Journal of Scientometric Res. 2020; 9(3):310-318
http://www.jscires.org

Research Article

Research Trends in the International Literature on Natural Language Processing, 2000-2019 — A Bibliometric Study

Roberto E Lopez-Martinez[1]* and Gerardo Sierra[2]

[1]Instituto de Ingeniería, Universidad Nacional Autonoma de Mexico.
[2]Circuito Exterior Ciudad Universitaria04510 Ciudad de México.

ABSTRACT
This work presents a review, using bibliometric methods, of the state of research on the whole field of natural language processing (NLP), understanding this as the methods to process human language, including semantic techniques, statistical techniques or a combination of both. Particularly we focus on the trends of research in NLP, since there are not in the literature studies that embrace in an integral way bibliometric studies about natural language processing, its applications and related topics. Our work includes an identification of the main sources where research is published, the most productive and influential countries and research institutions, the main actors involved in research, as well as the main topics that are investigated. We found that research in the field and subfields has increased continuously during the period under study; conference proceedings are the preferred media to communicate results and that biomedical informatics is one relevant field of application of NLP. We conclude with both, a synchronic and a diachronic characterization of research topics carried out internationally on natural language processing and related topics, which showed that several subfields of artificial intelligence are closely related to natural language processing in recent years.

Keywords: Bibliometrics, Natural language processing, Computational linguistics.

Correspondence
Roberto E Lopez-Martinez
Edificio 12 Instituto de Ingenieria,
Circuito Exterior Ciudad Universitaria,
04510 Ciudad de México, MEXICO.
Email: robertol@unam.mx

Received: 27-08-2020
Revised: 14-10-2020
Accepted: 23-11-2020
DOI: 10.5530/jscires.9.3.38

INTRODUCTION

Currently, a large amount of information flows through language, due to its function as a means of communication and interaction; language is therefore the best tool for the transmission of knowledge. We are also facing a radical change in which information technologies offer immediate access to contents from the entire world. Within this context, the function of language is deeply transforming and therefore the need to model it emerges, to generate informational systems to exploit it more easily.

Within the framework of informatics and language, natural language processing emerges as one of the links of communication between human beings and machines. Natural Language Processing (NLP) can be defined as techniques to process human language, including semantic or statistical techniques or a combination of both.[1] Natural language comprehends speech and written texts; according to it, NLP consist in technologies related to voice recognition and technologies to process texts. The fields of research of natural language processing are among others identification of named entities, filtering and classification of documents, the generation of automatic abstracts, information extraction, sentiment analysis, opinion mining, monitoring of reputation in social media, orthographic and grammatical correction, smart and optimized search, automatic response systems, personal assistants, automatic translation, voice recognition, text to voice systems, dialogue systems, etc.

These technologies are now mature enough to have an important impact in business processes and services of organizations, as well as in their strategies. NLP can be applied as new forms of communication and information systems or as part of the existing ones. Different types of NLP can also be combined to provide solutions for the benefit of organizations.

The present paper uses bibliometric methods to study the state of the international research on natural language processing. Its objective is to offer a vision of the main countries and agents that intervene in the development of NLP, their communications media and the research subjects that are of interest. After this introduction, the second section is devoted to a literature review of salient aspects of bibliometrics and its use in subjects related to NLP. The third section is focused on the methodological aspects of the work. Subsequently, in sections four and five, we present respectively the results and conclusions of our research.

LITERATURE REVIEW

Bibliometric methods

Bibliometric methods have been developed since the beginning of the Twentieth Century but they got a boost when Derek J de Solla Price[2] suggested the use of scientific methods to study science; this derived into a quantitative approach for the description, evaluation and monitoring of published research. These methods make use of bibliographic data from publication databases to construct images of the structure of scientific fields. With these it is possible introduce an objectivity measure for the evaluation of scientific literature and can be used to identify informal networks of research or "invisible colleges" that exist under the surface but are not formally linked.[2,3]

There exist two main types of use of bibliometric methods: performance analysis and the mapping of science.[4] The first one aims at the performance evaluation of research and publications of individuals and institutions, while the second attempts to reveal the structure and dynamics of scientific fields. The mapping of science is a combination of classification and visualization; its objective is to generate a representation of the structure of the research area by means of the partition of certain elements —documents, authors, sources, terms, etc., into different groups; subsequently, visualization techniques are used to create a representation of the classification that was obtained.

Bibliometric studies have among their main applications citation and co-citation analyses, but only the first one has been used for this particular work. Citation analysis is usually communicated in the form of top-N lists of the most cited studies, authors, institutions or journals in the area under examination. Citations are used as a measure of influence; it is assumed that if an article is heavily cited, it is considered to be important, since authors usually cite articles which are important for their work.[5] On the other hand, given that bibliographic data contain information about the institutional affiliation and geographic location of the authors, co-authorship analysis allows the examination of collaboration between institutions and countries.[6]

Regarding other methods, word co-occurrence analysis[7,8] is a technique of content analysis that uses terms of parts of documents to establish relationships and build the conceptual structure of the domain. The underlying idea of the method consists in the fact that when words frequently co-occur in documents, it means that the concepts behind them are closely related. This is the only method that uses the content of the documents to generate a similarity measure, compared with citation and co-authorship analyses.[9] The result of word co-occurrence analysis is a network of subjects and their interrelations representing the conceptual space of a field, which helps to comprehend its cognitive structure.[10,11]

Bibliometric studies about subthemes of natural language processing and related topics

There are, in recent years, several studies of diverse subthemes or topics related to natural language processing. For example, Chen *et al*.[12] employed a range of techniques from descriptive statistics and social network analysis to latent Dirichlet allocation (LDA) and affinity propagation clustering, to analyze and detect the status and trends in a set of relevant publications concerning NLP empowered mobile computing, indexed in the Web of Science (WoS) database. Using other techniques such as text mining and qualitative coding, Mäntylä *et al*.[13] analyzed 6, 996 papers from Scopus to study the recent evolution of sentiment analysis and opinion mining. Also, within this subject, bibliometric methods have been employed to evaluate quantitatively and qualitatively, research trends in the field;[14,15] their results include: number of publications and rate of growth, collaboration patterns, productive countries, institutions and authors, citation patterns, distribution of subject categories, languages that have been more investigated, geographic distributions and keywords bursts and trends. Similarly, other scholars have used various bibliometric techniques and the WoS database to study the research status of the field of big data in general[16,17] and its application in medicine.[18] These papers mainly differ in the number of publications analyzed and the timespan of the analysis as well as some of the indicators they offer.

One field of research which has been frequently studied with bibliometric methods is machine learning. Some of the analyses focus on particular applications such as the context of public health[19] and use various databases (Science direct, Scopus and Web of Science) to generate several indicators as well as identifying the most studied topics in public health and the machine learning techniques, programming languages and software applications used most frequently by researches. Similarly, the application of machine learning techniques used to deal with cyber security threats has been studied from this perspective.[20] Others concentrate for example in the historical progress and current situation of support vector machines in specific countries employing a comprehensive bibliometric analysis;[21] or in the in-depth study of a particular journal during several years.[22] Using a different approach, more general studies attempt to explain the intellectual structure and dynamics of the field of machine learning by means of analyzing publications indexed on the WoS database and also patents from the Derwent Innovation Index.[23] Related to the same subtheme, Mao *et al*.[24] report a comprehensive bibliometric study of the research status, trends and hotspots

of deep learning research by analyzing 3, 599 papers indexed in the WoS database from 1968 to 2018.

Artificial Intelligence (AI) is another field which has attracted the attention of bibliometric studies such as the work oriented to analyze the research landscape of AI applied to depression research and treatment to evaluate the productivity of researchers and institution in the field.[25] Similarly, the work devoted to providing an historical and comprehensive picture of research on AI use in health and medicine,[26] which analyzed 27, 451 papers, published between 1977 and 2018, retrieved from the WoS database. Also, on this subject, Alejo-Machado *et al*.[27] Carried out a bibliometric study of the scientific output on learning to rank, a leading research topic in the field of AI and information retrieval, covering the years from 2000 to 2013 and employing 627 records retrieved from Scopus.

With fewer bibliometric papers, there are studies devoted to other subfields of NLP, such as the one about the dynamic knowledge evolution of emerging information technologies in cancer literature[28] This analyzed 7,136 articles published between 2000 and 2017 and provides a visual display of knowledge evolution by means of the analysis of time sequence changes, spatial distributions, knowledge base and hotspots. Other research reviews some bibliometric indicators of automatic text processing employing records from the *Science Direct* database and the Russian e-library.[29] This uses differential indicators of speed and acceleration to evaluate the dynamics of NLP subdomains, identifying areas of high growth rates and those that have lost previous existing dynamics of growth. Using more common bibliometric indicators, there is a work that attempts to provide a theoretical clarification of the health informatics field[30] by conducting a quantitative analysis of the relevant literature. The study particularly aims to uncover the explicit and hidden patterns, knowledge structures and substructures in scientific networks; to track the flow and burst of scientific topics; and to discover the effects they have on the growth of health informatics. There is also a quantitative bibliometric analysis about conversational agents,[31] based on records from the WoS database and ProQuest.

Additionally, there are some close studies concerning the state of the art of particular topics, for instance the subject of automatic summarization in which new phrases are generated with the use of natural language processing.[32] Further work about the state of the art of recommendation expert systems has been carried out, studying those systems employing a combination of NLP methods to generate lists of persons with greater knowledge and experience on certain fields.[33] There is also the bibliometric study that compares the situation and characteristics of medical informatics in China, United States and Europe.[34] Furthermore, there are studies that do not use bibliometric methods but focus on analyzing the state of the art of specific topics of NLP, for example, on deep learning,[35] about sentiment analysis in Twitter[36] and the study about machine learning applied to energy systems.[37] However, as far as we know, there are not in the mainstream literature, studies that attempt to identify by bibliometric means, the general state and trends of research on the broad topic of NLP, specifically oriented to identify synchronically and diachronically the importance and relationships of natural language processing subthemes.

MATERIALS AND METHODS

We used the database Web of Science of Clarivate Analytics, since this source includes the journals and conference proceedings that are considered more relevant by the scientific communities, as well as for their constancy and periodicity. The WoS database is considered the highest quality index for scientific publications[38] and also include a high number of conference publications, which are an important source of recent and relevant knowledge in the field of natural language processing.

Our point of departure was to recover the bibliographic records of publications classified under the topic of natural language processing, from 2000 to 2019, to have an insight of the last twenty years of research in this area. Since we are interested in discovering the subjects and subthemes related to NLP, our query was simple: TOPIC: ("natural language processing"). Timespan: 2000-2019. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI. We retrieved 15,639 records within these criteria.

Subsequently, these records were processed in different programs for bibliometric and network analysis: Bibexcel, Science of Science Tool (Sci2) and VOSviewer.[39-41] Once the analyses were carried out, some preprocessing of the structured text was needed to normalize author names and to eliminate plurals. With these software tools, network graphs and different calculations were made to identify the countries, institutions and relevant authors, as well as the research areas that have been developed concerning natural language processing. We particularly analyzed the type of communication sources more frequently used to publish research results; the degrees of citation of countries and institutions; the relevant active authors during the period of analysis and cited authors (corresponding respectively to the research front and the knowledge base authors); and the co-occurrence of key words employed to identify their articles, revealing their research subjects. We opted to use co-occurrence analysis of keywords, which is a well-documented method in bibliometric studies.[2] It relies on the assumption that authors select at least three to six representative

keywords which describe the main topics included in their papers and therefore could be considered as an approximation to the contents of the whole article. This issue is analyzed in G. Chen & Xiao, 2016[42] and recent works using this method include Ellegaard & Wallin and Kalantari *et al*., [16,43] as well as several of the papers mentioned in the previous section. Additionally, we also used the WoS KeyWords Plus, which are derived using an algorithm that identifies the terms in the titles of cited articles in a paper but that do not appear in the title of the citing article itself.

The co-occurrence of keywords was carried out with VOSviewer, which uses a Visualization of Similarities algorithm (VOS) as an alternative to multidimensional scaling. This "aims to locate items in a low-dimensional space in such a way that the distance between any two items reflects the similarity or relatedness of the items as accurately as possible". [44] In addition, we performed a burst detection analysis on author's keywords, using the Kleinberg algorithm[45] implemented in Sci2. With this, it is possible to identify the bursts of activity of topics on a temporal line, revealing a latent hierarchical structure with a meaning in terms of the content of the stream of words.

RESULTS AND DISCUSSION

As mentioned before, the records of published works identified span from 2000 to 2019; these show an increasing trend of publication from 138 in 2000 to a peak of 2377 records in 2019 with a growth rate of 15.21% during the period of analysis (Figure 1). According to the type of documents, the publication on conference proceedings is the type of preference (54.57%), followed by journal articles (35.67%). There is a large number of sources in which papers concerning the topic of natural language processing are published, however, there is not a highly preferred one; Table 1 shows those media where the most influential papers, either advancing the knowledge on NLP or applying it, were published, according to the number of citations during the period. Regarding other indicators, the Journal of Machine Learning Research has a significantly higher average number of citations per paper while the Journal of The American Medical Informatics Association and the Journal of Biomedical Informatics have a higher *h*-index.

Regarding the countries where research is carried out, the most productive countries (with more than one thousand

Table 1: Main sources where research was published, 2000-2019.

Source	Documents	Citations	ACPP	*h*-index
Journal of the American Medical Informatics Association	224	8272	36.93	43
Journal of Biomedical Informatics	231	5739	24.84	38
Journal of Machine Learning Research	20	3978	198.90	14
Machine learning	20	2112	105.60	12
BMC Bioinformatics	94	1907	20.29	23
Bioinformatics	36	1760	48.89	17
Expert system with applications	94	1749	18.61	23
Journal of Artificial Intelligence Research	22	1723	78.32	11
Computational Linguistics	57	1488	26.11	17
International Journal of Medical Informatics	82	1297	15.82	21
IEEE Transactions on Knowledge and Data Engineering	35	1130	32.29	17
Journal of the American Society FOR Information Science and Technology	29	1112	38.34	12
Knowledge-based system	54	1108	20.52	15
Information Processing and Management	65	960	14.77	17
Neurocomputing	46	911	19.80	12
Language Resources and Evaluation	62	878	14.16	11
Artificial Intelligence	25	744	29.76	14
Nucleic Acids Research	7	729	104.14	5
Journal of Medical Internet Research	47	724	15.40	15
Briefings in Bioinformatics	10	719	71.90	9

ACPP = Average Citations Per Paper

publications) are the United States (4246 documents), the People's Republic of China (2337) and India (1114). However, if we have a look at the average number of citations per paper, we can see that more influential papers have been published in Northern Ireland (18.72 citations per paper), Scotland (16.38), Singapore (16.30) and Israel (15.31). These results are most likely related to the fact that an influential researcher or group were active in the country during the period under study. Concerning the academic organizations which perform research on natural language processing or its applications, Table 2 shows those which are more productive and have a larger number of citations (more than 2000 citations). According to the number of citations, researchers at Columbia University, Harvard University and Mayo Clinic have published the more influential papers on the field. However, if we look at the *h*-index of the organizations, the best ranked are Vanderbilt University, Harvard University and Mayo Clinic. It has to be observed that in the case of organizations, one researcher or group or one highly cited paper can be responsible for the high rank of the organization.

Similarly, Table 3 depicts the most productive authors and those with a higher number of citations during the period under study, with Jason Weston, Christopher G. Chute and Carol Friedman as the more cited authors. The first works as Research Scientist at Facebook and is Visiting Research Professor at New York University. His research interests are statistical machine learning with a focus on reasoning, memory, perception, interaction and communication. The second is the Bloomberg Distinguished Professor of Health Informatics, Professor of Medicine, Public Health and Nursing

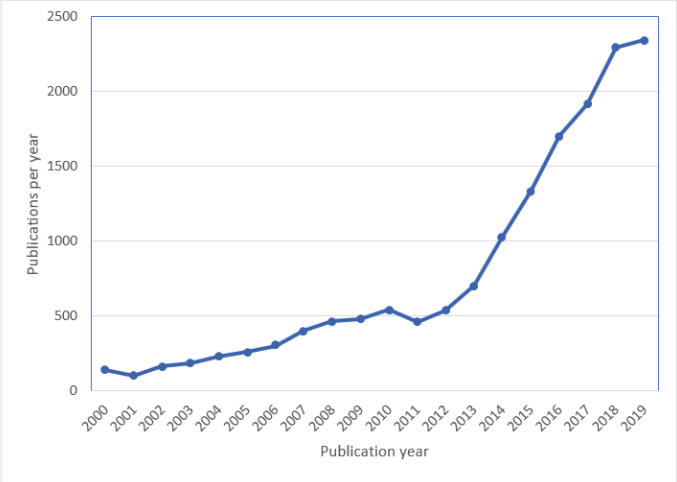

Figure 1: Number of publications on NLP, 2000-2019.

Table 2: Main organizations where research is carried out, 2000-2019.

Organization	Documents	Citations	ACPP	*h*-index
Columbia University	149	4779	32.07	27
Harvard University	108	3196	29.59	33
Mayo Clinic	113	2931	25.94	31
University of Pittsburgh	99	2644	26.71	18
University of Utah	165	2576	15.61	23
Vanderbilt University	94	2387	25.39	45
Stanford University	132	2341	17.73	26
Nanyang Technological University	88	2256	25.64	26
New York University	45	2216	49.24	9
Rutgers State University	24	2211	92.13	6
Massachusetts IT	106	2202	20.77	21
Cornell University	25	2095	83.80	5
University of Colorado	90	2014	22.38	22

ACPP = Average Citations Per Paper

Table 3: Main authors (active) publishing during 2000-2019.

Author	Documents	Citations	ACPP	*h*-index
Weston, Jason	5	2158	431.60	5
Chute, Christopher G.	17	1980	116.47	16
Friedman, Carol	47	1950	41.49	25
Chapman, Wendy W.	34	1761	51.79	17
Denny, Joshua C.	38	1709	44.97	21
Savova, Guergana	34	1627	47.85	13
Cambria, Erik	27	1315	48.70	17
Hripcsak, George	21	1311	62.43	16
XU, Hua	63	1167	18.52	19
Aronson, AR	7	1042	148.86	11
Sohn, Sunghwan	27	915	33.89	11
Uzuner, Oezlem	7	889	127.00	12
Poria, Soujanya	10	707	70.70	7
Liu, hongfang	65	684	10.52	8
Szolovits, Peter	18	588	32.67	12
Demner-Fushman, Dina	15	573	38.20	9
Cai, Tianxi	24	570	23.75	4
South, Brett R.	14	558	39.86	10
Murphy, Shawn N.	18	543	30.17	12
Pathak, Jyotishman	12	541	45.08	9
Roden, Dan M.	8	532	66.50	8
Cardie, C	5	517	103.40	3
Shen, Shuying	13	513	39.46	2
Luo, Yuan	17	509	29.94	1
Duvall, Scott L.	24	504	21.00	9

ACPP = Average Citations Per Paper

at Johns Hopkins University and Chief Research Information Officer for Johns Hopkins Medicine. He has focused on how to represent clinical information to support analyses and inferencing and has a deep interest in semantic consistency, harmonized information models and ontology. The last, from Columbia University, is a Professor of Biomedical Informatics and Director of the Department's Graduate Training Program. She was a pioneer in the use of NLP in the field of medical informatics and her current research is devoted to the use of natural language processing to obtain executable data and knowledge from clinical reports and biomedical text. If we focus our attention on the Average Citations Per Paper (ACPP), we find that the order of the authors changes a little, with Jason Weston, A. R. Aronson, Oezlem Uzuner and Christopher Chute as those with a higher ACPP. Regarding the *h*-index, the authors with higher ranks are Carol Friedman, Joshua Denny and Hua Xu. It is important to note that the number of times cited could represent citations in papers which are outside of the topic of natural language processing which we are investigating. In addition, some of these authors are either directly working on natural language processing subjects or applying NLP within other domains, which is why it is important to further process the results as we shall see below.

Therefore, another important indicator is who are the authors whose work is more frequently cited during the period under study. In this regard, the most cited author is Tomas Mikolov, currently at the Czech Institute of Informatics, Robotics and Cybernetics, whose paper (with four coauthors): "Distributed Representations of Words and Phrases and their Compositionality" (2013) is the most cited in the retrieved records with 748 citations. However, this and other works were cited a total of 2319 times. The second most cited author is Christopher D Manning of Stanford University, who has three highly cited papers among which: "The Stanford CoreNLP Natural Language Processing Toolkit" (2014), written with five coauthors, is the most cited with 314 citations and a total of 958. The third most cited author is David Blei, working at Columbia University, with 621 citations of the paper: "Latent Dirichlet allocation" (2003) and a total of 874 citations in the period 2000–2019 (Table 4 and Figure 2). If we take a look at the average citations per paper, the order of the author's changes, with Sepp Hochreiter, Christiane Fellbaum and Ronan Collobert as those with a higher rank. Paying attention to the *h*-index, the best ranked cited authors are Tomas Mikolov and Carol Friedman. In Table 4 we have included, for reference, the *h*-index calculated in the Web of Science for the most cited authors.

Regarding the content of the records obtained we performed an analysis and visual representation of the frequency of terms and their co-occurrences, that describe the publications during the period under study. To do this we used the keywords that the authors suggest characterizing their contributions as well as the WoS KeyWords Plus (Figure 3). We can identify five

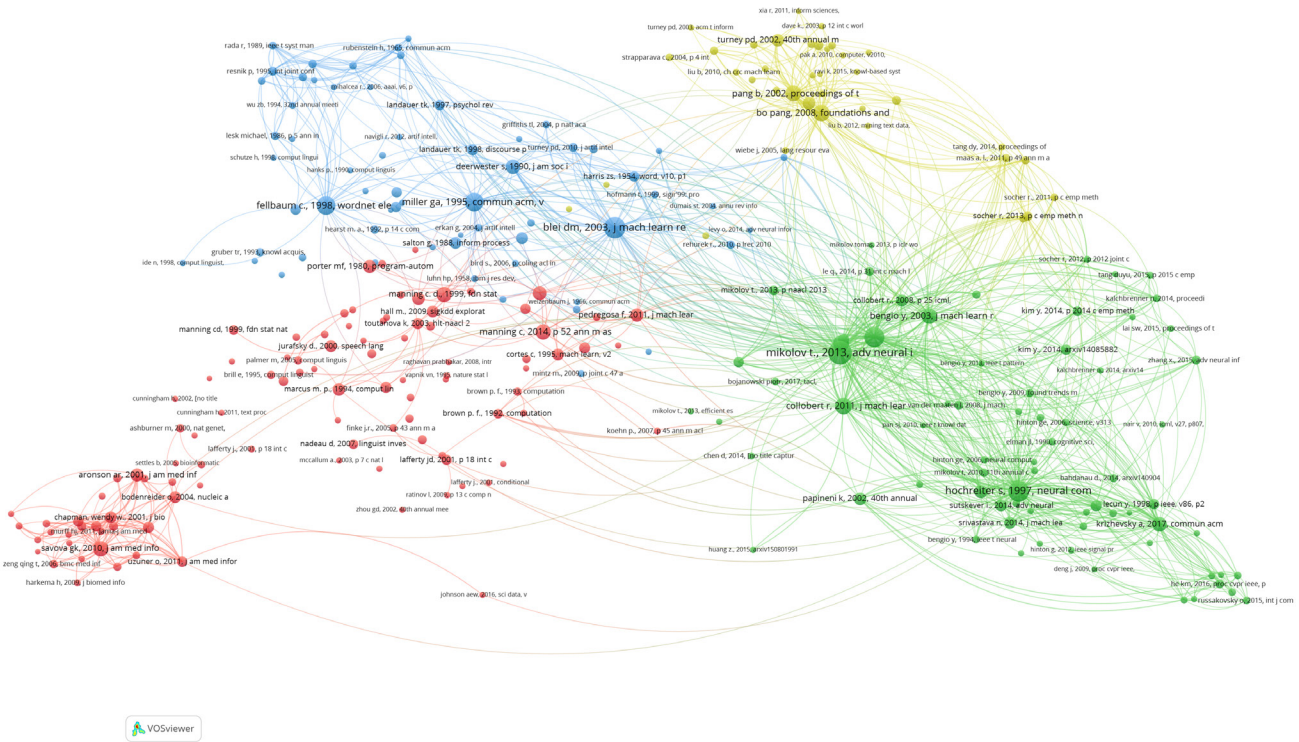

Figure 2: Most cited authors and their works in the papers published during 2000-2019.

major clusters and in the center of them, for obvious reasons, the most occurring keyword is natural language processing (6521 occurrences). The first cluster (bottom right) and with strong linkages to NLP there are the subthemes of deep learning (561), artificial intelligence[1] (291) word embedding's (253) and neural networks (258); also in the same cluster, but with less occurrences, we find text classification (211) convolutional neural networks (168) and machine translation (178). It should be noted that many of these terms correspond also to subthemes of artificial intelligence (AI). A second cluster (bottom middle) includes sub-topics such as machine learning (1077), also related to AI, sentiment analysis (724) and opinion mining (247), text mining (623), classification (501), social media (275) and data mining (278) (Figure 4). The majority of these terms correspond to direct applications of natural language processing. The third cluster (right) mainly consists of terms corresponding to keywords and subthemes such as systems (486), information (391), text (322), electronic health records (238), extraction (253) and big data (203). The fourth cluster (upper right) has nodes with less occurrences including subtopics and keywords such as information retrieval (413), ontology (388), knowledge (232), word sense disambiguation (152), semantic web (153) and semantic similarity (142). The fifth cluster (upper left) has several links with other groups (Figure 5), it consists of keywords such as information extraction (542), named entity recognition (245), machine translation (178) and corpus (166).

Regarding the temporal analysis involving burst of keywords in particular periods (2000 to 2019) (Figure 6), we can see that between 2000 and 2010, research interests concentrated on areas with low weight or intensity but large periods, related to information retrieval, knowledge representation and management, machine translation, algorithms, semantic web and with more strength, ontologies. From 2010 to 2014 there are sub-themes that continue from previous years such as

1 Hidden label in the general graph.

Table 4: Cited authors in the literature published during 2000 to 2019.

Author	Papers	Citations	ACPP	*h*-index	*h*-index global WoS
Mikolov, Tomas	15	2319	154.60	15	10
Manning, Christopher D	7	958	136.86	7	27
Blei, David M	4	874	218.50	4	32
Miller, George A	4	750	187.50	4	16
Bengio, Yoshua	6	746	124.33	6	54
Pang, BO	4	735	183.75	4	N/A
Friedman, Carol	12	721	60.08	12	37
Salton, Gerald	6	716	119.33	6	N/A
Socher, Richard	9	711	79.00	9	9
Fellbaum, Christiane	2	685	342.50	2	9
Pennington, jeffrey	4	668	167.00	4	11
Collobert, Ronan	2	665	332.50	2	13
Turney, Peter D	6	659	109.83	6	18
Hochreiter, Sepp	1	657	657.00	1	25
Liu, Bing	8	614	76.75	8	25
Mihalcea, Rada	4	582	145.50	4	17
Jurafsky, Dan	3	568	189.33	3	21
Aronson, Alan R	3	564	188.00	3	21
Lecun, Yann	3	545	181.67	3	46
Kim, Yoon	5	532	106.40	5	2
Uzuner, Ozlem	6	529	88.17	6	18
Navigli, Roberto	5	507	101.40	5	26
Cambria, Erik	4	468	117.00	4	44
Joachims, Thorsten	4	463	115.75	4	21
Landauer, Thomas K	2	459	229.50	2	31

ACPP = Average citations per paper

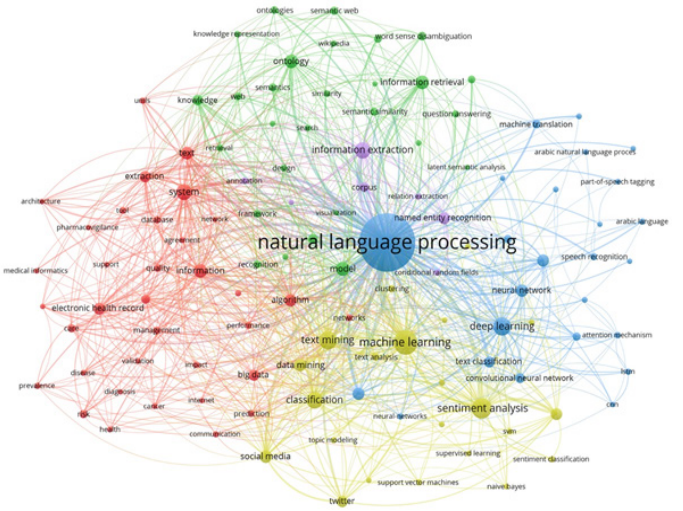

Figure 3: Graph of author and WoS keywords for papers classified under the topic of NLP, 2000-2019.

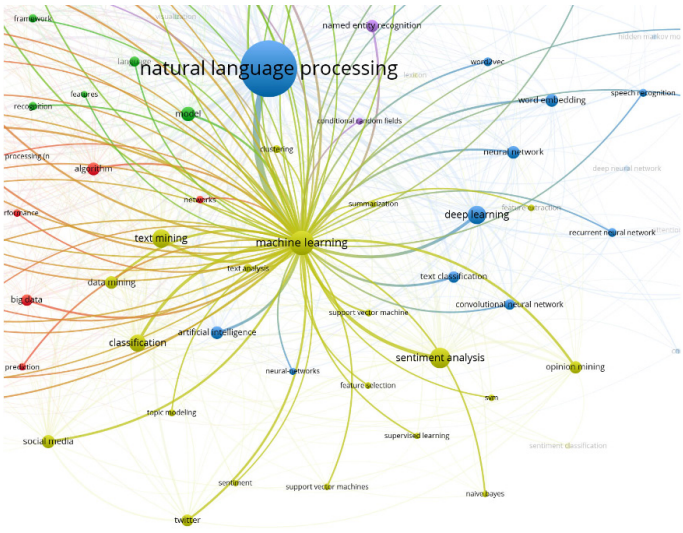

Figure 4: Cluster where machine learning is the most relevant term.

ontology and ontology learning, annotation and algorithms. Between 2015 and 2017 the terms that emerge as research trends are big data, opinion mining and sentiment analysis. From 2017 to 2019 and with a little higher weighting, techniques and applications related to artificial intelligence as word embedding's, recurrent neural networks have interest among researchers. From 2018 to 2019 several interrelated sub-topics emerge some of them with very high weighting, such as attention mechanism, LSTM (Long Ahort-Term Memory), convolutional neural networks and deep learning. Finally, between 2019 and 2020 there appears some interest on artificial intelligence, electronic health records and social media.

As mentioned above, while there are a large number of studies about particular subthemes of natural language processing or topics related to it, there are not, as far as we know, integral studies of NLP showing the areas of research interest and the emerging trends in the field as we have presented in this article. These results are of particular interest for researchers in the field of natural language processing and computational linguistics.

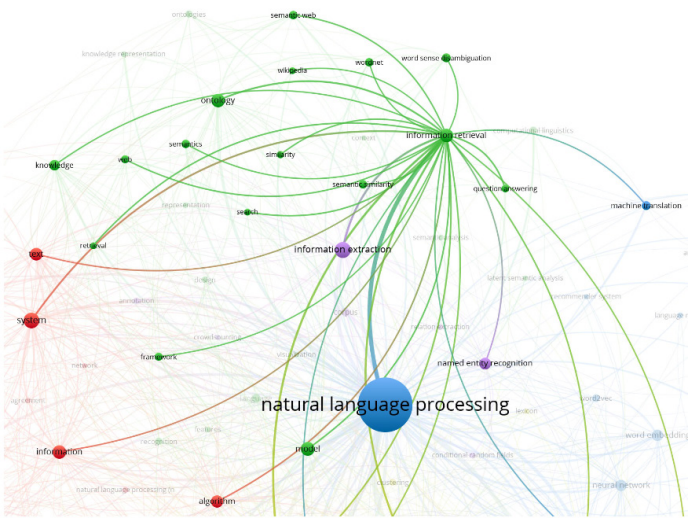

Figure 5: Cluster where information retrieval and extraction are the more relevant terms.

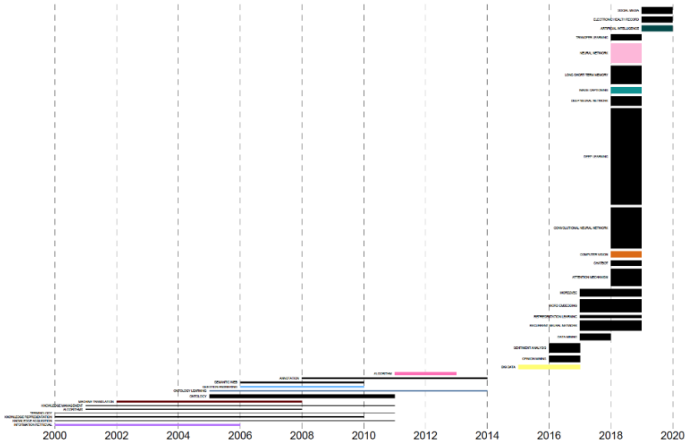

Figure 6: Temporal bar graph of bursting author keywords, 2000-2019.

CONCLUSION

In this work, we used bibliometric methods to study the state of research of a particular knowledge area. Specifically, the development of natural language processing, identifying the main sources where research papers are published, the most important countries and institutions where research is carried out on NLP and its applications, the individual top active researchers, the most cited authors in the field, as well as the main topics of interest on natural language processing.

Firstly, during the period under study we have identified a continuous growth of research on natural language processing internationally; publications tend to be communicated mainly through media such as conference proceedings –which is the most preferred communication media for NLP research and scientific journals.

Secondly, we presented the main academic sources where NLP research has been published in the last ten years, as well as the countries, research organizations and authors that have carried out research on natural language processing. The conferences and journals where more papers are published are dedicated to subjects related to artificial intelligence, computational linguistics and biomedical informatics. The most prolific countries are the United States, People's Republic of China and India, but the most influential research is carried out in Northern Ireland, Scotland, Singapore and Israel. The organizations which perform a higher number of research projects reflected in publications are in the United States.

Thirdly, we performed a synchronic analysis of the main keywords that describe research in the field; natural language processing, machine learning, sentiment analysis, text mining and deep learning –some of them related to artificial intelligence, are the research subjects which are more frequently used to describe publications in the area. Finally, a temporal analysis of bursting author keywords showed that research on convolutional neural networks and deep learning are the areas of major interest in recent years, these two topics have also a relationship with research on artificial intelligence.

Our study is a comprehensive view of the whole area of natural language processing, covering all of the subjects that comprise this research topic. NLP is a field which has had an important growth particularly during the last ten years. It has applications in multiple areas among which medical informatics is one of the most important as we have shown in this work. We have also shown that, particularly in recent years, it has frequent interactions with subjects related to artificial intelligence, such as deep learning, machine learning and neural networks. Our findings can be of relevance for researchers of the diverse areas of NLP, as well as its applications.

ACKNOWLEDGEMENT

We are greatly indebted to Karla Lopez and Christopher Russell for their comments on the final draft of this paper.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Henkel M, Perjons E, Sneiders E. Examining the potential of language technologies in public organizations by means of a business and IT architecture model. International Journal of Information Management. 2017;37(1 Part A):1507-16. doi: 10.1016/j.ijinfomgt. 2016.05.008
2. DeSolla PDJ. Networks of Scientific Papers. Science. 1965;149(3683):510-15. doi:10.1126/science.149.3683.510
3. Crane D. Invisible colleges; diffusion of knowledge in scientific communities. Chicago, IL: University of Chicago Press. 1972.
4. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. An approach for detecting, quantifying and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. Journal of Informetrics. 2011;5(1):146-66. doi: 10.1016/j.joi.2010.10.002
5. Zupic I, Čater T. Bibliometric Methods in Management and Organization. Organizational Research Methods. 2015;18(3):429-72. doi:10.1177/1094428114562629
6. Yeung AWK, Goto TK, Leung WK. A bibliometric review of research trends in neuroimaging. Current Science. 2017;112(4):725-34.
7. Callon M, Courtial JP, Turner WA, Bauin S. From translations to problematic networks: An introduction to co-word analysis. Information (International Social Science Council). 1983;22(2):191-235. doi:10.1177/053901883022002003
8. Callon M, Courtial JP, Laville F. Co–word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. Scientometrics. 1991;22(1):155-205.
9. Randhawa K, Wilden R, Hohberger J. A Bibliometric Review of Open Innovation: Setting a Research Agenda. Journal of Product Innovation Management. 2016;33(6):750-72. doi:10.1111/jpim.12312
10. Börner K, Chen C, Boyack KW. Visualizing knowledge domains. Annual Review of Information Science and Technology. 2003;37(1):179-255. doi:10.1002/aris.1440370106
11. Liang TP, Liu YH. Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study. Expert Systems with Applications. 2003;111:2-10. doi: 10.1016/j.eswa.2018.05.018
12. Chen X, Ding R, Xu K, Wang S, Hao T, Zhou Y. A Bibliometric Review of Natural Language Processing Empowered Mobile Computing. Wireless Communications and Mobile Computing. 2018.
13. Mäntylä MV, Graziotin D, Kuutila M. The evolution of sentiment analysis: A review of research topics, venues and top cited papers. Computer Science Review. 2018;27:16-33. doi: 10.1016/j.cosrev.2017.10.002
14. Keramatfar A, Amirkhani H. Bibliometrics of sentiment analysis literature. Journal of Information Science. 2018;45(1):3-15. doi:10.1177/0165551518761013
15. Piryani R, Madhavi D, Singh VK. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. Information Processing and Management. 2017;53(1):122-50. doi: 10.1016/j.ipm.2016.07.001
16. Kalantari A, Kamsin A, Kamaruddin HS, AleEbrahim N, Gani A, Ebrahimi A, *et al*. A bibliometric approach to tracking big data research trends. Journal of Big Data. 2017;4(1):30. doi:10.1186/s40537-017-0088-1
17. Xu Z, Yu D. A Bibliometrics analysis on big data research (2009–2018). Journal of Data, Information and Management. 2019;1(1):3-15. doi:10.1007/s42488-019-00001-2
18. Liao H, Tang M, Luo L, Li C, Chiclana F, Zeng XJ. A Bibliometric Analysis and Visualization of Medical Big Data Research. Sustainability. 2018;10(1):166.
19. DosSantos BS, Steiner MTA, Fenerich AT, Lima RHP. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. Computers and Industrial Engineering. 2019;138:106120. doi:10.1016/j.cie.2019.
20. Makawana PR, Jhaveri RH. A Bibliometric Analysis of Recent Research on Machine Learning for Cyber Security. Intelligent Communication and Computational Technologies. 2018;19:213-26.
21. Yu DJ, Xu ZS, Wang XZ. Bibliometric analysis of support vector machines research trend: A case study in China. International Journal of Machine Learning and Cybernetics. doi:10.1007/s13042-019-01028-y
22. Xu ZS, Yu DJ, Wang XZ. A bibliometric overview of International Journal of Machine Learning and Cybernetics between 2010 and 2017. International Journal of Machine Learning and Cybernetics. 2019;10(9):2375-87. doi:10.1007/s13042-018-0875-9
23. Bhattacharya S. Some Salient Aspects of Machine Learning Research: A Bibliometric Analysis. Journal of Scientometric Research. 2019;8(2):S85-92. doi:10.5530/jscires.8.2.26
24. Mao M, Li Z, Zhao Z, Zeng L. Bibliometric Analysis of the Deep Learning Research Status with the Data from Web of Science. Data Mining and Big Data. Cham: Springer. 2018;10943.
25. Tran BX, McIntyre RS, Latkin CA, Phan HT, Vu GT, Nguyen HLT, *et al*. The Current Research Landscape on the Artificial Intelligence Application in the Management of Depressive Disorders: A Bibliometric Analysis. International Journal of Environmental Research and Public Health. 2019;16(12):2150. doi:10.3390/ijerph16122150
26. Tran BX, Vu GT, Ha GI, Vuong QH, Ho MT, Vuong TT, *et al*. Global Evolution of Research in Artificial Intelligence in Health and Medicine: A Bibliometric Study. Journal of Clinical Medicine. 2019;8(3):360. doi:10.3390/jcm8030360
27. Alejo-Machado OJ, Fernandez-Luna JM, Huete JF. Bibliometric study of the scientific research on "Learning to Rank" between 2000 and 2013. Scientometrics. 2015;102(2):1669-86. doi:10.1007/s11192-014-1467-4
28. Wang XY, Guo JJ, Gu DX, Yang Y, Yang XJ, Zhu KY. Tracking knowledge evolution, hotspots and future directions of emerging technologies in cancers research: A bibliometrics review. Journal of Cancer. 2019;10(12):2643-53. doi:10.7150/jca.32739
29. Barakhnin VB, Duisenbayeva AN, Kozhemyakina OY, Yergaliyev YN, Muhamedyev RI. The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area. Journal of Physics: Conference Series. 2018;1117(1):012001. doi:10.1088/1742-6596/1117/1/012001
30. Saheb T, Saheb M. Analyzing and Visualizing Knowledge Structures of Health Informatics from 1974 to 2018: A Bibliometric and Social Network Analysis. Healthcare Informatics Research. 2019;25(2):61-72. doi:10.4258/hir.2019.25.2.61
31. Io HN, Lee CB. Chatbots and Conversational Agents: A Bibliometric Analysis. In 2017 Ieee International Conference on Industrial Engineering and Engineering Management New York: IEEE; 2017. p.215-9.
32. Gupta S, Gupta SK. Abstractive summarization: An overview of the state of the art. Expert Systems with Applications. 2019;121:49-65. doi: 10.1016/j.eswa.2018.12.011
33. Nikzad–Khasmakhi N, Balafar MA, Feizi–Derakhshi RM. The state-of-the-art in expert recommendation systems. Engineering Applications of Artificial Intelligence. 2019;82:126-47. doi: 10.1016/j.engappai.2019.03.020
34. Jia Y, Wang W, Liang J, Liu L, Chen Z, Zhang J, *et al*. Trends and characteristics of global medical informatics conferences from 2007 to 2017: A bibliometric comparison of conference publications from Chinese, American, European and the Global Conferences. Computer Methods and Programs in Biomedicine. 2018;166:19-32. doi: 10.1016/j.cmpb.2018.08.017
35. Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, *et al*. A State-of-the-Art Survey on Deep Learning Theory and Architectures. Electronics. 2019;8(3):292. doi:10.3390/electronics8030292
36. Zimbra D, Abbasi A, Zeng D, Chen HC. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. Acm Transactions on Management Information Systems. 2018;9(2):29. doi:10.1145/3185045
37. Mosavi A, Salimi M, Ardabili SF, Rabczuk T, Shamshirband S, Varkonyi-Koczy AR. State of the Art of Machine Learning Models in Energy Systems: A Systematic Review. Energies. 2019;12(7):42. doi:10.3390/en12071301
38. Li K, Rollins J, Nad YE. Web of Science use in published research and review papers 1997–2017: A selective, dynamic, cross-domain, content-based analysis. Scientometrics. 2018;115(1):1-20. doi:10.1007/s11192-017-2622-5
39. DeNooy W, Mrvar A, Batagelj V. Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software. Third Edition. New York: Cambridge University Press. 2018.
40. Persson OR, Danell J, Schneider W. How to use Bibexcel for various types of bibliometric analysis. Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday. Leuven, Belgium: International Society for Scientometrics and Informetrics. 2009;9-24.
41. Eck NJV, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. ;84(2), 523-538. doi:10.1007/s11192-009-0146-3
42. Chen G, Xiao L. Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. Journal of Informetrics. 2016;10(1):212-23. doi: 10.1016/j.joi.2016.01.006
43. Ellegaard O, Wallin JA. The bibliometric analysis of scholarly production: How great is the impact?. Scientometrics. 2015;105(3):1809-31. doi:10.1007/s11192-015-1645-z
44. Eck NJ, Waltman L, Dekker R, DenBerg JV. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. Journal of the American Society for Information Science and Technology. 2010;61(12):2405-16. doi:10.1002/asi.21421
45. Kleinberg J. Bursty and Hierarchical Structure in Streams. Paper presented at the Proc. 8th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining. 2002.