Advances in Technology Innovation, vol. 9, no. 3, 2024, pp. 224-238

Current Trends in Named Entity Recognition from Automatic Speech Recognition: A Bibliometric Analysis Using Scopus Database

Thu Hien Nguyen[1], Tuan Linh Nguyen[2], Thanh Binh Nguyen[3,*]

[1]Faculty of Mathematics, Thai Nguyen University of Education, Thai Nguyen, Vietnam

[2]Faculty of Electronics Enginee, Thai Nguyen University of Technology, Thai Nguyen, Vietnam

[3]Faculty of Physics, Thai Nguyen University of Education, Thai Nguyen, Vietnam

Received 23 April 2024; received in revised form 08 July 2024; accepted 09 July 2024

DOI: https://doi.org/10.46604/aiti.2024.13619

Abstract

Named entity recognition (NER) is critical for language understanding and text mining systems, such as event extraction and automatic question-and-answer systems. However, NER from automatic speech recognition (ASR) outputs remains challenging due to errors and lack of textual cues. This study aims to provide a comprehensive bibliometric analysis of research on NER from ASR, focusing on publications indexed in the Scopus database before 2024 to understand the research field. Using Biblioshiny and VOSviewer tools, this research identifies the key trends, prominent authors, and international collaborations in the research network. The results show steady growth in this research area, while conference papers are the predominant source type. Additionally, the study highlights the increasing intervention of deep learning approaches to enhance NER accuracy, suggesting potential research directions to reduce error rates, and developing more robust NER algorithms. Finally, the findings underscore the importance of cross-disciplinary collaborations to document any current challenges.

Keywords: named entity recognition, automatic speech recognition, bibliometrics, Scopus, potential research direction

1. Introduction

Named entity recognition (NER) is identifying named entities from free-text documents and classifying them into predefined types such as person names, organizations, and locations [1]. In 2011, the Quaero project proposed an extended definition of entity identification, where basic entities are combined to identify more complex entities. For example, the organization name is further categorized into government organizations, educational institutions, or commercial organizations [2]. Automatic speech recognition (ASR) is defined by Yu and Deng [3] as the processes, technologies, and methods that enable better human-computer interaction through the translation of human speech into text format. NER from the ASR output text is more challenging than written text. This difficulty arises because the ASR output often contains numerous errors (insertions, deletions, and word substitutions) and lacks some important indicators for NER, such as capitalization and punctuation. Furthermore, the scarcity of large, standardized speech data sources for training purposes demonstrates the challenges [4].

In ASR systems, the NER information illustrates the significant meaning in information extraction systems (Fig. 1) and is useful in various applications such as optimizing search engines, content categorization for news providers, and content recommendations. Sometimes, NER from speech is also used for privacy support applications, such as concealing patient

* Corresponding author. E-mail address: binhnt@tnue.edu.vn

English language proofreader: Yen-Chun Hsieh

names in healthcare. Some companies employ NER systems to detect negative customer feedback. In addition, applications like Netflix, YouTube, and Facebook rely on NER to provide recommendations based on user search history [5]. Until now, most research on NER from ASR (NER–ASR) has traditionally followed the pipeline approach. However, errors propagated through the different stages in this method might directly affect the performance of the NER system. To minimize the disadvantage generated by pipelines, researchers recently have been exploring an end-to-end approach to directly label named entities from the ASR system [6].

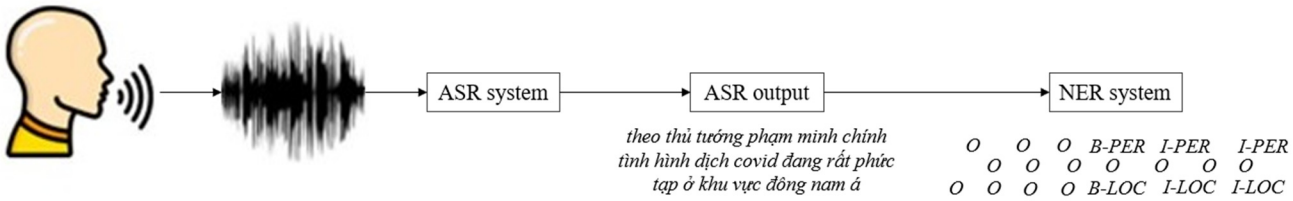


Fig. 1 NER system from ASR

Various methods have been proposed in recent decades to solve the NER–ASR problem. However, both approaches still documented the difficulties. Research to gain a comprehensive overview of this issue worldwide is necessary. This research employs an effective method using bibliometrics to measure the quality of research by statistically analyzing quantitative data from scientific publications [7]. Bibliometrics refers to the mathematical and statistical methods for monitoring and analyzing the structure of a scientific field, identifying research areas and trends, evaluating research development, and examining patterns related to regional and authorship characteristics in publications and citations [8]. The investigation from bibliometrics might point out the current research trends and enhance research quality for innovative future research.

The overview studies for natural language processing (NLP) have also utilized bibliometrics as a statistical method [9-10]. Among these studies, only research on NER for Bahasa and Indonesian languages for regular text [11]. The data was only collected from SCOPUS, ACM, IEEEXplore, and Science Direct within a short time frame, the past five years, from 2016 to 2021. Moreover, to the best of the author's knowledge, there is currently no comprehensive bibliometric study on NER for ASR, which provides a holistic view of the development of research in this area worldwide and identifies key research directions, opportunities, and challenges for the future.

To continue with the bibliometric analysis research for NER–ASR and address the shortcomings of previous studies, this research will undertake the important aspects, including offering a standardized five-step methodological framework for conducting quantitative scientific studies in this field; conducting a comprehensive bibliometric analysis of NER–ASR publications indexed in the Scopus database before 2024; identifying and visualizing research trends, prominent authors, and international collaborations using Biblioshiny and VOSviewer tools; providing insights into the challenges and limitations of NER–ASR, emphasizing the importance of deep learning approaches; Finally, this work proposed potential research directions and highlighted the significance of cross-cultural and interdisciplinary research collaborations.

2. Methodology

The research will be conducted by applying a scientific mapping process [12], which consists of five stages including (1) Research design; (2) Data collection; (3) Data analysis and visualization; (4) Result interpretation.

(1) Research design: The research design stage has been directed by identifying the main research questions: How have the studies been distributed over the past ten years? Who are the most prominent authors in this field? Which countries have shown the highest productivity and activity in this area? What are the research trends of NER from speech? What are the approaches, research methods, and data used in the publications?

(2) Data collection: The data collection phase consists of three separate steps: collection, filtering, and cleaning.

Step 1: Collection

This task performed a search using the Scopus database (http://www.scopus.com), utilizing the advanced search options to input search conditions and appropriate operators based on the syntax of this search tool. The Scopus database was used for this bibliometric analysis due to its large amount of indexed documents to the Web of Science and Dimensions.

The data query was executed on January 2, 2024. Keyword selection is an important factor in the article retrieval process, and the possibility of incorrect or incomplete keywords can skew the results. Therefore, the keywords have been carefully selected based on the advice of specialized research and experts, reflecting key concepts in NER for ASR. The identification of relevant keywords for NER includes: “named entity recognition,” “named entity extraction,” “named entities,” and “name entity.” For ASR, the keywords were “automatic speech recognition,” “speech recognition,” “speech recognizer,” “speech processing,” and “speech data" and found rational 885 papers. Additionally, the search scope was limited to introduce only English-language documents in Computer Science, including scientific articles and conference papers. The filtering result was 797 documents.

Finally, research restricted the search to include only documents published before 2024, with the keywords appearing in the title and abstract. The data investigation statement to Scopus was: TITLE-ABS (("Name* Entit*") AND ("Speech Recognition*" OR "Speech Processing" OR "Speech Data")) AND PUBYEAR < 2024 AND (LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English")). The filtering data consisted of 244 documents.

Step 2: Filtering

The data filtering process involved a systematic approach to ensure the final dataset consisted of relevant and high-quality articles. The process of reading and filtering unrelated articles was conducted thoroughly in two rounds: independent evaluation of titles, abstracts, and keywords by each researcher to eliminate documents which not directly relevant to the research topic, and followed by a group review to achieve the consensus and resolve any discrepancies. This meticulous approach resulted in the final selection of 55 appropriate documents deemed directly relevant to the mentioned problem, enhancing the accuracy and reliability of the bibliometric analysis.

Step 3: Cleaning

This step has been conducted by addressing inconsistencies in certain information within the obtained dataset, such as author names and affiliations. The following quest was Data analysis and visualization: Various analytical techniques were applied to extract information from the collection of publications. General information about the published collection was summarized, and the annual publication count was analyzed to identify trends in the research field. Keyword analysis techniques were used to identify research trends in the field.

(3) Data analysis and visualization: For data analysis, several widely used open-source scientific bibliometric tools were utilized including CiteSpace, Science of Science (Sci2) Tool, BibExcel, CoPalRed, Workbench Tool, BiblioTools, VOSviewer, SciMAT, CitNetExplorer, Biblioshiny, among others. In this study, authors have employed VOSviewer (version 1.6.20) and Biblioshiny (version 4.0) to identify and visualize collaboration networks between authors and countries and to identify trends using keywords.

(4) The interpretation of the results is presented in Section 3.

3. Results and Discussion

According to Table 1, a total of 55 documents from Scopus, in which 7 journal articles and 48 conference papers have been listed. These documents were published in 35 different sources before 2024. The steady growth rate indicates sustained interest in ASR and NER. The total number of citations was 497 times, corresponding to an average of 9.036 citations per

document, indicating the significance and impact of the research. A total of 176 authors have contributed to these documents. Interestingly, single-authored works are rare, with only 1 such instance. Collaborative efforts are more common, with an average of 3.69 co-authors per document. 10.91% of the collaborations involve international authors. This global collaboration fosters cross-cultural perspectives and knowledge exchange.

Table 1 Investigation of articles on NER–ASR published in Scopus before 2024

Attribute	Number of magnitudes
Sources (Journals, Books, etc)	35
Documents	55
Article	7
Conference paper	48
Annual growth rate	4.49%
Document average age	11.4
Average citations per document	9.036
References	1179
Keywords plus (ID)	357
Author's keywords (DE)	92
Authors	176
Single-authored (with no co-author)	1
Co-Authors per document	3.69
International co-authorships	10.91%

The first publication on NER–ASR was introduced in 1998 in the proceedings of the annual meeting of the Association for Computational Linguistics [13]. There challenges in researching NER–ASR can be attributed to various factors. Firstly, the output text from ASR systems often lacks structure, such as punctuation marks, capitalization or proper nouns, and location names. This leads to difficulties in understanding and limits the ability to exploit the ASR output text for applications. So, the Automatic NER–ASR output text always contains recognition errors, especially with out-of-vocabulary (OOV) named entities. Additionally, ASR errors often occur in the constituent words of named entities or the context of those words, directly affecting the performance of NER. Furthermore, NER systems have to address issues related to the lack of important cues such as capitalization and punctuation marks. In particular, the scarcity of sufficiently large labeled speech datasets for training NER models is one of the most significant challenges in research.

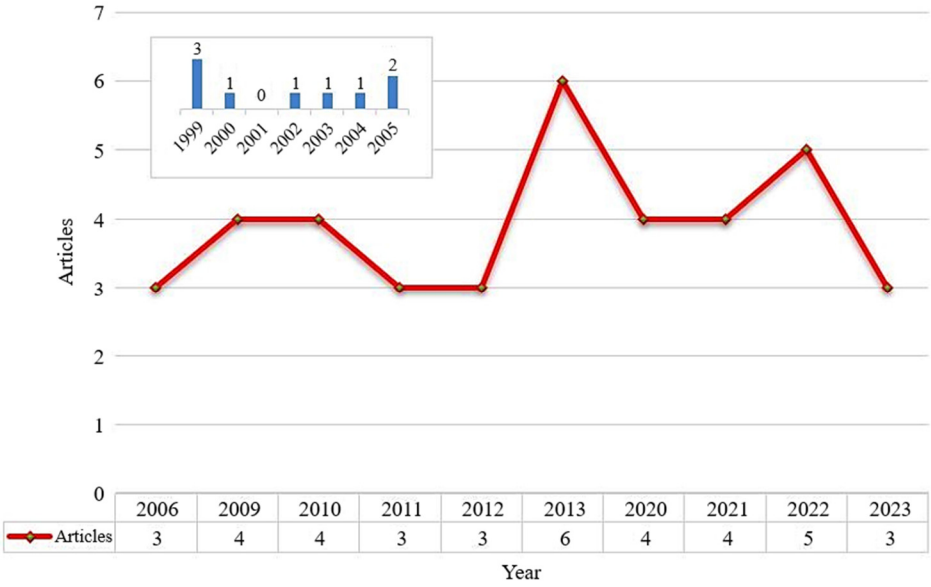


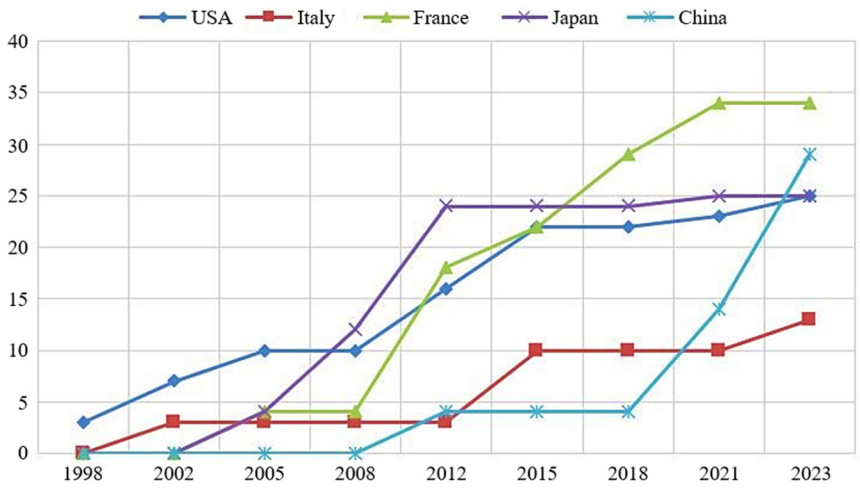
Fig. 2 Annual scientific publication on NER–ASR

Until 2005, there was minimal attention to this topic, resulting in very few publications. The bar chart within the graph in Fig. 2 likely corresponds to this early period, showing a low number of articles per year. Starting from 2006, there has been a significant surge in NER-related publications with an average of around 3-6 publications per year. The graph's red line shows

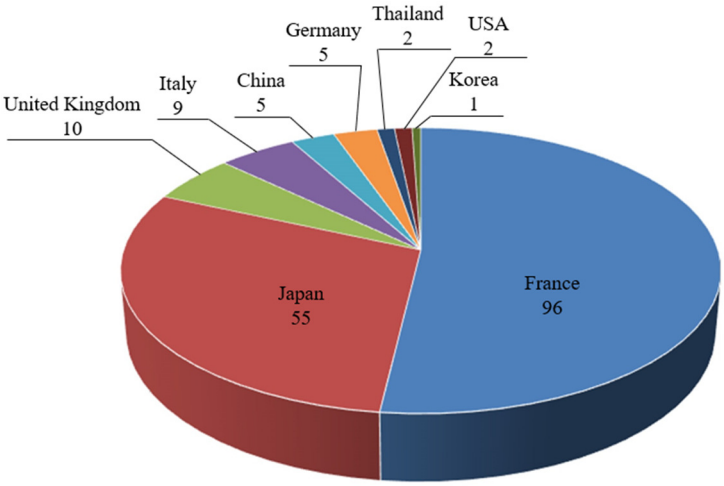
a gradual increase in the number of articles per year. Researchers have increasingly explored NER in the context of ASR, leading to a growing body of literature. The surge in publications likely reflects advancements in ASR technology. Researchers' increased interest in NER applications has contributed to this trend. NER–ASR is gaining prominence due to its relevance in various domains.

The peaks in the line (around 2010, 2013, and 2022) indicate periods of heightened research output. In 2013, the field of NLP witnessed significant advancements. Researchers began adopting deep learning techniques, including neural networks, for various NLP tasks. This shift likely influenced ASR research, leading to improved NER methods. In 2022, pre-trained language models (such as BERT, T5, and their variants) had gained prominence. These models demonstrated exceptional performance across NLP tasks, including NER. The fluctuations between peaks represent variations in annual publication numbers. Despite fluctuations, the overall trend is upward. The line representing the number of articles suggests sustained interest and progress in NER research from ASR.

Fig. 3 shows the number of publications by different countries over time. Fig. 3(a) highlights the global interest in NER research within the ASR domain and shows that France consistently has the highest number of publications, followed by China, Germany, Japan, and the United States. The upward trend suggests a growing interest in NER–ASR research globally. Recent research on the end-to-end approach for NER–ASR also follows studies for French [14], English [15], and China [16]. Fig. 3(b) highlights the countries with the most citations for their NER–ASR work. France and Japan were the first and second-largest segments reflecting significant citation impact, followed by the United States, China, Germany, and the United Kingdom, respectively. These countries likely produce influential research or have well-established research networks.



(a) Five highest publication countries



(b) Most cited countries on NER–ASR before 2024

Fig. 3 Please add captions for figures.

The combination of both graphs suggests that while several countries actively produce NER-related research, France stands out in terms of both quantity and impact. France's dominance in both production and citations indicates a strong research ecosystem. The concentration of publications in France, China, and Japan may be attributed to several factors:

(1) Rich data sources: These countries may have access to diverse and extensive speech data, enabling comprehensive research.

(2) Collaboration networks: Collaborations among researchers in these countries could lead to more impactful work.

(3) Research infrastructure: Well-established research institutions and funding support contribute to their prominence. Researchers worldwide can learn from France's success in NER–ASR. Collaboration across borders can enhance research quality and impact. Access to rich data remains crucial for advancing NER–ASR studies.

The receiver operating characteristic (ROC) curve is an important tool in evaluating the performance of classification models. In the bibliometric field, the ROC curve can be used to evaluate the ability of a certain model or index to classify scientific articles into "highly cited" and "undercited" groups of many leads. The calculated number of citations and CiteScore for each document is shown in Table 2.

Table 2 Statistics on the number of citations and CiteScore for documents

Documents	Publication year	Number of citations	CiteScore
D1	2018	59	2.000
D2	2005	48	3.587
D3	2000	36	0.600
D4	2004	35	3.684
D5	2013	29	0.256
		⋮	
D55	2022	0	1.973

Fig. 4 shows the ROC curve based on the CiteScore classification. The model's performance, indicated by an area under the curve (AUC) of 0.63, demonstrates a superior classification ability compared to a random model. However, this performance is not very high and requires the collection step to gather more data from various databases to improve the analysis process and thereby enhance the ROC performance.

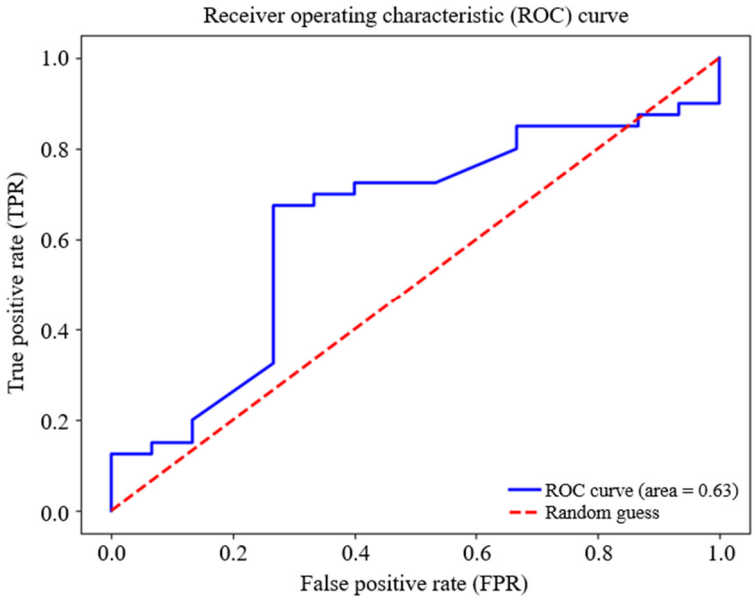


Fig. 4 The ROC curve plotted against the FDR based on CiteScore

Fig. 5 illustrates the international collaboration network research of NER–ASR which consists of 27 countries. In each country, the strength of authorship will be calculated by the co-authorship links with other countries. The thickness of these links may indicate the strength of collaboration. The network's structure reflects global research partnerships in NER from

speech. Collaborations across clusters may lead to knowledge exchange, and advancements, and can guide future research initiatives and foster international cooperation. These clusters likely represent regions with strong research ties and collaborative efforts in NER from speech. The results show that the network can be divided into two main clusters or groups:

(1) The United States, France, Qatar, Japan, and India. The United States, as a central node, plays a significant role in NER research. France, Qatar, Japan, and India are closely connected to the United States, suggesting active collaboration. These countries likely share expertise, resources, and research findings. Joint projects, conferences, and knowledge exchange may be common. The United States plays a significant role, possibly leading in NER research.

(2) Germany, Canada, and Switzerland form another cohesive group. Their collaboration may involve joint projects, conferences, or shared research interests. The proximity of these countries in the network indicates their strong ties.

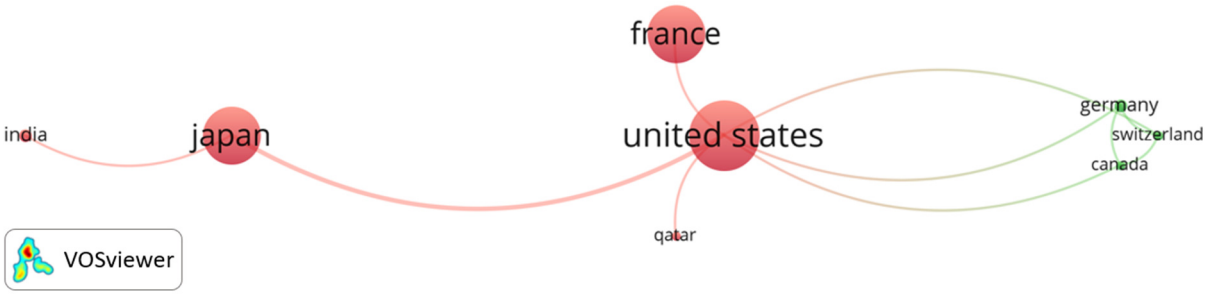


Fig. 5 Significant collaborative relationships and partnerships among the countries on the research of NER–ASR

Fig. 6 presents an analysis of the top 10 sources in the field of NER–ASR. These sources comprise eight conference proceedings and two journals. These sources play a crucial role in shaping research and advancing knowledge in NER–ASR. The chart visually represents each source's relevance and impact, with the distance from the center indicating its significance. Notably, sources closer to the outer edge hold greater relevance and influence in NER–ASR research.

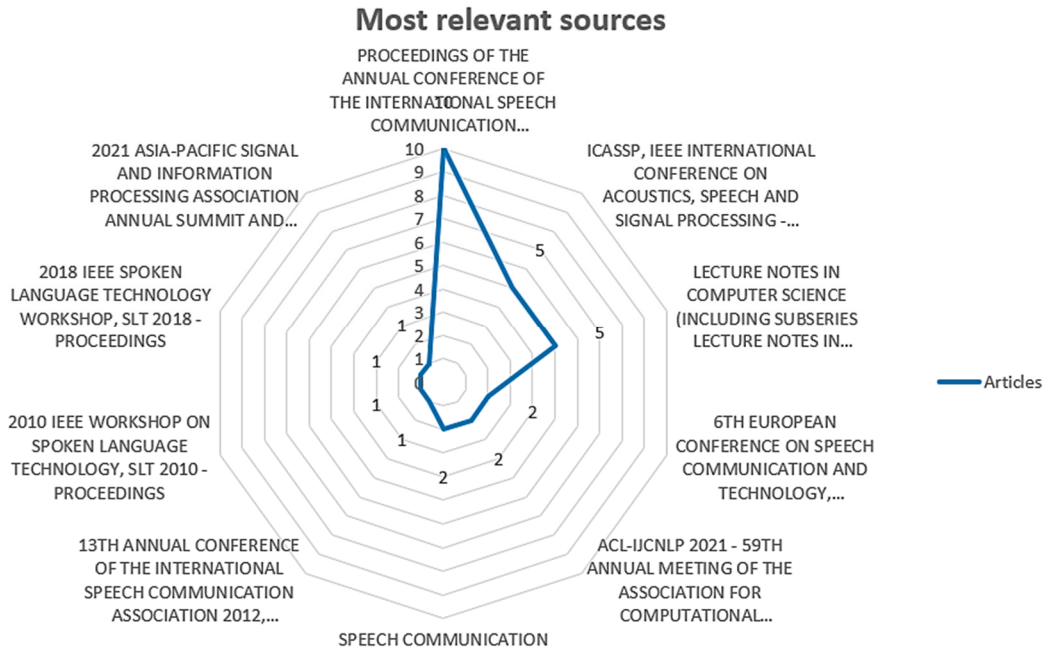


Fig. 6 Top 10 sources ranked by the output published by them on NER–ASR

Specifically, the proceedings of the annual conference of the International Speech Communication Association (Interspeech) contain 10 articles. As a leading conference in the field, Interspeech serves as a platform for researchers to present their findings and exchange ideas. The high number of articles indicates its significance in NER–ASR research. proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) feature 5 articles related to NER–ASR. ICASSP is a prestigious conference where experts discuss cutting-edge research in speech and signal processing. Its inclusion underscores its impact on NER–ASR advancements.

Notably, scholarly publications in NER–ASR predominantly appear in conference and workshop proceedings. These venues provide a fertile ground for researchers to share their latest findings, collaborate, and contribute to the field. Researchers often present their work at conferences like Interspeech and ICASSP, leading to a rich body of knowledge in NER–ASR. A total of 55 documents in the investigated dataset were contributed by 176 authors. Table 3 reveals the top 10 highest in terms of publications, along with affiliations and country. The authors' names are ranked from 1 to 10 based on their NER–ASR publications.

Table 3 Top 10 authors ranked by the number of publications on NER–ASR and their affiliations

Rank	Authors	Affiliation	Total publications	Total citations	Total publications/ total citations
1	Béchet F.	University of Avignon, France	3	87	3.4
2	Isozaki H.	NTT Communication Science Labs, Japan	3	29	10.3
3	Morin E.	University of Nantes, France	3	80	3.8
4	Rosset S.	University of Paris-Saclay, France	3	50	6.0
5	Sudoh K.	NTT Communication Science Labs, Japan	3	29	10.3
6	Tsukada H.	NTT Communication Science Labs, Japan	3	29	10.3
7	Glotin H.	Université du Sud Toulon-Var, France	2	9	22.2
8	Itoh N.	IBM Research - Tokyo, IBM Japan	2	14	14.3
9	Kim J. H.	Sogang University, Seoul, South Korea	2	37	5.4
10	Kurata G.	IBM Research - Tokyo, IBM Japan	2	14	14.3

Béchet F. has 3 publications and is affiliated with the University of Avignon, France. They have received 87 citations, indicating significant impact. Morin E. also stands out with 80 citations, demonstrating substantial influence. Isozaki H., Sudoh K., and Tsukada H., all affiliated with NTT Communication Science Labs, Japan. These authors, affiliated with NTT Communication Science Labs, have each contributed three publications. Their moderate citation count of 29 suggests active research within a focused group. Itou K., Kurata G., Glotin H., and Kim J. H. have two publications each. Kim J. H. received 37 citations, indicating impactful work from Seoul National University in South Korea. The remaining two authors have lower citation counts (14 and 9), but their contributions are valuable.

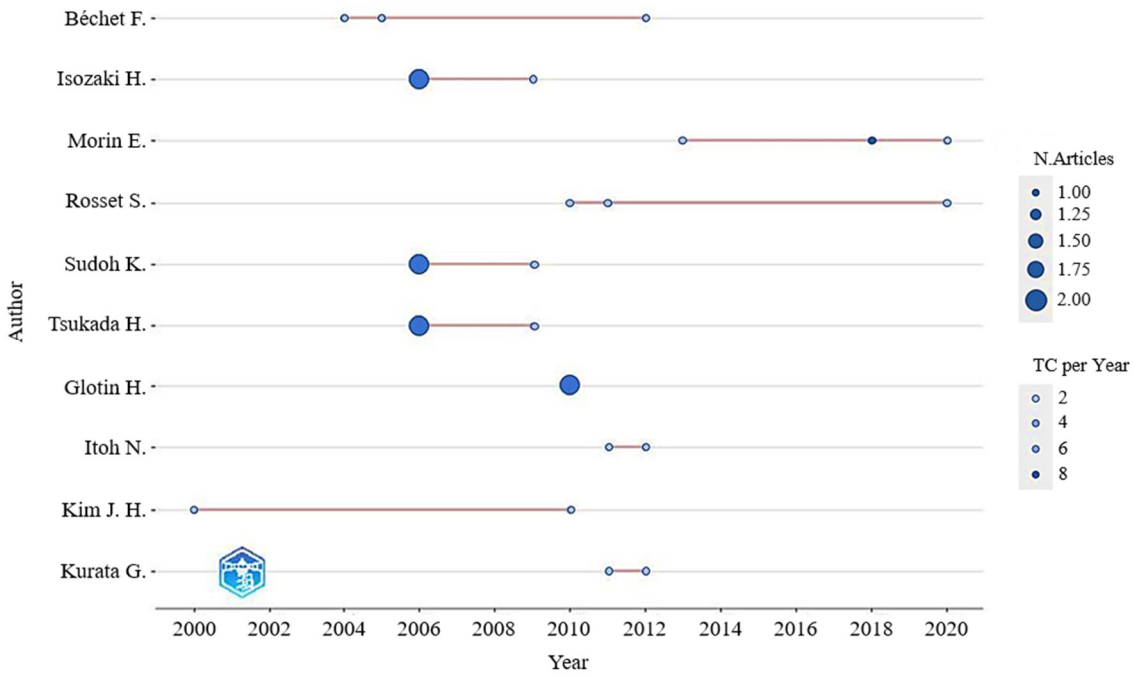


Fig. 7 Top 10 most prolific authors' publication production in the mentioned period

The collaboration among authors reveals a mix of small research groups. Notably, two main clusters emerge—one from Japan and the other from France. While collaboration within these groups is evident, the overall diversity remains relatively low. Future research could explore ways to foster cross-group collaboration and enhance diversity in NER–ASR studies. The

analysis of Fig. 7 provides valuable insights into the annual publication trends of the top 10 authors in the field of NER–ASR from 2000 to 2022. The visualization effectively represents the publication output of each author through the use of node size, providing an immediate understanding of their productivity. By representing collaborations through connecting lines, the visualization provides insight into the collaborative nature of the field. Thicker lines indicate more collaborative works, emphasizing the importance of collaborations among researchers in NER–ASR. Collaborative efforts can lead to the exchange of ideas, the pooling of resources, and the development of more comprehensive research outcomes.

Isozaki H., Sudoh K., and Tsukada H. stand out as the most prolific authors, as indicated by their larger circles. This suggests that they have made significant contributions to the field and have produced a substantial amount of research. However, it is noteworthy that their publications are concentrated within a short timeframe. This concentration of publications could indicate a period of intense research activity or specific projects that resulted in a high number of publications. In contrast, Béchet F. demonstrates consistent research efforts throughout the years, as indicated by a steady publication output. This signifies a sustained level of productivity over an extended period, suggesting a long-term commitment to NER–ASR research.

The visualization also highlights emerging authors Morin E. and Rosset S., whose publication circles are growing over time. This indicates their increasing productivity and suggests that they are actively contributing to the field. The growing circles of these authors may indicate a rising influence and potential for becoming influential figures in the future of NER–ASR research.

In Fig. 8, the analysis of keyword criterion, the unrelated keywords such as “state of the art,” “semantics,” “character recognition,” “computational linguistics,” “text processing,” and “broadcast news” have been excluded. For each year, the author has selected three keywords to represent the annual research trends, each keyword should appear at least five times to meet the criteria requirements.

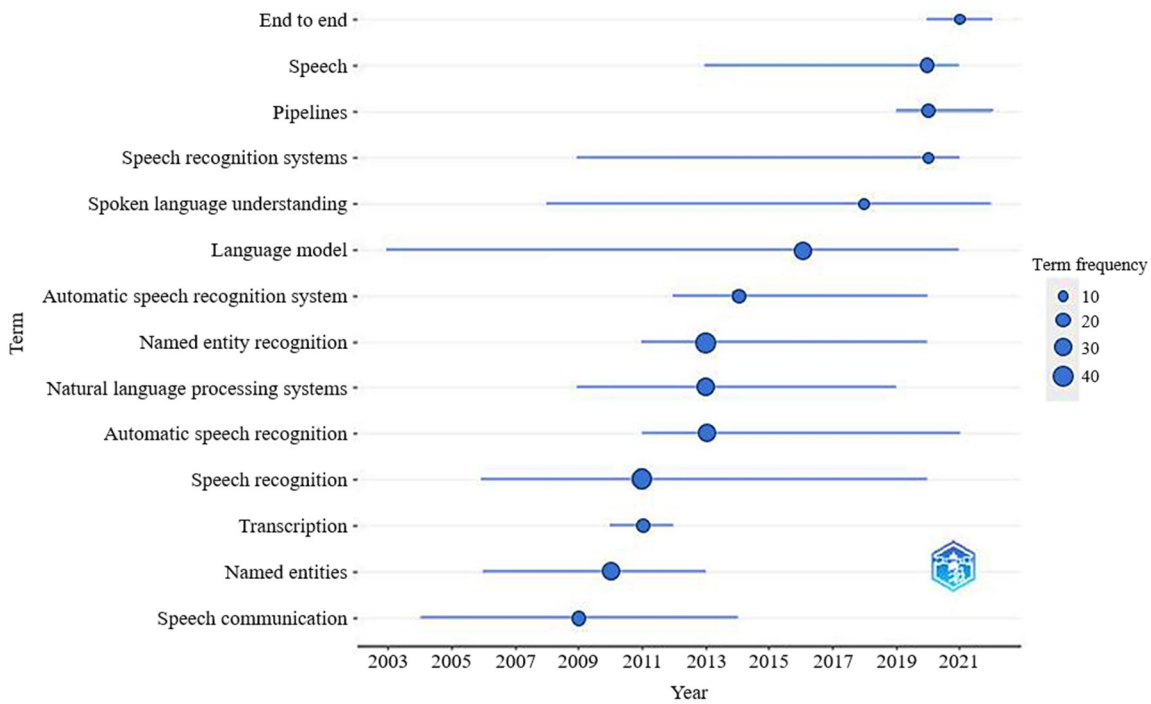


Fig. 8 The research trends topic in NER–ASR

The bar graph provides an overview of the research trends in NER–ASR ascertained through the co-occurrence of keywords in publications within the research field. Deep neural network models have played a crucial role in achieving state-of-the-art results in NER. Researchers have achieved these results by exploring different aspects of NER systems, including deep learning model architectures, training methods, training data, and the encoding of NER system outputs. These advancements highlight the adaptability and flexibility of deep learning models in improving NER performance.

Despite the progress made in NER, substantial amounts of human-annotated training data are still required. Efforts have been made to address this challenge by exploring the use of external knowledge sources, such as Named Entity dictionaries and part-of-speech tags, to replace human annotations. However, obtaining effective external resources remains a significant challenge. The shift from linear learning methods to deep learning architectures has been a pivotal development in NER research. Researchers have not only focused on refining algorithms and enhancing NER performance but have also explored novel approaches to address challenges specific to NER. Additionally, the influence of upstream and downstream tasks related to NER, such as sequence tagging and entity linking, has further shaped the trajectory of NER advancements.

The word cloud visually represents the prominence of specific terms within the context of NER–ASR. The size of each word in the cloud indicates its frequency or importance. Prominent terms in the word cloud, such as "named entities," "speech recognition," and "natural language processing systems," highlight critical concepts in NER–ASR research. These terms suggest that researchers in this field may be focused on refining algorithms to handle ASR errors, improving entity recognition accuracy, and developing NER systems that are specifically designed for ASR applications.

Fig. 9 analyzes the co-occurrence of author keywords, with the minimum number of occurrences of a keyword being two. The network of keywords is based on their co-occurrence and represents those most frequently used in publications on NER–ASR. Relevant keywords were grouped and given the same color. The links between keywords represent their co-occurrences, and the size of the keyword is proportional to its frequency of occurrence. Out of a total of 92 keywords, the fact that only 13 meet the threshold suggests that these concepts are particularly relevant and prominent in NER–ASR research. The analysis confirms that the research on NER for ASR has been focusing on end-to-end deep learning approaches since 2016. This finding aligns with the broader trend in NER research, where the adoption of deep learning models has significantly impacted the field's development.

The network diagram underscores the significance of deep learning in NER–ASR research, with "deep learning" being one of the identified keywords. This suggests that researchers in this field have been leveraging deep learning techniques to enhance NER performance within ASR systems. Additionally, the network highlights the importance of "named entity recognition" and "automatic speech recognition" as core concepts in this research area. The interconnectedness of these keywords in the network diagram signifies their relevance and the interdependency between NER and ASR.

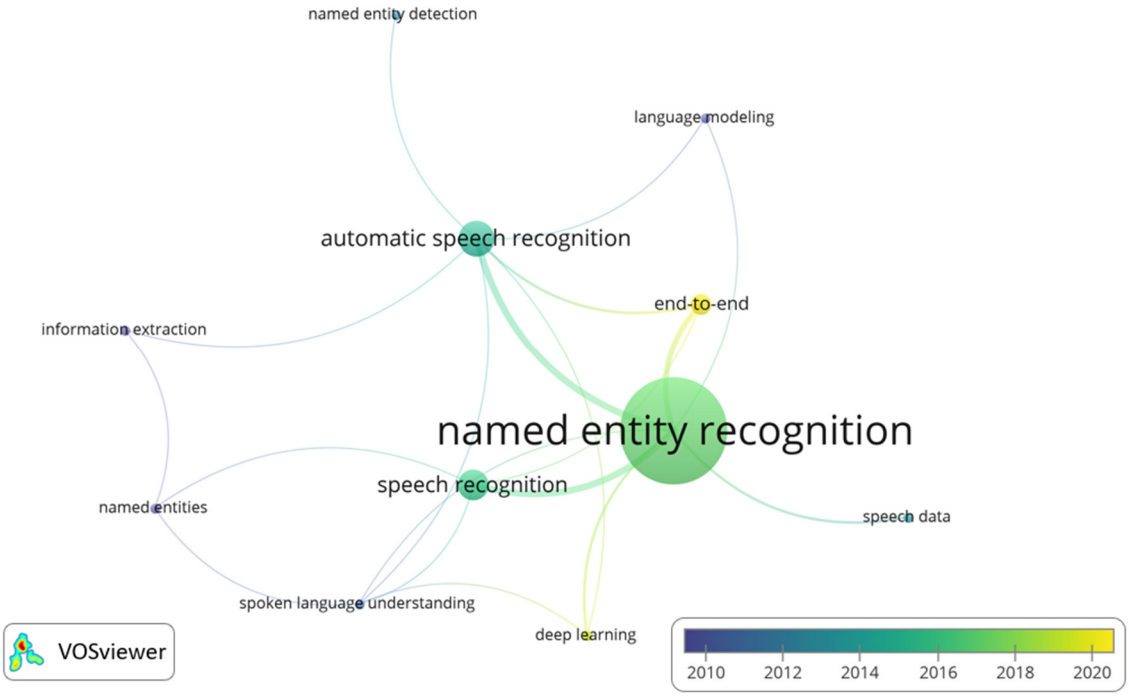


Fig. 9 Network of co-occurring keywords in NER–ASR publications

In addition to the bibliometric study, the study also carried out analyzing in detail the content of outstanding recent publications indexed by the Scopus database. The results are shown in Table 4 and illustrate the most challenging belongs to the database criterion. The majority focus on countries with rich data sources such as English [1, 15], Chinese [16-17], French [14, 18] and Japanese [19]. Very few studies have focused on low-resource languages [20-21]. The studies with limited databases have been proposed using standard text datasets, converting uppercase letters to lowercase, and removing punctuation marks to obtain output text data from ASR [20]. Especially, some have suggested methods to augment data [22], including Text-to-Speech systems to improve the data for model training, demonstrating a significant improvement in the performance of the recognition model [5, 20], and knowledge transfer to enrich training data [23].

In the initial stage, the method is mainly based on rule sets [24]. However, when the input is the ASR output text, the capitalization information for named entities is no longer available, making it challenging to gather the necessary linguistic information to construct the rules. So, many machine learning-based approaches have been proposed such as the SVM [19], HMM [25], Maximum Entropy [26], and CRF [27], which focused on English, Chinese, Japanese, and French. Researchers in NLP and ASR are actively exploring ways to improve NER within ASR systems. Techniques such as deep learning architectures have significantly impacted NER performance. Understanding the co-occurrence patterns of relevant terms can guide further research and system development. In particular, in terms of research methods, the main trend is also shifting to using Deep learning models. These approaches offer advantages in vector representation, computational capability, non-linear mapping from input to output, and the ability to learn high-dimensional latent semantic information [28].

Table 4 demonstrates the dominance of the bidirectional encoder representations from the transformers (BERT) model which has been exploited critically in recent studies [16-17, 20]. BERT addresses context in NER by utilizing pre-trained language models to extract text features at the word level, capturing contextual information effectively. By integrating BERT with models like BiLSTM and CRF, BERT can consider surrounding words and their sequential relationships, ensuring a better understanding of the context and dependencies between adjacent entities for coherent entity recognition.

The BERT method enhances semantic information enrichment, resolving issues like unclear entity boundaries and semantic ambiguity during training, leading to improved accuracy in entity recognition tasks. Furthermore, BERT might encode individual characters and extract sequential information through dual-channel networks contributing to achieving high accuracy, recall, and F1 scores in NER tasks [17, 20]. Some proposals addressed the NER system's issue with data by pre-training the prediction of capitalization [29] or using punctuation and uppercase recovery model [20] before combining it with the NER model.

Only 6 out of 55 publications have an end-to-end, while the rest all use a pipeline approach. Almost all early studies mostly employed the pipeline approach due to its simplicity in design. However, this approach involves training individual components separately, requiring separate training algorithms and loss functions for each component, therefore a large number of hyperparameters are needed, leading to training complexity. Furthermore, the errors occurring in each component are not computed when combined with other components, resulting in significant accumulated errors.

Some studies have demonstrated the effectiveness of the end-to-end model when combined with language models [6] or augmenting training data [22]. Almost all studies proved that the end-to-end model was not better than the pipeline model in terms of performance [7] and confirmed that "filtering" data through each step was still possible and therefore improved results. It also documented the need to improve the ASR model to reduce "anomaly" errors of insertion, deletion, substitution, and addition of words in the ASR output text [7, 20]. However, an end-to-end approach still shows the potential to optimize the ASR system. This is a complex process that spans from the beginning to the end. This model demonstrates the advantages of integrating the system into a single model, which facilitates the training process, minimizes errors between components, improves execution speed, and enhances deployability in practical applications.

Table 4 Statistics of some recent studies on NER for ASR indexed in Scopus database

Author	Year	Data	Approaches/methods/techniques	Results
Yadav et al. [6]	2020	English speech	• Two-step pipeline and end-to-end • End-to-end: CNN - BiLSTM - FC and Softmax • CTC loss	• The E2E approach provides better results compared to the two-step pipeline approach
Caubrière et al. [7]	2020	French speech (ETAPE 2012)	• Pipeline and end-to-end • End-to-end: CNN - BiLSTM – Softmax • CTC loss	• In comparison with the best result of ETAPE 2012, the E2E system improved by 4%. • The E2E approach shows interest but is below the updated pipeline approach.
Porjazovski et al. [23]	2020	Finnish speech Knowledge transfer from the Estonian dataset to enrich training data.	• Pipeline • BiLSTM-CRF • Rule-based	• The proposed model is better than the neural network architecture, and worse than the rule-based system. • Converted the training set to lowercase and removed the punctuation, which yielded significant improvement.
Porjazovski et al. [15]	2021	Finnish, Swedish, and English speech	• End-to-end • Attention-based encoder-decoder model for ASR with NER tags. • Multitask learning for speech transcription and named entity annotation.	• The multi-task approach allowing additional fine-tuning of the NER branch, outperforms the augmented labels approach
Chen et al. [16]	2022	Chinese speech (AISHELL-NER built upon AISHELL-1)	• End-to-end • Transformer: Entity-aware ASR • BERT: Pretrained NER tagger	• Conformer ASR outperforms Transformer ASR in CER. • Transformer EA-ASR faces a small loss in performance.
Nguyen et al. [20]	2022	Vietnamese speech	• End-to-end • Multi-task learning with the punctuation and uppercase (CaPu) recovery model • ViBERT: Pretrained NER tagger	• Multi-task learning model with CaPu recovery for improved 5% F1 score
Wang et al. [17]	2023	Chinese speech (CLUENER2020)	• BERT-BiLSTM-CRF • The BIO annotation method	• The BERT-BiLSTM-CRF model offers superior performance in NER for controlled speech compared to traditional methods.
Liu et al. [30]	2023	Chinese speech (Aishell3-NER, CNERTA, and MSRA)	• Proposes a multimodal Chinese NER method called USAF (Using Synthesized Acoustic Features). • USAF uses synthesized acoustic features and a multi-head attention mechanism	• USAF improves the performance of Chinese-named entity recognition. • USAF outperforms the SOTA external-vocabulary-based method on two datasets
Olatunji et al. [21]	2023	AfriSpeech-200 dataset	• Multilingual pre-training, data augmentation, fine-tuning on African accents. • Addressed problem as distribution shifts to mitigate model bias.	• The baseline model shows a significant decrease in performance on samples with African-named entities. • Fine-tuned models demonstrate an 81.5% relative WER improvement on samples with African-named entities.

Despite the valuable insights yielded by this bibliometric analysis, several inherent limitations warrant careful consideration for further investigations. Firstly, the exclusive reliance on the Scopus database, although recognized for its extensive coverage, and reliability, commonly used in bibliometric analysis, presents a significant limitation. This might be attributed to the relevant research indexed in other academic databases, such as IEEE Xplore, Web of Science or Google Scholar may be omitted, and therefore can result in a partial and biased view of the research landscape. This limitation is particularly critical in rapidly evolving fields like NER and ASR, where innovative work may appear in varied sources not included in Scopus. Additionally, bibliometric methods primarily focus on quantitative metrics such as citation counts, publication trends, and co-author networks. This approach, while useful for identifying general trends, can overlook crucial qualitative aspects of research, such as the novelty of findings, methodological rigor, and practical applications.

As a result, important but less cited contributions, often from niche or emerging research, may be underrepresented. Another limitation is temporal analysis. By focusing on a specific period, the study may not adequately reflect the historical development of the field or recent advances that have not yet accumulated a significant number of citations. This time lag can lead to outdated conclusions, especially in dynamic areas where knowledge evolves quickly.

The dominance of citation metrics also poses problems. By favoring well-established research areas and highly cited articles, this approach may overlook innovative but less visible studies. This trend is accentuated by the under-representation of research published in languages other than English or from less academically visible regions, which can distort the overall perspective of research activities.

Finally, bibliometric analysis, by highlighting collaboration patterns and prominent researchers, can overshadow the contributions of lesser-known researchers or institutions producing high-quality work. Lack of ongoing monitoring and integration of diverse data sources is essential to maintaining a current and comprehensive understanding of the research landscape, particularly in rapidly evolving areas like NER and ASR.

In general, while bibliometrics provides robust tools for analyzing research trends, it is essential to acknowledge its limitations. Implementing more in-depth content analysis methods, such as systematic reviews, and integrating qualitative approaches and data from multiple sources is particularly necessary. These steps, especially when conducted regularly, are critical for obtaining a more accurate and comprehensive understanding of the studied field.

4. Conclusions

In this work, Biblioshiny and VOSviewer have been used to conduct a quantitative analysis of scientific publications related to NER–ASR systems, published in the Scopus-indexed database before 2024. The analysis revealed stable growth in the field, with notable publication spikes in 2013 and 2022. Key findings include identifying countries with the highest number of publications and significant international collaborations, highlighting the diversity and global nature of NER–ASR research. Most publications were found in prominent conference and workshop proceedings, which are known as major forums for disseminating research in this domain. The study also underscored the impact of advanced deep learning models, including BERT and its variants, which have significantly improved the accuracy of NER systems by enabling the adaptation of pre-trained models.

The findings provide a comprehensive overview of the research landscape in NER for ASR, offering valuable insights into publication trends, the scientific impact of various contributions, and the network of international collaborations. Based on the analysis, several promising research directions have been proposed including the integration of Multimodal Data to enhance NER systems by combining audio, text, and visual data; cross-lingual NER for ASR to develop models that can effectively handle multiple languages and dialects; real-time NER in ASR with purpose improve the efficiency and speed of NER systems to enable real-time applications; domain-specific NER purpose to tailor NER systems to specific domains such as healthcare, finance, or legal sectors, and contextual adaptation to enhance NER accuracy by adapting to the context of the conversation.

Future investigations will aim to integrate data from multiple database sources to provide more comprehensive analyses and expand the scope of data extraction, such as language, document type, etc. To enhance the accuracy and reliability of bibliometric analyses for the continuous development of NER–ASR systems.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] G. Attardi, G. Berardi, S. Dei Rossi, and M. Simi, "The Tanl Tagger for Named Entity Recognition on Transcribed Broadcast News at Evalita 2011," Evaluation of Natural Language and Speech Tool for Italian, vol. 7689, pp. 116-125, 2013.

[2] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, and L. Quintard, "Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview," Proceedings of the 5th Linguistic Annotation Workshop, pp. 92-100, June 2011.

[3] D. Yu and L. Deng, Automatic Speech Recognition, vol. 1, Berlin: Springer, 2016.

[4] M. Hatmi, C. Jacquin, E. Morin, and S. Meigner, "Incorporating Named Entity Recognition Into the Speech Transcription Process," Proceedings of the 14th Annual Conference of the International Speech Communication Association, pp. 3732-3736, August 2013.

[5] I. Cohn, I. Laish, G. Beryozkin, G. Li, I. Shafran, I. Szpektor, et al., "Audio De-Identification: A New Entity Recognition Task," https://doi.org/10.48550/arXiv.1903.07037, March 17, 2019.

[6] H. Yadav, S. Ghosh, Y. Yu, and R. R. Shah, "End-To-End Named Entity Recognition From English Speech," https://doi.org/10.48550/arXiv.2005.11184, May 22, 2020.

[7] A. Caubrière, S. Rosset, Y. Estève, A. Laurent, and E. Morin, "Where Are We in Named Entity Recognition From Speech?" Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4514-4520, May 2020.

[8] S. Thanuskodi, "Journal of Social Sciences: A Bibliometric Study," Journal of Social Sciences, vol. 24, no. 2, pp. 77-80, 2010.

[9] R. E. Lopez Martinez and G. Sierra, "Research Trends in the International Literature on Natural Language Processing, 2000-2019 — A Bibliometric Study," Journal of Scientometric Research, vol. 9, no. 3, pp. 310-318, 2000.

[10] N. Khadivi and S. Sato, "A Bibliometric Study of Natural Language Processing Using Dimensions Database: Development, Research Trend, and Future Research Directions," Journal of Data Science, Informetrics, and Citation Studies, vol. 2, no. 2, pp. 77-89, 2023.

[11] I. Budi and R. R. Suryono, "Application of Named Entity Recognition Method for Indonesian Datasets: A Review," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 2, pp. 969-978, April 2023.

[12] I. Zupic and T. Čater, "Bibliometric Methods in Management and Organization," Organizational Research Methods, vol. 18, no. 3, pp. 429-472, July 2015.

[13] J. D. Burger, D. Palmer, and L. Hirschman, "Named Entity Scoring for Speech Input," Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 201-205, August 1998.

[14] S. Ghannay, A. Caubrière, Y. Estève, N. Camelin, E. Simonnet, A. Laurent, et al., "End-To-End Named Entity and Semantic Concept Extraction From Speech," IEEE Spoken Language Technology Workshop, pp. 692-699, December 2018.

[15] D. Porjazovski, J. Leinonen, and M. Kurimo, "Attention-Based End-To-End Named Entity Recognition From Speech," 24th International Conference on Text, Speech, and Dialogue, pp. 469-480, September 2021.

[16] B. Chen, G. Xu, X. Wang, P. Xie, M. Zhang, and F. Huang, "AISHELL-NER: Named Entity Recognition From Chinese Speech," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8352-8356, May 2022.

[17] Z. Wang, Y. Wang, X. Wang, and Q. He, "BERT-BiLSTM-CRF Based Named Entity Recognition Method for Controlled Speech," 6th International Conference on Artificial Intelligence and Big Data, pp. 270-275, May 2023.

[18] B. Favre, F. Béchet, and P. Nocéra, "Robust Named Entity Extraction From Large Spoken Archives," Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 491-498, October 2005.

[19] K. Sudoh, H. Tsukada, and H. Isozaki, "Named Entity Recognition From Speech Using Discriminative Models and Speech Recognition Confidence," Journal of Information Processing, vol. 17, pp. 72-81, 2009.

[20] T. H. Nguyen, T. B. Nguyen, Q. T. Do, and T. L. Nguyen, "End-To-End Named Entity Recognition for Vietnamese Speech," 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, pp. 1-5, November 2022.

[21] T. Olatunji, T. Afonja, B. F. Dossou, A. L. Tonja, C. C. Emezue, A. M. Rufai, et al., "AfriNames: Most ASR Models "Butcher" African Names," https://doi.org/10.48550/arXiv.2306.00253, June 01, 2023.

[22] A. Pasad, F. Wu, S. Shon, K. Livescu, and K. J. Han, "On the Use of External Data for Spoken Named Entity Recognition," https://doi.org/10.48550/arXiv.2112.07648, July 09, 2022.

[23] D. Porjazovski, J. Leinonen, and M. Kurimo, "Named Entity Recognition for Spoken Finnish," Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery, pp. 25-29, October 2020.

[24] J. H. Kim and P. C. Woodland, "A Rule-Based Named Entity Recognition System for Speech Input," 6th International Conference on Spoken Language Processing, vol. 1, pp. 528-531, October 2000.

[25] D. D. Palmer, M. Ostendorf, and J. D. Burger, "Robust Information Extraction From Spoken Language Data," Sixth European Conference on Speech Communication and Technology, pp. 1035-1038, September 1999.

[26] G. Kurata, N. Itoh, M. Nishimura, A. Sethy, and B. Ramabhadran, "Leveraging Word Confusion Networks for Named Entity Modeling and Detection From Conversational Telephone Speech," Speech Communication, vol. 54, no. 3, pp. 491-502, March 2012.

[27] M. Hatmi, C. Jacquin, E. Morin, and S. Meignier, "Named Entity Recognition in Speech Transcripts Following an Extended Taxonomy," First Workshop on Speech, Language and Audio in Multimedia, pp. 61-65, August 2013.

[28] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, January 2022.

[29] S. Mayhew, G. Nitish, and D. Roth, "Robust Named Entity Recognition With Truecasing Pretraining," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 8480-8487, 2020.

[30] Y. Liu, S. Huang, R. Li, N. Yan, and Z. Du, "USAF: Multimodal Chinese Named Entity Recognition Using Synthesized Acoustic Features," Information Processing & Management, vol. 60, no. 3, article no. 103290, May 2023.