Year 2022 in Medical Natural Language Processing: Availability of Language Models as a Step in the Democratization of NLP in the Biomedical Area

Cyril Grouin1, Natalia Grabar2, Section Editors for the IMIA Yearbook Section on Natural Language Processing

1 Université Paris Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400 Orsay, France
2 UMR8163 STL, CNRS, Université de Lille, Domaine du Pont-de-bois, 59653 Villeneuve-d’Ascq cedex, France

Summary

Objectives: To analyse the content of publications within the medical Natural Language Processing (NLP) domain in 2022.

Methods: Automatic and manual preselection of publications to be reviewed, and selection of the best NLP papers of the year. Analysis of the important issues.

Results: Three best papers have been selected. We also propose an analysis of the content of the NLP publications in 2022, stressing on some of the topics.

Conclusion: The main trend in 2022 is certainly related to the availability of large language models, especially those based on Transformers, and to their use by non-NLP researchers. This leads to the democratization of the NLP methods. We also observe the renewal of interest to languages other than English, the continuation of research on information extraction and prediction, the massive use of data from social media, and the consideration of needs and interests of patients.

Keywords
Natural language processing; best paper; topics; issues; 2022

1 Introduction

Natural Language Processing (NLP) aims at providing methods, tools and resources designed in order to mine textual and narrative documents, and to make it possible to access the information they convey [1]. While human languages are complex (as an example, learning a human language requires many years in order to be fluent), the importance of using NLP approaches to mine documents produced by humans has been pointed out for a long time [2]. In this synopsis, we first present the selection process applied this year and then we analyze the content of some publications. More particularly, we will focus on several important trends, as well as the originality of the research questions addressed in 2022.

2 The Selection Process

In order to identify all papers published during the year 2022 in the field of NLP for the biomedical domain, we queried two databases: MEDLINE1, specifically dedicated to the biomedical domain, and the ACL anthology2, a database that brings together the major NLP conferences (ACL, COLING, EMNLP, LREC, NAACL, etc.) and journals, since some NLP studies concerning the biomedical domain are published in conferences and journals which are not indexed by PubMed.

We applied on MEDLINE the basic query we defined in 2018 and we used since then (Figure 1): all journal papers published in English in 2022, having abstract, and composed of sequences “clinical language processing” or “medical language processing” or “natural language processing”.

We collected 1,670 entries, which is much more than last year (1,204 entries found in 2021). We applied a similar query on the ACL anthology database and collected 20 additional entries. In order to process those 1,690 papers, we automatically scored the papers. While we defined a set of rules in 2018 to filter out each year the papers, we refined the automatic scoring this year: a score ranging from 0 to 1 is given to each candidate, for five criteria.

Fig. 1 Query used for collecting candidate publications for review.

2 https://www.aclweb.org/anthology/
(journal name, objectives, methods/corpus/resources used in the paper, evaluation/metrics used, special concepts/key phrases used in the abstract). The underlying idea is to focus on the NLP community’s custom in terms of place where papers are published, keywords and metrics the community generally used, and phrases that highlight the use or the design of a method. Since “language” may refer either to the study object (i.e., language productions) or to a cognitive aspect, we considered positively papers published in the NLP-related conferences (AMIA, MedInfo, MIE, etc.) or some journals (Journal of the American Medical Informatics Association, Journal of Biomedical Informatics, International Journal of Medical Informatics, etc.) rather than cognitive journals (concerning brain, neuroscience, or psychiatrics research). We also excluded yearbooks and survey papers since they do not fit criterion for best paper candidates. In order to focus on original contributions, we gave a lower score to abstracts that specifically mention phrases like “using natural language processing” or “perform a natural language processing analysis” since the NLP dimension is not central in those submissions, and consequently, those papers are not good candidates for NLP best papers.

For each of the 1,690 candidate papers, the final score ranked from 0.05 to 1 (Figure 2). This score has been used as a meta element during the manual selection of the top-13 papers. Indeed, the section editors did not fully rely on the scores but only used them as additional information. Hence, both section editors independently browsed the abstracts, keywords and automatic scores, and assigned the Yes/Maybe/No score to each paper. All papers having at least one Yes or Maybe score have been kept for the next step of the selection. At this stage, 116 candidate papers remained (i.e., a subset of 9.35% of the whole dataset). We then performed an adjudication process, in order to choose the final 13 candidates to be proofread by external reviewers.

This year, we noticed that our process produces an unexpected selection of candidates mainly from English-speaking countries, with an over-representation of papers from the USA. We hypothesize that, in this post-COVID-19 period, researchers from the US maintained their level of submissions while European researchers decreased the number of submitted papers.

In the next sections, we present the main issues and approaches addressed in the pre-selected publications.

3 Current Trends in Biomedical NLP

To present the current trends in biomedical NLP observed over the last year, we propose an analysis of the top 200 citations according to the scores computed automatically. We focus on the topics studied by researchers. For this, we first analyze the keywords provided by the authors (Section 4.1). Second, we analyze the languages addressed (Section 4.2) in this year’s publications. Then, we group our observations along two lines: main approaches used (Section 4.3) and some frequent topics investigated (Sections 4.4 to 4.6). Useless to say that many other topics are also investigated but we cannot mention them all in this synoptic paper.

3.1 Analysis of Keywords from Publications

We studied the distribution of keywords used to index the 200 top papers. This distribution provides a global overview of the keywords. By comparison with the last year, we have observed the following trends in 2022:

- Among the key-words similar to those from last year, we can distinguish two series: source of data (social media, social network, Twitter, Twitter analysis, digital pharmacovigilance, reddit, Twitch) and methods (transformers, learning transfer, word2vec). Let’s notice that the frequency of transformers is significantly increased in 2022;
in a continuation of last year’s observations, several studies address the pandemics, and especially the COVID-19 period. The researchers have now more global data and can provide more exhaustive analyses related to the pandemics, the vaccines, their safety, and global information on this period. Several keywords (Covid-19, covid-19 surveillance, Covid-19 vaccine, Sars-cov-2, vaccine hesitancy, vaccine safety, vaccine adverse events, vaccine misinformation, vaccine acceptance, patient safety, vaccination) confirm the situation. Besides, other keywords represent the methods used for the analysis of these data (sentiment analysis, emotion analysis, emotion detection, opinion mining, patient experience, patient reported outcome measures). Hence, we can say that the COVID-19 period revealed the concern of patients towards the vaccination, which may be explained by the technology used for the creation of vaccines and little knowledge about their efficiency and safety. The first need of population is related to the security, which is now studied through sentiment and emotion analysis. Another need is to check out the veracity of information, which is now studied through the detection of fake news and misinformation, and through fact-checking. Social media remain the main source of content created by patients;

among the new trends, we can find indexing terms related to methods (reinforcement learning, explainable IA, knowledge graph, knowledge graph embeddings) and tasks (disinformation, fake news detection, misinformation, medical errors). We can also notice technological evolutions of AI methods towards the reinforcement learning [3, 4] and knowledge graphs [5, 6].

Even if those keywords give a global view of work done in these papers, they only reflect a small and general part of the work that has been done by the authors. Indeed, keywords are essentially used for indexing papers while we used them to draw an overview of the research done in 2022. Nevertheless, the trends observed within this set of keywords are also reflected in the analysis provided in the following sections.

3.2 Languages Addressed

Contrary to previous COVID-19 and post-COVID-19 years, we finally witness another wave of research on non-English-language data. Indeed, several languages are addressed in the publications in 2022, which we indicate in what follows together with the topics addressed:

- Arabic: sentiment analysis in COVID-19 tweets [7], assessment of web page credibility [8];
- Bengali: detection of depression severity in social media posts [9];
- Chinese: ICD-coding [10], classification of patient-doctor dialogues [11], question-answering [12], clinical named entity recognition [13, 14, 15], prescription recommendation of traditional Chinese medicine [16];
- Danish: extraction of adverse drug reactions (ADRs) from clinical narratives [17];
- Dutch: predicting COVID-19 symptoms from clinical narratives [18], detecting changes in help seeker conversations on a suicide prevention during the COVID-19 pandemic [19];
- French: analysis of tweets about COVID-19 vaccines [20], coding of ADRs from patient reports [21], extraction of explicit and implicit cause-effect relationships from tweets [22], ICD-10 coding [23], construction of cohorts of similar patients [24], processing of electronic medical records [25], understanding of patient’s answers in a French medical chatbot [26];
- German: evaluation of Transformers on clinical notes [27];
- Greek: improving the performance of localized healthcare virtual assistants [28];
- Hindi: classification of COVID-19 texts [29], chatbot for information sexual and reproductive health for young people [30];
- Italian: analysis of social media for quality of life in Parkinson’s patients [31], sentiment analysis of opinion on COVID-19 vaccines [32, 33], estimation of the incidence of infectious disease cases [34];
- Japanese: understanding psychiatric illness [35], detection of adverse events from narrative clinical documents [36];
- Korean: BERT model for processing medical documents [37], sentiment analysis of tweets about COVID-19 vaccines [38];
- Malayalam: generation of synoptic clinical reports [39];
- Polish: prediction of cardiovascular diseases in electronic health records [40];
- (Brazilian) Portuguese: description of an annotated clinical corpus [41], ICD-10 coding [42];
- Serbian: sentiment analysis in COVID-19 tweets [43];
- Spanish: ICD-coding [10, 44], negation and uncertainty detection in clinical narratives [45], training and evaluation of word embeddings for the clinical domain [46];
- Swedish: ICD-10 coding [44];
- Turkish: sentiment analysis of tweets about COVID-19 vaccines [47].

Aside from English, which is addressed in a huge number of publications, the most frequently processed languages are Chinese, Japanese, Spanish, Arabic, and French. We can also notice new and rare languages like Bengali, Malayalam, Hindi, Greek, or Serbian. Another interesting fact is that some publications address multilingual data or data in several languages: ICD-10 coding in English, Spanish and Swedish [44], analysis of social media for quality of life in Parkinson’s patients and their caregivers in English, French, Italian, Spanish, and German [31], term normalization using the UMLS [48].

The research work in English takes undoubtedly advantage of the existing datasets annotated within various challenges (I2B2, N2C2, KDD, etc.) and institutions (like MIMIC-III), as well as data from social media, hospitals, bibliographical datasets, clinical trials, etc. The research in other languages is possible mainly thanks to the availability of data from social media [7, 9, 19, 20, 22, 38, 43, 47] and documents from local hospitals [10, 13, 14, 17, 18, 23, 25, 27, 36, 37, 40, 42]. Besides, this set of works in languages other than English relies on the dedicated language models, which cover a great variety of languages by now. We expect that this trend will continue. We also hope that large annotated datasets in languages other than English will become available for the research.
3.3. Availability of Large Language Models as a Step towards the Democratization of NLP

Language models group together a set of methods based on word embeddings. Such models often encode several levels of linguistic knowledge, hence their efficiency and self-sufficiency. In the past few years, large language models based on Transformers have been created starting with the BERT (Bidirectional Encoder Representations from Transformers) model [49]. Hence, BERT has been conjugated within the biomedical domain through models like SciBERT [50], Clinical BERT [51], BioBERT [52], PubMedBERT [53], or BioM [54]. At the same time, the BERT model has been adapted to languages other than English: French [55, 56], Spanish [57], Dutch [58], Finish [59], Italian [60, 61], Portuguese [62], Japanese [63], etc. In some languages, medicine-specific BERT models have also been proposed, such as French DrBERT [64] or Japanese clinical BERT [65]. This set of Transformer models is successfully exploited for several tasks such as categorization, POS-tagging, semantic similarity, named entity and relation extraction.

It is important to notice that these language models are freely available for research purpose, which opens new possibilities for those interested in their testing and use within various clinical and health-related tasks. We can even tell that the availability of such models leads to the democratization of this NLP approach. Indeed, without the help of NLP researchers, biologists, radiologists, pharmacists and other clinicians can now use these models within the clinical context and we can find several such experiments, such as: analysis of tweets for user opinions and side effects on COVID-19 vaccines [66, 33], fact-checking of posts on COVID-19 vaccines [67], identification of social determinants of health in EHRs [68], analysis of literature for drug-induced liver injury [69], labeling of diagnosis in cardiovascular Magnetic resonance imaging (MRI) [70], analysis of social media on the quality of life in Parkinson’s patients [31], extraction of biomedical relations from the scientific literature [71].

More recently, the GPT (Generative Pre-trained Transformer) models [72] have been proposed for various generation tasks. These models are also coming into the clinical domain but with only few works published in 2022: creation of BioGPT (Generative Pre-trained Transformer for biomedical text generation and mining) [73], prediction and suggestion of medical text in dental medical notes [74], challenges for GPT-3 in ophthalmology [75]. This generative model proposes text on the basis of the training corpora. Even if huge training corpora are used, the models do not cover the whole language and, more specifically, all situations. For instance, such models cannot describe a given patient, his lab or MRI results. This limits seriously the use of GPTs within the clinical domain, in which reliable and realistic data are required. Yet, the GPT-generated text can be used as suggestions which must be verified and approved by human experts. For instance, the application related to the automatic text completion can show its utility since the text is approved by human users.

3.4 Social Media as the Preferred Source of Information

Social media prove to remain the preferred source of information for researchers in several situations:

- when other sources of information are not freely available, such as in languages other than English [7, 9, 19, 20, 22, 38, 43, 47];
- when researchers investigate questions related to patients and population, while these questions are not discussed with medical doctors or require large population samples. We can mention for instance sentiment analysis on medication and vaccines [7, 20, 33, 38, 43, 47], and adverse drug effects [76, 77];
- when mental health of patients is concerned in cases like depression [9, 4, 78], eating disorders [79], suicide detection and prevention [19, 80], quality of life of patients [31, 81], and drug misuse [82-84].

We expect that, for these topics at least, social media continue to occupy an important place in the research community.

3.5 Typical NLP tasks: IE and Prediction

Information extraction (IE) has the purpose to localize within narrative documents exact pieces of information (drug, disorder, age of patients...) and to extract them for further processing. Prediction is related to the classification of texts or text spans into a given class. For several years now, these two tasks are well processed and provide reliable results for a given corpus, class, hospital, etc. Several clinical questions are addressed, such as: extraction of clinical information [15, 17, 85, 86], ICD-10 coding of medical records [10, 23, 42, 44], prediction of diseases [18, 40, 87, 88, 89], mortality [90], risks [91], and patient outcomes [89, 92, 93]. IE and prediction tasks will certainly continue to perform on clinical and health data in the next years. Providing reliable and robust models, which may work on data from different sources, should be one of the objectives for future research.

3.6 Patients as Actors of the Healthcare Process

Recently, patients became one of the true actors within the healthcare process. For instance, their opinions matter in the decision-making process, therapeutic choice, procedures, etc. Besides, medical doctors are interested in discovering the everyday life of patients, their quality of life, their opinion on the healthcare-related aspects, etc. Publications from 2022 represent these research questions. We have already mentioned various issues on the COVID-19 vaccine perception and hesitancy [7, 20, 33, 66]. Other observations on patients are also important, such as classification of self-harm behaviors in EHRs [94], analysis of unnoticed and unresolved safety incidents in patients [95], analysis of web-based reviews of sanctioned physicians [96], and patient experience and satisfaction in online reviews [97].

Since patients are part of the healthcare process, they can access their medical records. In relation with this, one work proposes to study stigmatizing language in the EHRs [98] in order to protect patients. More generally, patients should also be able
to understand the content of medical and clinical documents. The main difficulty is that training and knowledge of patients do not provide the necessary basis for such understanding. One solution is to make clinical methods more patient-friendly thanks to the text simplification [99, 3, 100, 101, 102]. This research question is addressed more frequently now. Even if these works are done on English-language documents, the interest in this issue is important. We expect that, in future, needs and requirements of patients will continue to attract researchers.

4 Conclusion

In 2022, we can observe the evolution of the research in several directions. The most interesting direction is certainly related to the democratization of the NLP methods. Indeed, large language models built with Transformers, which encode several levels of linguistic knowledge, are now freely available for research. Moreover, several of such models are adapted or fine-tuned to the data from the medical domain. This means that clinicians, often without the help of NLP researchers, can use the available language models for different tasks, such as IE, categorization and prediction, coding, or computing of text similarity. This trend does not really contribute to the NLP research, yet it permits to test the available models in different contexts and on different datasets, and to reveal the current limitations of these models.

The technological evolution of AI methods towards the reinforcement learning and knowledge graphs is yet another new trend, which may develop in the following years.

Another evolution is related to the renewal of interest to languages other than English. First, the language models are also adapted to or created for these languages. Second, while standard annotated datasets are rarely available for languages other than English, researchers exploit data from social media and from local hospitals instead. In continuation of the past trends, IE and prediction tasks are widely addressed by researchers and clinicians on various datasets. These methods are quite efficient in precise contexts and on some categories. Let’s also notice that patients continue to occupy an important role within the healthcare process and that several works are dedicated to the needs of patients.

References


Bhasuran B. BioBERT and similar approaches


Grouin et al.
Year 2022 in Medical Natural Language Processing: Availability of Language Models as a Step in the Democratization of NLP in the Biomedical Area


Correspondence to:
Cyril Grouin
Université Paris-Saclay
CNRS, UCN
Campus universitaire
91405 Orsay, France
E-mail: cyril.grouin@lisn.upsaclay.fr
Appendix: Content Summaries of Best Papers for the Natural Language Processing Section of the 2023 IMIA Yearbook

Ahne A, Khetan V, Tannier X, Rizvi MdIH, Czernichow T, Orchard F, Bour C, Aano A, Fagherazzi G

Extraction of explicit and implicit cause-effect relationships in patient-reported diabetes-related tweets from 2017 to 2021: Deep learning approach

JMI Med Inform 2022;10(7):e37201. doi:10.2196/37201

In this paper, the authors aim at providing a deep learning-based method to extract implicit and explicit relations of cause/effect for diabetes from tweets, and a methodology to understand opinions/feelings reported by patients from a causality perspective. They fine-tuned a BERTweet model on 562,000 tweets annotated with emotion information in order to detect causal sentences. Then, they designed a CRF model using BERTweet features to identify possible cause-effects associations from 265,000 causal sentences. This method allows the authors to obtain several clusters for cause-effect (diabetes, death, insulin), including emotions (anger, fear, sadness) reported by diabetes patients.

Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y

A comparative study of pretrained language models for long clinical text


This paper proposes to enrich Transformers models with clinical knowledge, which allows the authors to achieve state-of-the-art results on biomedical NLP tasks. Nevertheless, the authors highlight that the self-attention mechanism uses a lot of memory and does not allow to process long texts (limitation of 512 sub-units: e.g., discharge summaries from MIMIC have 2,984 tokens on average). They produced two domain-enriched language models based on Longformer (Clinical-Longformer) and BigBird (Clinical-BigBird) to process up to 4,096 sub-units. Those models outperformed existing models (BERT, RoBERTa, BioBERT, and ClinicalBERT) on three tasks (NLI @ medNLI; QA @ emrQA-relations; NER @ i2b2 2014). We notice that the source code is available.

Phatak A, Savage DW, Ohle R, Smith J, Mago V

Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning-Based Text Simplification Approach

JMI Med Inform 2022 Nov 18;10(11):e38095. doi: 10.2196/38095

The authors of this paper highlight that abstracts of scientific papers are publicly available, but they are hard to understand due to the use of medical vocabulary. They develop a text simplification method based on deep-learning trained on 3,568 complex-simple paragraphs (training) and evaluated on 480 paragraphs. Several scores are used to evaluate all aspects: FKGL (Flesch-Kincaid Grade Level), ROUGE, SARI (Simplified Automatic Readability Index), Likert scale. In addition, several examples of generated medical paragraphs are given in the paper, including texts generated by other systems (BART fine-tuned, BART-UL, MUSS, Keep-It-Simple, PEGASUS), which allows to compare all produced outputs.