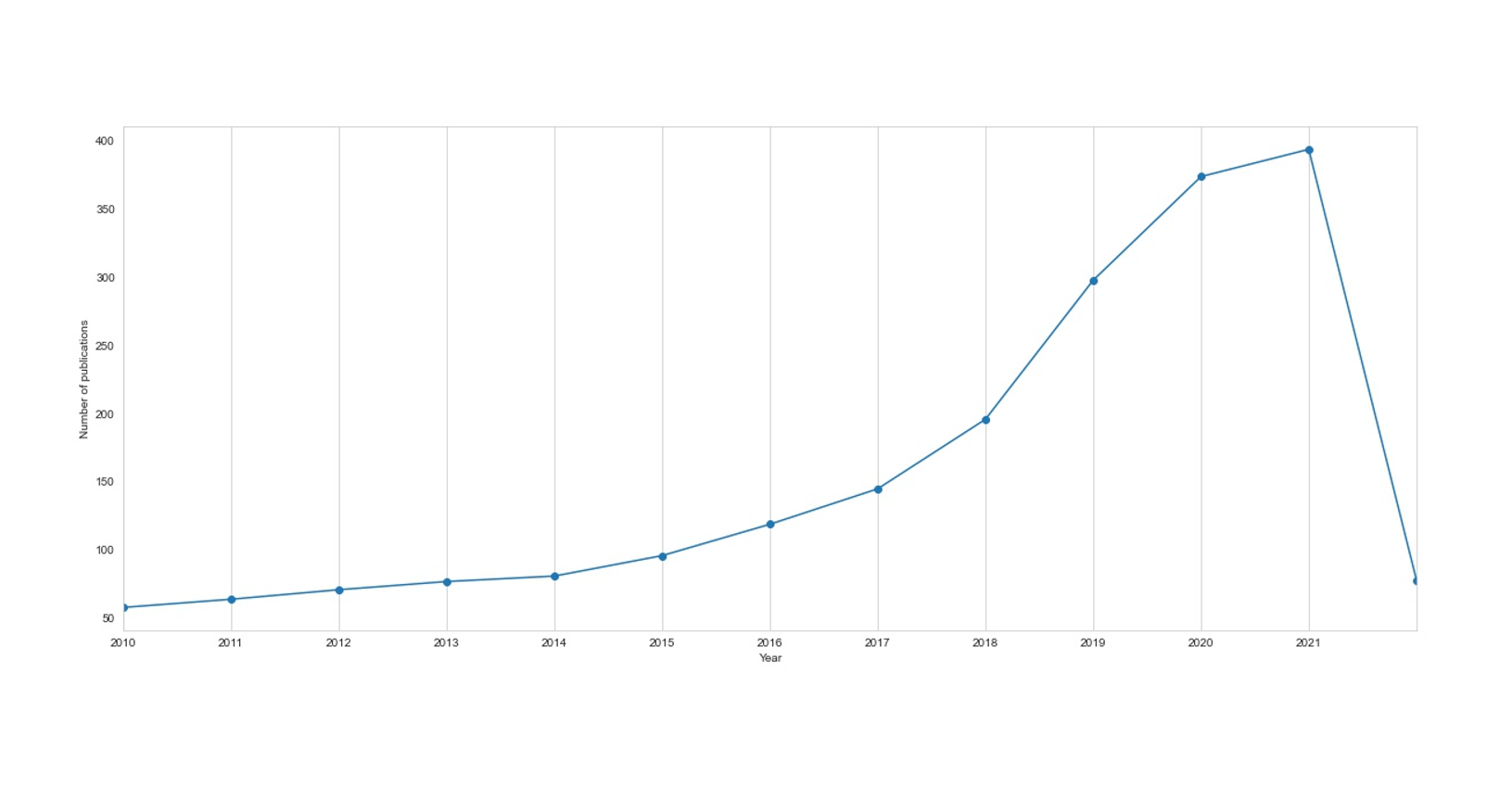
The fast growth in the number of NLP papers and resources has made it more difficult to find and extract useful information from papers. To handle this issue, researchers widely apply bibliometric analysis, which is defined as a powerful tool for quantifying research output using geometrical and statistical analysis. Bibliometric analysis has several advantages, including organizing information in a specific field, assessing scientific advancements of a particular topic, and comparing the quality of the research throughout affiliations and countries. For newcomers in the research field, bibliometric analysis can help them choose relevant study topics by providing an assessment of recent research studies and developments in the interested field. The bibliometric analysis of named entity recognition is conducted in this study utilizing the Scopus database. Information about NER publications in recent 12 years was retrieved from this database which includes citations, bibliographical information, and also funding details of articles. In this paper, the influence of journals, affiliations, and authors was assessed by citation analysis, whereas network analysis was used to determine the links between keywords, authors, and countries.



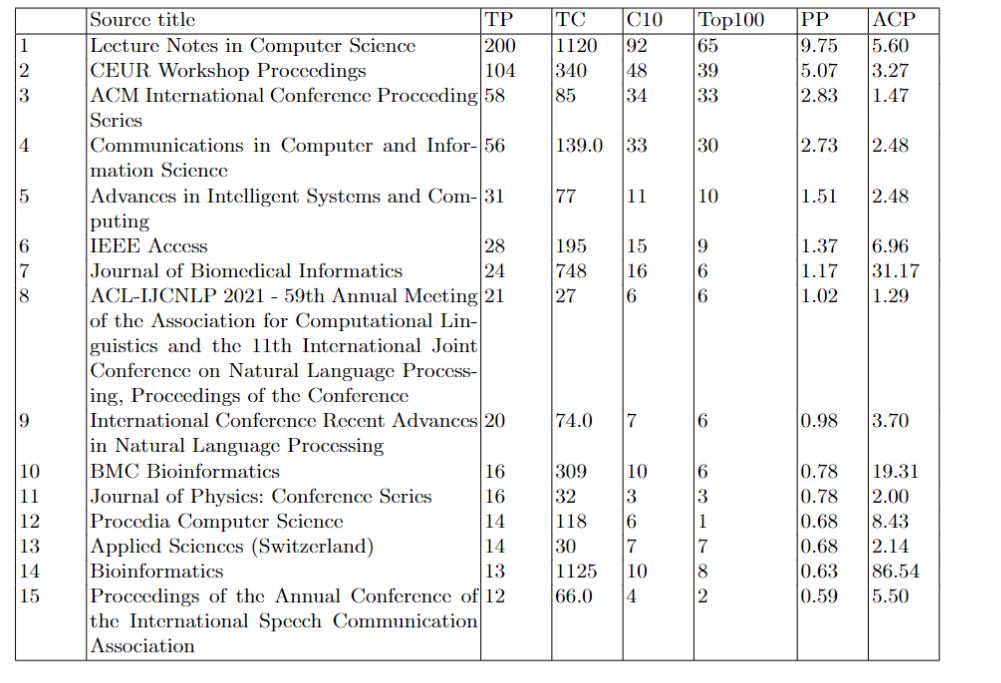
Results

There is an increase in the number of publications in the study span, as shown in Fig1. Only 58 papers were published in the year 2010, and then there is stable progress till 2015. After that, the annual number of publications increase dramatically and extreme value appear in 2021 as 394. The results show an overall upward trend and demonstrate that the Named Entity Recognition has been becoming popular in the research field in recent years.

The total publications, total citations, average number of citations per publication, and the number of annual citations are demonstrated in Figure 2. The results show that the research in the NLP empowered mobile computing field exhibits an overall upward trend in fluctuation (from 12 publications in 2000 to 55 publications in 2016). The publication number presents a stable increasing trend since 2010. Based on the data for years 2010–2016, we developed a regression model by setting the independent variables as time/1000 and (time/1000)2.The estimatedregression model is calculated as 𝑦 = 6.7143 ∗ 103 − 1.34777 ∗ 104𝑥.The adjusted goodness-of-fit 𝑅2 of the model is 0.9468. With the regression model, publication number in 2017 is predicted as 65, while the actual number of publications on WoS in 2017 is 66. The trend of citations does not keep step with publication number, and extreme values appear in 2002 as 431, 2007 as 503, and 2010 as 490. The average number of citations per publication is calculated as total citations/total publications. It shows an overall downward trend in fluctuation from 21.92 in 2000 to 2.53 in 2016. We eliminated the influence of duration since first publication using the formula: the number of annual citations (C/Y) = total citations/(2016 + 1-publishing year). The number of annual citations increases in fluctuation from 15.47 in 2000 to 139 in 2016.

Figure 2 illustrates the number of papers released by various sources relevant to our study topic.

3.2. Productive Journals. The top 11 contributing journals in the research field are presented in Table 3. These journals contribute about 21% of the total publications and 29.20% of the total citations. The most productive 3 are IEEE/ACM Transactions on Audio Speech and Language Processing (25 publications, 447 citations, 17.88ACP, and 11𝐻-index), Speech Communication (11 publications, 179 citations, 16.27ACP, 6 𝐻-index), and Computer Speech and Language (10 publications, 93 citations, 9.30 ACP, 6𝐻-index). Expert Systems with Applications has the highest ACP of 40.00. We found that 32 of the 100 most influential publications are published in the 11 journals. According to subject category of these 11 journals, computer science possesses the widest influence in the research field. In order to better measure the overall scientific importance of these 11 journals, 5 assessment indicators acquired fromScientific Journal Rankingswere used, including Impact Factor (IF), SCImago Journal Rank (SJR), 5-Year IF, Source Normalized Impact per Paper (SNIP), and CiteScore. IF is a measure for reflecting the yearly average number of citations to recent publications published in a journal. It is the primary and widely used indicator on assessing one journal’s significance. SJR is a measure of scientific influence of scholarly journals. It accounts for both the number of citations received by a journal and the importance or prestige of the journals where such citations come from. 5-Year IF is calculated by dividing the number of citations to the journal in a given year by the number of publications published in that journal in the previous five years. SNIP is defined as the ratio of the journal’s citation count per publication and the citation potential in its subject field. CiteScore index, launched by Elsevier in December 2016, is calculated as the ratio of total citations received in a given year by all publications published in a given journal in three previous years and the number of publications published in the journal in three previous years. Therefore, the 11 productive journals were compared by using their IF, SJR, 5-Year IF, SNIP, and CiteScore for year 2016, as shown in Figure 3. As for IF, SJR, and CiteScore, the top 3 are Information Sciences (IF 4.832, SJR 1.91, and CiteScore 5.37), Expert Systems with Applications (IF 3.928, SJR 1.433, and CiteScore 4.7), and IEEE/ACM Transactions on Audio Speech and Language Processing (IF 2.491, SJR 0.813, and CiteScore 3.5).As for 5-Year IF, the top 3 are Information Sciences (5-Year IF 4.731), Expert SystemswithApplications (5- Year IF 3.526), and Personal and Ubiquitous Computing (5- Year IF 2.512). As for SNIP score, the top 3 are IEEE/ACM Transactions on Audio Speech and Language Processing (SNIP 3.143), Information Sciences (SNIP 2.537), and Expert Systems with Applications (SNIP 2.492).



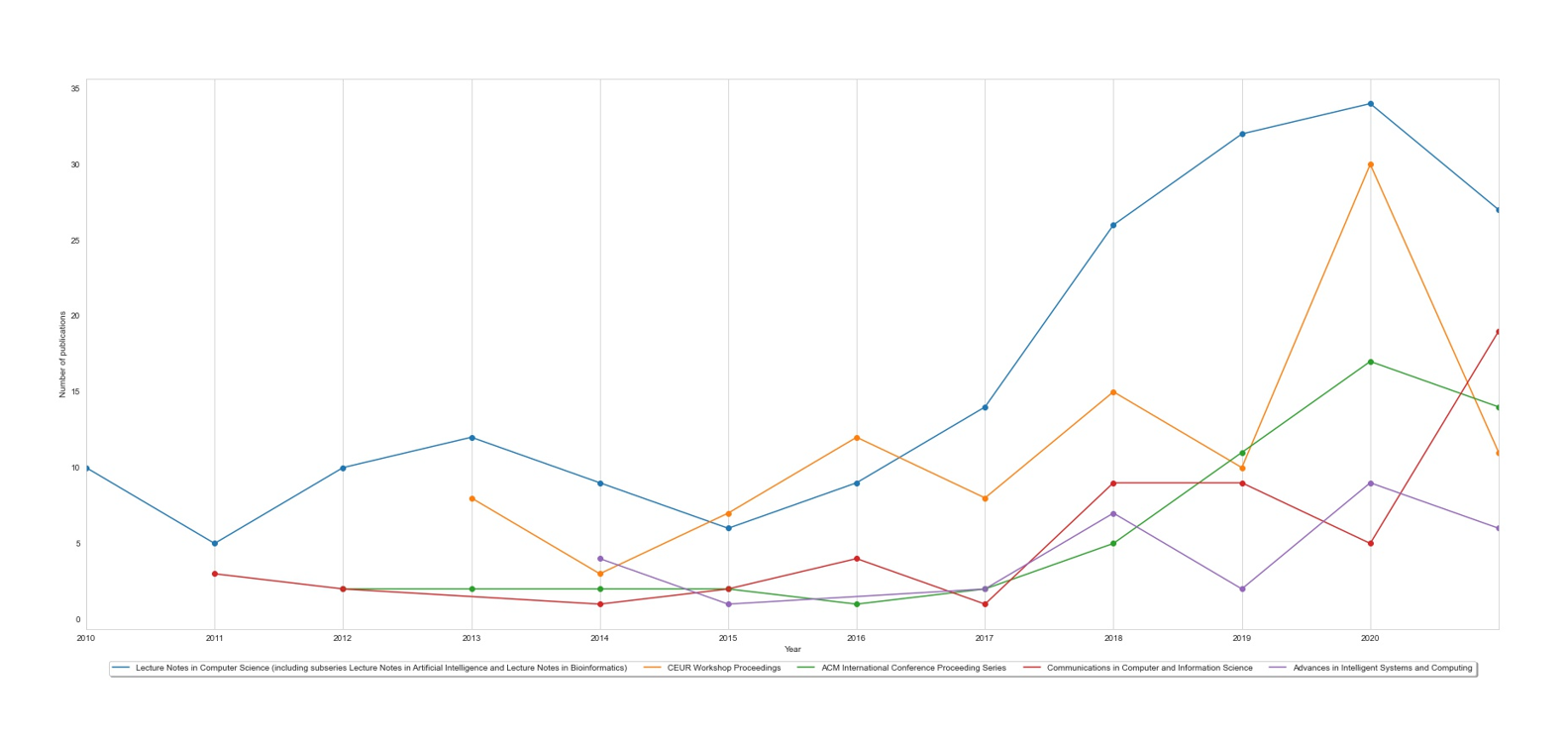
There are top 15 productive journals in Table 2. To evaluate contribution of the journal we used several measumerents: total publications(TP); percentage of the publications (PP); total citations (TC); average number of citations per publication(ATC), which calculated as TC/TP; number of publications with citations ≥10(C10); number of publications which in the top 100 most cited publications (T100). Computer science field most effect on the research sector, according to the topic categories of these journals.

According to the categories of these journals , computer science is the most common area of NER . Beside computer science domain NER is also applied in biomedicine. There are three journals in top ranks which specialized in the biomedicine called ‘Journal of Biomedical Informatics’, ‘BMS Bioinformatics’ and ‘Bioinformatics’ with 24, 16 and 13 total publications respectively.

There are the top 15 productive journals in Table 2. To evaluate the contribution of the journal, we used several measurements:

* Total publications(TP).
* Percentage of the publications (PP).
* Total citations (TC).
* The average number of citations per publication(ATC) is calculated as TC/TP.
* The number of publications with citations ≥10(C10).
* The number of publications which in the top 100 most cited publications (T100).

According to the categories of these journals, computer science is the most common area of NER. However, besides the computer science domain, NER is also applied in biomedicine. Three journals in the top ranks specialized in biomedicine called "Journal of Biomedical Informatics," "BMS Bioinformatics," and "Bioinformatics," with 24, 16, and 13 total publications, respectively.



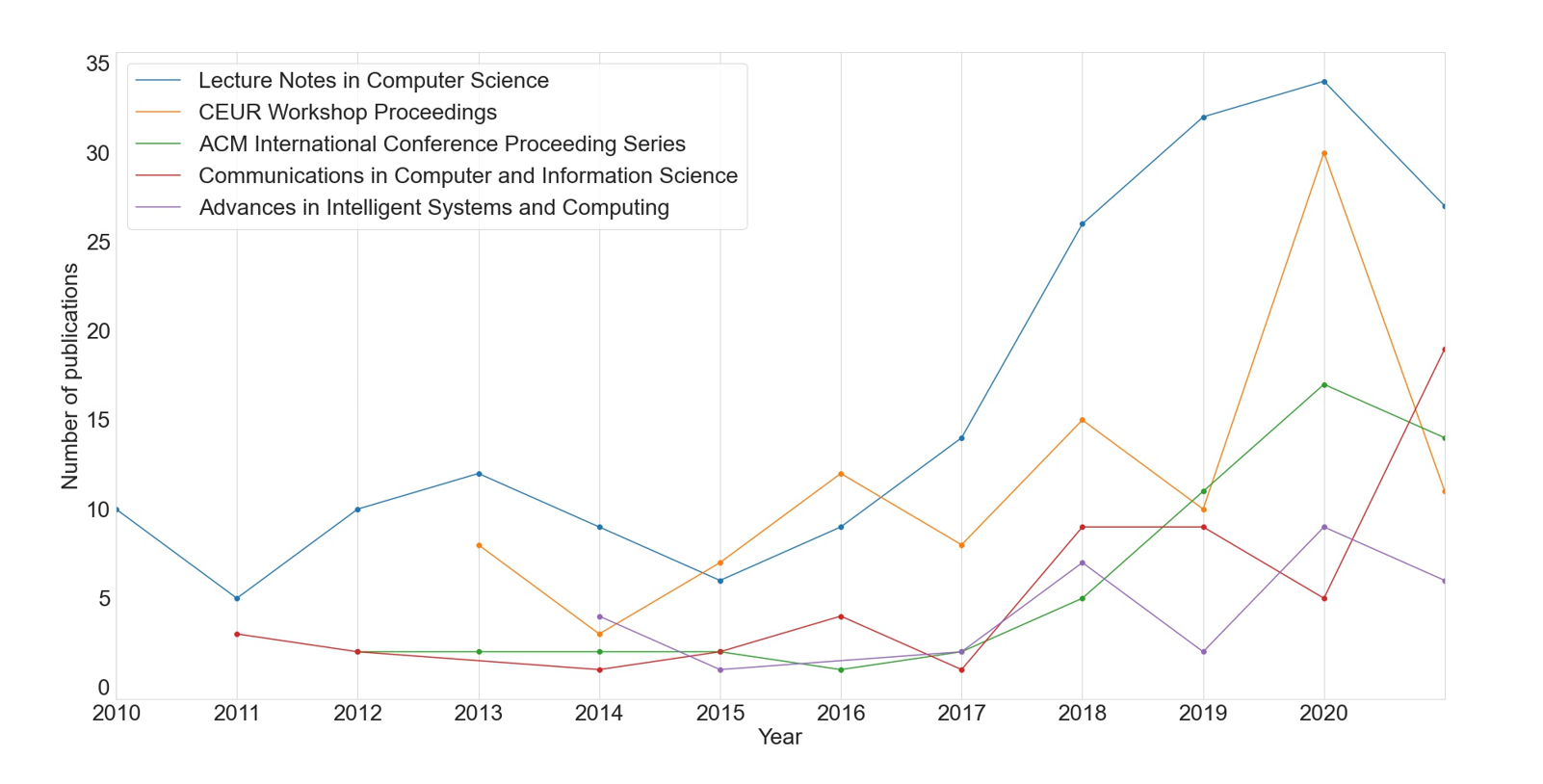


Figure 2 shows the number of publications last 11 years, related to our research domain, published by top 5 influential journals. The results show that the research in the NLP exhibits an overall upward trend in fluctuation from 2010 to 2021.

1) Lecture Notes in Computer Science source: Analysis shows that the oldest paper is from 2010. In comparison to the other four sources in the graph, they have published the most articles, 34 documents in the year 2020. The graph's second highest peak, 31 documents, is likewise from the same source.

2) CEUR Workshop Proceedings source: Since 2013, they've been publishing papers that are specific to our field of study. In the year 2020, this journal published the most articles with 30. Figure 3 shows a decline in the number of papers published from 30 to 11 papers  between 2020-2021

3) ACM International Conference Proceeding Series: They published their first paper in 2012, which was about NER. As illustrated in Figure 2, there was a significant increase from 2017 to 2020.

4) Communications in Computer and Information Science source: In the year 2021, they published the most papers, which totaled 19 . They began publishing articles relevant to the study in 2011.

5) Advances in Intelligent Systems and Computing source: The highest peak of this journal related to our domain was in 2020 with 9 papers. The source is quite new in the NER area, having only published its first paper in 2014.

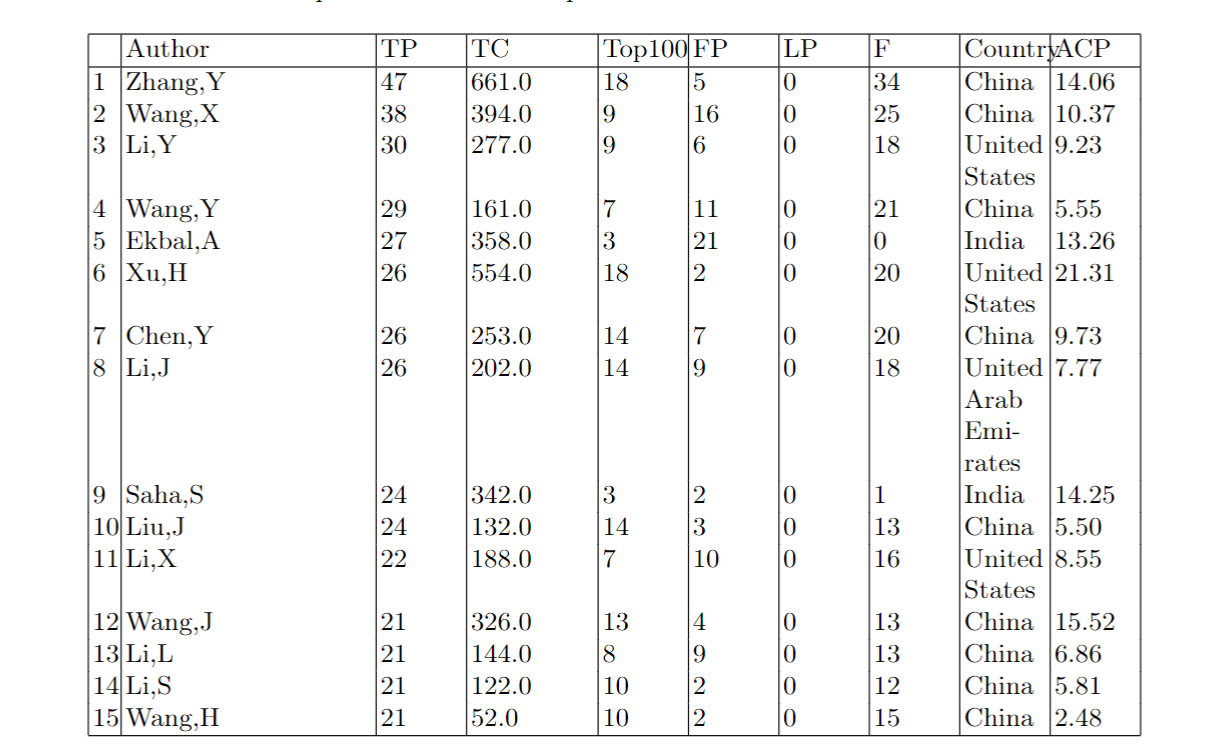
The number of citations reflects the popularity and influence of a publication in the scientific community [10].Thus, we used the total citations as a measurement of influence.There are 69 and 129 publications with the number of citations ≥20 and ≥10. Top 15 most influential publications are listed in Table 4. The publication by Miao et al. [35] in 2010 (376 citations) is the most influential one, followed by [36] published by MacKenzie and Soukoreff in 2002 (172 citations) and [37] by Strayer and Drews in 2007 (148 citations). We further consider the number of annual citations of the 15 publications. The top 3 publications measured by this indicator are [38] by Cao et al. published in 2015 (𝐶/𝑌 = 56), [35] by Miao et al. in 2010 (𝐶/𝑌 = 53.71), and [39] byMostafa in 2013 (𝐶/𝑌 = 22). These 3 publications rank 14th, 1st, and 6th, respectively, according

to total citations.

The amount of citations represents a publication's popularity and influence in the research world.

As a result, we utilized total citations as a metric of influence. Based on this metric, we listed the top 15 publications in Table 3. The most cited paper is "Neural architectures for named entity recognition" by Lample et al. in 2016, with 1501 citations. The article "Named entity recognition in tweets: An experimental study" by Ritter et al. took second place in most influential publications with 832 citations and followed by "Deep learning with word embeddings improves biomedical named entity recognition," published by Habibi et al. in 2017. These sorted publications are also measured by an indicator of the number of annual citations. This indicator is calculated as TC(total citation)/(2021-published year) 2,Named entity recognition in tweets: An experimental study,"Ritter A., Sam C., Mausam, Etzioni O.",2011,832.0,83.2

3,Deep learning with word embeddings improves biomedical named entity recognition,"Habibi M., Weber L., Neves M., Wiegandt D.L., Leser U.",2017,270.0,67.5



From overall 2051 publications there are 4467 authors who contribute to the field of Named Entity Recognition. 1392 authors published papers as first authors. To assess the contribution of the authors to the study , we used several measurements: total publications (TP), total citation (TC) , number of publications in top 100 most cited papers (Top100), number of funded publications (F), number of publications as first author (FP), number of publications as last author (LP), average number of citations per publication(ACP). Based on these indicators top 15 most productive authors are listed in Table 4. The most productive author, according to statistics, is Zhang, Y with 47 publications and 34 of them are funded. 661 times cited these publications and also 18 of them are in top 100 most influential papers. Followed by Wang, X (25 publications from 38 overall supported by funds, 394 citations,10.37 ACP) from China and Li, Y (18 publications from 30 overall supported by funds, 277 citations, 9.23 ACP) from United States. 9 papers of Wang, X and Li, Y are part of top publications. Ekbal, A from India has 27 papers and 21 of them published as first author. Xu, H from United States published total 27 papers and has the most high ACP with 21.31.

From overall 2051 publications, there are 4467 authors who contribute to the field of Named Entity Recognition. 1392 authors published papers as first authors. To assess the contribution of the authors to the study, we used several measurements: total publications (TP), total citation (TC), number of publications in the top 100 most cited papers (Top100), number of funded publications (F), number of publications as first author (FP), number of publications as last author (LP), the average number of citations per publication(ACP). Based on these indicators top 15 most productive authors are listed in Table 4. The most productive author, according to statistics, is Zhang, Y., with 47 publications, and 34 of them are funded. 661 times cited these publications, and also 18 of them are in the top 100 most influential papers. Followed by Wang, X. (25 publications from 38 overall supported by funds, 394 citations,10.37 ACP) from China and Li, Y. (18 publications from 30 overall supported by funds, 277 citations, 9.23 ACP) from the United States. 9 papers of Wang, X., and Li, Y. are part of top publications. Ekbal, A. from India, has 27 papers, and 21 of them were published as the first author. Xu, H. from the United States published total 27 papers and has the highest ACP with 21.31.

Co-authorship relations represents the collaboration between two authors if they have worked together. Analysis of co-authorship details helped identify groups of people who worked together closely. The maximum number of the co-occurrences is attributed to four pairs of authors: Zhang Y., Wang X., Chen Y., Xu H. Fig. 7 shows a co-authorship network map generated based on the number of published documents by the authors. The total number of keywords was 4290, which were too many to fit on a chart. Therefore, a threshold of five documents and five citations per author was set, and top 50 authors were selected .

If two authors have worked together, co-authorship relations represent their collaboration. Co-authorship details were examined to find groups of persons who collaborated closely. Four pairs of researchers of the greatest number of co-occurrences: Zhang Y., Wang X., Chen Y., and Xu H. A co-authorship network  based on the number of research papers by the authors is shown in Figure 7. There were 4290 keywords in total, which was too much to fit on a map. As a result, the top 50 authors were chosen based on a requirement of five papers and five citations each author.

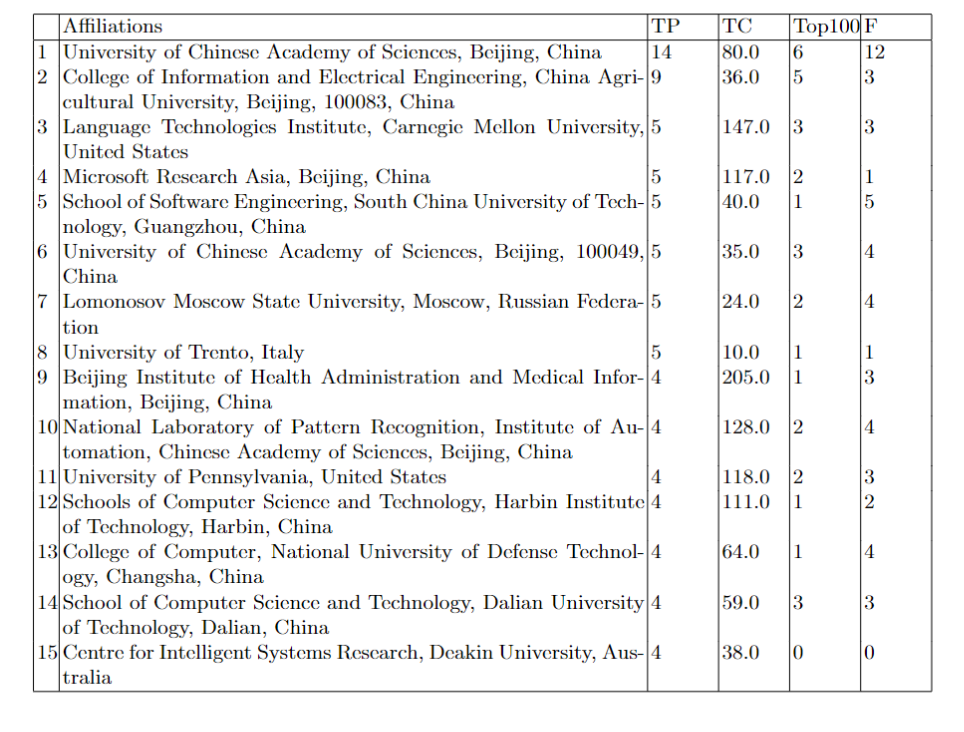


Table shows the top productive affiliation of recent years which hihly contribute to the field of NER. University of Chinese Academy of Sciences ranked first by amount of published papers with 14. It is much more than other influential affiliations papers number, but it has average number of total citation in comparison with others. 12 papers from total 14 are funded in China and 6 papers considered as influential publications. If we sort affiliations by total citation number, Beijing Institute of Health Administration and Medical Information is the first with 205 citations.This is the evidence that NER system is common in conducting research in medicine recent years. But this organization only has 1 paper in top ranking publications. Language Technologies Institute of Carnegie Mellon University takes the second place by the number of total citation with 147 and has 3 top papers that totally funded.

,Affiliations,TP,TC,Top100,F

1,"University of Chinese Academy of Sciences, Beijing, China",14,80.0,6,12

2,"College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China",9,36.0,5,3

3,"Language Technologies Institute, Carnegie Mellon University, United States",5,147.0,3,3

4,"Microsoft Research Asia, Beijing, China",5,117.0,2,1

5,"School of Software Engineering, South China University of Technology, Guangzhou, China",5,40.0,1,5

6,"University of Chinese Academy of Sciences, Beijing, 100049, China",5,35.0,3,4

7,"Lomonosov Moscow State University, Moscow, Russian Federation",5,24.0,2,4

8,,5,20.0,3,1

9,"University of Trento, Italy",5,10.0,1,1

10,"Beijing Institute of Health Administration and Medical Information, Beijing, 100850, China",4,205.0,1,3

11,"National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China",4,128.0,2,4

12,"University of Pennsylvania, United States",4,118.0,2,3

13,"Schools of Computer Science and Technology, Harbin Institute of Technology, Harbin, China",4,111.0,1,2

14,"College of Computer, National University of Defense Technology, Changsha, China",4,64.0,1,4

15,"School of Computer Science and Technology, Dalian University of Technology, Dalian, China",4,59.0,3,3

16,"Centre for Intelligent Systems Research, Deakin University, Australia",4,38.0,0,0

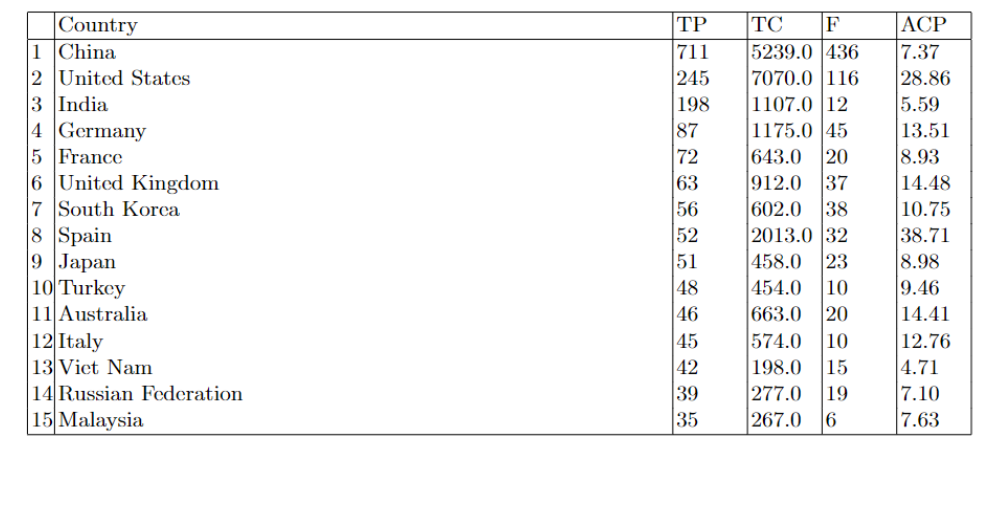
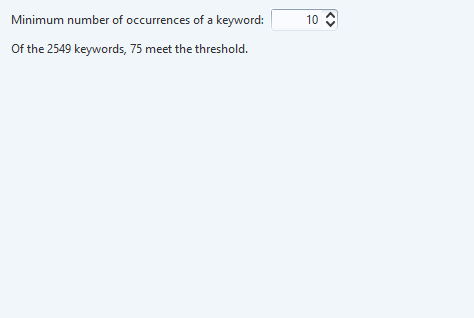
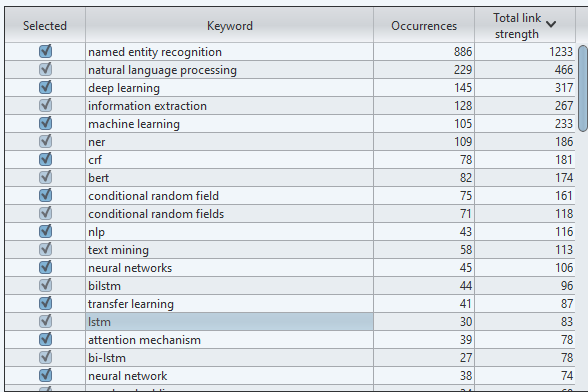


Table shows top productive countries sorted by number of total publications and number of total citation in the NER field. The 2051 publications are from less that 50 countries. The country where NER is the most poplular is China with 711 total publications. More than 50 percentage of publications by China were funded. But number of citations and ACP much lower than United States. Results shows that the top countries after China are the USA with 245 publications, 7070 citations and 28.86 ACP, India with 198 publications, 1107 citations, 5.59 ACP, Germany with 87 publications, 1175 citations, 13.51 ACP, and United Kingdom with 63 publications, 912 citations, 14.48 ACP. Among the countries the USA take second place mostly funding , almost half of publications are funded. However, India has the lowest number of funded publications.





Keywords

Keywords give information about a document's main content and can also be used to identify research trend in a particular domain. Table shows 15 keywords with most occurrences . Obviously the firs one is “named entity recognition” with 886 occurences and also “ner” with 109, total is 995. The next most common words are “deep learning” with 145 and “machine learning” with 105. “Conditional random fields” , “crf”, “lstm”,”bilstm” are name of methods which are commonly used in NER systems. As a result, their occurrences in author keywords a lot.

Keywords give information about a paper's leading content and can also be used to identify research trends in a particular domain. The table shows 15 keywords with the most occurrences. The first one is “named entity recognition” with 886 occurrences and “ner” with 109, total is 995. The following most common words are “deep learning” with 145 and “machine learning” with 105. “Conditional random fields,” “CRF,” “lstm”, ”bilstm” are the name of methods that are commonly used in NER systems. As a result, their occurrences in author keywords a lot.

In named entity recognition for ovarian cancer research dominant countries, organizations and authors were determined. Also the most cited article, the most productive journals were revealed. For identifying the research distribution around the globe, bibliometric maps using VOSviewer mapping software are displayed in this paper[6].

In this study, we presented several criteria including influential publications, keywords frequency, productive institutions, most influential authors, leading publishing journals, collaboration between authors. These criteria helped reveal the global trends related to research on named entity recognition. Bibliometric maps using VOSviewer mapping software are displayed in this paper.

Abstract

With the massive amount of data that has been generated in the form of unstructured text documents, Biomedical Named Entity Recognition (BioNER) is becoming increasingly important in the field of biomedical research. Since currently there does not exist any automatic archiving of the obtained results, a lot of this information remains hidden in the textual details and is not easily accessible for further analysis. Hence, text mining methods and natural language processing techniques are used for the extraction of information from such publications.Named entity recognition, is a subtask that comes under information extraction that focuses on finding and categorizing specific entities in text. In this paper, bibliometric analysis of named entity recognition of ovarian cancer is carried out using information about publications from Scopus. The most productive journals, countries and authors are determined. The most frequently cited article and its citation history has been described. Also bibliometric maps based on citation network among countries are constructed. This study can assist people in the medical field to get a comprehensive understanding of the study of BioNER.

Named Entity Recognition is becoming increasingly significant in the various fields as a result of the vast volume of unstructured data . A bibliometric study of named entity recognition  is performed in this paper retrieving  information about publications from Scopus. It is determined which journals, countries, and writers are the most influential.  The most commonly publishing journal  has been detailed, as well as its publication history.   This survey can help researchers  gain a thorough understanding of NER .