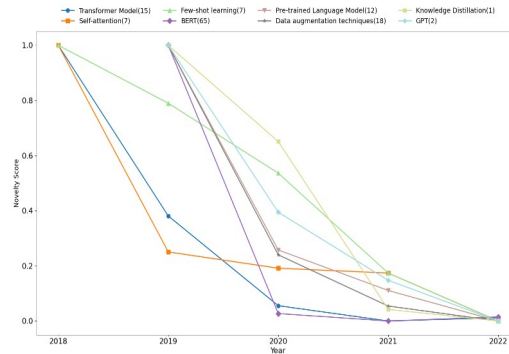
### 5.3. Results of emerging topic identification based on emerging entities in future work sentences

In the previous section, we validated the reliability of future work sentences for identifying emerging topics. In this section, we present the emerging entities found in future work sentences in the NLP field in recent years, as well as the emerging topics.

### 5.3.1. Emerging entities in future work sentences

Based on the above research findings, we filter emerging entities from the future work sentences of the years 2018 to 2021. To ensure the novelty of these entities and their impact on subsequent research, we only select those that first appear in a specific year with a total frequency of at least five occurrences and continue to appear in subsequent years. Statistical analysis reveals that novel entities related to tools, evaluations, and datasets rarely have consistent occurrences. Therefore, we ultimately focus only on novel entities related to methods.

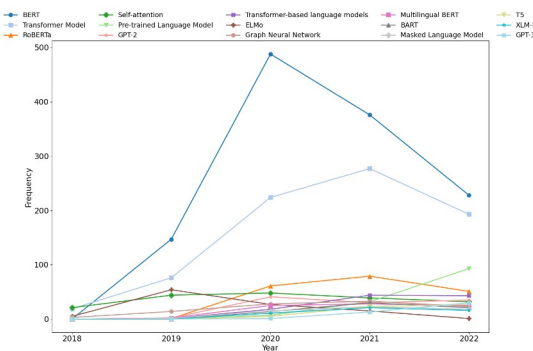
Based on the above screening criteria, we identify emerging entities in future work sentences from 2018 to 2021. Figure 6 illustrates the novelty score changes for some of these entities. It can be seen that a rapid decline in novelty over the years for these entities. This indicates that these entities have been widely referenced in subsequent years, representing current research interests. Notably, the BERT entity, in its first year of appearance in future work sentences, is mentioned in 65 research papers, highlighting its significance.



**Note:** The numbers in parentheses indicate the total frequency of each entity when it first appears in the future work sentences.

**Figure 6:** Changes in novelty scores of some emerging entities in future work sentences from 2018 and 2019

Additionally, we observe a few emerging entities with a decline in frequency in future work sentences, so we analyze their frequency changes in abstracts. Statistics show that 17 of the top 20 entities appearing in abstracts after 2017 are among the emerging entities we selected, indicating our method effectively captures current research trends.



**Figure 7:** Changes in the frequency of certain emerging entities in academic articles

Figure 7 presents the frequency changes of the top 15 emerging entities over the past five years, revealing a gradual increase, which suggests a rise in related research papers. Entities like BERT and Transformer are frequently mentioned, but their mention rates have declined recently, likely due to fewer research papers published since 2020, as well as the emergence of some well-performing model variants, which led researchers to only mention those optimized models instead. Notably, ELMo shows a growth trend in 2018-2019 but declines afterward, appearing only once in

2022. Thus, we conclude that the influence of ELMo has decreased and it should not be regarded as an emerging research entity.

In summary, we consider the novelty and impact of entities in future work sentences to identify emerging entities, as shown in Table 4. We find that most entities relate to research methods associated with pre-trained language models, with the Transformer and BERT models being the most prominent. Many of the subsequent emerging entities evolve from these two foundations.

**Table 4**  
Emerging entities in future work sentences

| Year | Emerging Entities  |
|------|--|
| 2018 | <b>Transformer model</b> (15), Self-attention(7), Few-shot learning(7), Back-translation(3), Meta-learning(2), <b>Generative adversarial network</b> (2), <b>Graph neural network</b> (1), Knowledge graph embedding model(1), <b>Multi-headed attention mechanism</b> (1) |
| 2019 | <b>BERT</b> (65), Data augmentation techniques(18), Pre-trained language model(12), <b>Masked language model</b> (3), Curriculum learning(3), GPT(2), XLNet(2), <b>RoBERTa</b> (1), <b>Knowledge distillation</b> (1)  |
| 2020 | <b>GPT-2</b> (15), <b>T5</b> (6), <b>ALBERT</b> (5), <b>BART</b> (4), <b>GPT-3</b> (4), <b>XLNet</b> (4), DistillBERT(4), Multilingual NMT(3), <b>Non-autoregressive Transformer</b> (1), <b>Multilingual BERT</b> (1), <b>Transformer-based language models</b> (1)       |
| 2021 | Longformer(3), ELECTRA(2), DeBERTa(1), CharacterBERT(1), VisualBERT (1), SpanBERT(1)   |

**Note:** The number following the entity represents the frequency of that entity appearing in that year. The bolded entities are those ranked among the top 20 in frequency within abstracts published after 2017.

### 5.3.2. Results of emerging topic identification

Finally, we filter out the research papers from 2021 and 2022 that contain emerging entities, resulting in a total of 1,629 papers. We annotate the research tasks for each paper. Next, we construct a co-occurrence network of emerging entities and research tasks, as shown in Figure 8.

It can be observed that most emerging entities are concentrated on the analysis and optimization of pre-trained language models, particularly focusing on model compression techniques like knowledge distillation and applying techniques such as data augmentation and post-editing to enhance the transferability of models to specific tasks. Additionally, researchers widely apply pre-trained models to research tasks such as machine translation, sentiment analysis, and named entity recognition, while also optimizing models to meet the demands of specific domains like Healthcare and Finance. Overall, the emerging research topics in the field of NLP primarily center around pre-trained language models. Based on the co-occurrence network and literature review, we summarize the following emerging research topics.

#### (1) Optimization and development of existing pre-trained language models

Since 2018, with the introduction of Transformer and pre-trained language models like BERT and GPT, researchers have evaluated the reliability of these models from various perspectives. To ensure model performance in specific tasks—especially in low-shot, zero-shot, and cross-lingual tasks—researchers have implemented improvements such as few-shot or zero-shot learning techniques to reduce reliance on data annotation. Additionally, model compression methods like knowledge distillation and pruning have been employed to lower the cost of pre-trained models.

These enhancements are widely reflected in tasks such as machine translation, sentiment analysis, and text generation.

## (2) Application of pre-trained language models and large models in other fields

The transferability of pre-trained models to other domains has also been further explored, particularly with large pre-trained models like GPT-3, which possess vast amounts of training data and demonstrate strong transfer capabilities. These models not only excel in NLP tasks but also find effective applications in specific fields such as Social Sciences and Healthcare. Especially, as researchers' interest in applying NLP to the social sciences grows, numerous research tasks, such as those addressing harmful comments, have emerged [39]. Therefore, the success of pre-trained language models drives further advancements in other domains, which in turn stimulates researchers to pursue more in-depth studies on domain adaptability.

## (3) Development of multimodal and multilingual models

Additionally, we observe that recent optimization models integrate more diverse data sources, such as audio and video. These optimized models are widely applied in tasks like machine translation, dialogue, and interactive systems. Therefore, how to combine multiple modalities to overcome the limitations of text data and achieve good results in low-resource language tasks is currently a key concern for researchers in the field of natural language processing.



**Note:** Blue circles represent entities, red circles represent research tasks, and only research tasks that appear five times or more between 2021 and 2022 are shown.

**Figure 8:** Emerging entities and tasks co-occurrence network

## 6. Conclusion and future works

With advancements in technology and the continuous increase in academic articles, the demand for grasping the latest research trends in specific fields is also on the rise. This paper focused on the field of NLP, constructing a corpus of future work sentences within this domain and employing entity extraction techniques to collect a set of entities from these sentences. Based on this, we validated the reliability of future work sentences in recognizing emerging topics by comparing the similarities among future work sentence sets and with subsequent research content. Finally, we filtered the emerging entities for each year based on entity novelty analysis and summarized the emerging topics in the NLP field in conjunction with the research tasks in this domain. The results indicated that pre-trained language models have garnered widespread attention in the NLP field, suggesting that more research and analysis in this area will emerge in the future.

However, this paper also has some limitations. Firstly, relying solely on entity extraction makes it difficult to capture the true semantic information, resulting in somewhat rough predictions. Therefore, fine-grained analysis of the semantic content in future work sentences is needed.

Secondly, we have only considered the novelty of entities in future work sentences; in the future, we will incorporate other content from the text and use entity co-occurrence networks to better assess entity novelty [40]. Finally, we have only identified emerging research topics in the NLP field without evaluating the quality of the research outcomes. Consequently, we will further evaluate the reliability of the research results in the future, such as by using topic modeling for analysis and comparison [41].

## Acknowledgements

This study is supported by the National Natural Science Foundation of China (Grant No. 72074113). The paper is presented at the second Workshop on “Innovation Measurement for Scientific Communication (IMSC) in the Era of Big Data” at 2024 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

## Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to correct grammatical errors, typos, and other writing mistakes. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication’s content.

## References

- [1] Q. Wang, A bibliometric model for identifying emerging research topics, *Journal of the Association for Information Science and Technology*, 69 (2018) 290–304.
- [2] Z. Liang, et al., Combining deep neural network and bibliometric indicator for emerging research topic prediction, *Information Processing & Management*, 58 (2021) 102611.
- [3] S. Teufel, Do "Future Work" sections have a real purpose? Citation links and entailment for global scientometric questions, *Proceedings of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries*, Tokyo, Japan, (2017) 7–13.
- [4] C. Zhang, et al., Automatic recognition and classification of future work sentences from academic articles in a specific domain, *Journal of Informetrics*, 17 (2023) 101373.
- [5] W. Hao, et al., The ACL FWS-RC: A dataset for recognition and classification of sentences about future works, *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, (2020) 261-269.
- [6] Z. Zhu, D. Wang, and S. Shen, Recognizing sentences concerning future research from the full text of JASIST, *Proceedings of the Association for Information Science and Technology*, 56 (2019) 858-859.
- [7] Y. Hu, X. Wan, Mining and analyzing the future works in scientific articles, *arXiv preprint arXiv:1507.02140*, (2015).
- [8] K. Li, E. Yan, Using a keyword extraction pipeline to understand concepts in future work sections of research papers, *Proceedings of International Conference on Scientometrics & Informetrics*, (2019) 87-98.
- [9] Y. Qian, et al., Using future work sentences to explore research trends of different tasks in a special domain, *Proceedings of the Association for Information Science and Technology*, 58 (2021) 532–536.
- [10] R. Song, L. Qian, Y. Du, Identifying academic creative concept topics based on future work of scientific papers, *Data Analysis and Knowledge Discovery*, 5 (2021) 10–20. doi:10.11925/infotech.2096-3467.2020.1275.
- [11] L. Xie, Y. Xiang, C. Zhang, Research on mining future work sentences of academic papers for discovering cutting-edge topics in integrated publishing, *Information Engineering*, 9 (2023) 123–138.

- [12] J. Suray, et al., How the future works at SOUPS: Analyzing future work statements and their impact on usable security and privacy research, arXiv preprint arXiv:2405.20785, (2024).
- [13] S. Xu, L. Hao, X. An, H. Pang, T. Li, Review on emerging research topics with key-route main path analysis, *Scientometrics*, 122 (2020) 607–624.
- [14] D. Rotolo, D. Hicks, B. R. Martin, What is an emerging technology?, *Research Policy*, 44 (2015) 1827–1843.
- [15] D. J. D. S. Price, Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front, *Science*, 149 (1965) 510–515.
- [16] Y. Kajikawa, et al., Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy, *Technological Forecasting and Social Change*, 75 (2008) 771–782.
- [17] H. Small, K. W. Boyack, R. Klavans, Identifying emerging topics in science and technology, *Research Policy*, 43 (2014) 1450–1467.
- [18] M.-H. Huang, C.-P. Chang, Detecting research fronts in OLED field using bibliographic coupling with sliding window, *Scientometrics*, 98 (2014) 1721–1744.
- [19] N. Shibata, et al., Comparative study on methods of detecting research fronts using different types of citation, *Journal of the American Society for Information Science and Technology*, 60 (2009) 571–580.
- [20] S. Kwon, et al., Research addressing emerging technological ideas has greater scientific impact, *Research Policy*, 48 (2019) 103834.
- [21] K. W. Boyack, R. Klavans, Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?, *Journal of the American Society for Information Science and Technology*, 61 (2010) 2389–2404.
- [22] R. Ohniwa, A. Hibino, K. Takeyasu, Trends in research foci in life science fields over the last 30 years monitored by emerging topics, *Scientometrics*, 85 (2010) 111–127.
- [23] Z. Liu, et al., Visualizing the intellectual structure and evolution of innovation systems research: A bibliometric analysis, *Scientometrics*, 103 (2015) 135–158.
- [24] S. Xu, et al., Emerging research topics detection with multiple machine learning models, *Journal of Informetrics*, 13 (2019) 100983.
- [25] T. Ma, et al., Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies, *Technological Forecasting and Social Change*, 173 (2021) 121159.
- [26] F. Alattar, K. Shaalan, Emerging research topic detection using filtered-LDA, *AI*, 2 (2021) 578–599.
- [27] J. Yang, et al., A novel emerging topic detection method: A knowledge ecology perspective, *Information Processing & Management*, 59 (2022) 102843.
- [28] W. Wei, H. Liu, Z. Sun, Cover papers of top journals are reliable source for emerging topics detection: A machine learning based prediction framework, *Scientometrics*, 127 (2022) 4315–4333.
- [29] B. Song, C. Luan, D. Liang, Identification of emerging technology topics (ETTs) using BERT-based model and semantic analysis: A perspective of multiple-field characteristics of patented inventions (MFCOPs), *Scientometrics*, 128 (2023) 5883–5904.
- [30] S. Jung, R. Datta, A. Segev, Identification and prediction of emerging topics through their relationships to existing topics, in *2020 IEEE International Conference on Big Data (Big Data)*, (2020) 5078–5087.
- [31] Z. Yang, W. Zhang, Z. Wang, et al., A deep learning-based method for predicting the emerging degree of research topics using emerging index, *Scientometrics*, 129 (2024) 4021–4042. doi:10.1007/s11192-024-05068-2.
- [32] Bollmann, Marcel, et al. Two decades of the ACL Anthology: Development, impact, and open challenges, *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, (2023) 83–94.
- [33] S. Rohatgi, et al., The ACL OCL corpus: Advancing open science in computational linguistics, arXiv preprint arXiv:2305.14996, (2023).

- [34] H. Zhang, C. Zhang, Y. Wang, Revealing the technology development of natural language processing: A scientific entity-centric perspective, *Information Processing & Management*, 61 (2024) 103574.
- [35] A. Pramanick, et al., A diachronic analysis of paradigm shifts in NLP research: When, how, and why?, *arXiv preprint arXiv:2305.12920*, (2023).
- [36] B. Uzzi, et al., Atypical combinations and scientific impact, *Science*, 342.6157 (2013) 468-472.
- [37] T. Ghosal, et al., Novelty detection: A perspective from natural language processing, *Computational Linguistics*, 48.1 (2022) 77-117.
- [38] Z. Luo, et al., Combination of research questions and methods: A new measurement of scientific novelty, *Journal of Informetrics*, 16 (2022) 101282.
- [39] A. Pramanick, et al., The Nature of NLP: Analyzing Contributions in NLP Papers, *arXiv preprint arXiv:2409.19505*, (2024).
- [40] Z. Wang, et al., Content-based quality evaluation of scientific papers using coarse feature and knowledge entity network, *Journal of King Saud University-Computer and Information Sciences* 36.6, (2024) 102119.
- [41] P. Savov, et al., Identifying breakthrough scientific papers, *Information Processing & Management* 57.2, (2020) 102168.