Current Reviews in Musculoskeletal Medicine (2021) 14:392–396
https://doi.org/10.1007/s12178-021-09734-3

THE USE OF TECHNOLOGY IN ORTHOPAEDIC SURGERY—INTRAOPERATIVE AND POST OPERATIVE MANAGEMENT (C KRUEGER AND S BINI, SECTION EDITORS)

Natural Language Processing and Its Use in Orthopaedic Research

John M. Wyatt[1] · Gregory J. Booth[2,3,4] · Ashton H. Goldman[1,3]

Accepted: 19 October 2021 / Published online: 10 November 2021

Abstract

Purpose of Review This review aims to demonstrate how natural language processing is used in orthopaedic research.

Recent Findings Natural language processing is a form of artificial intelligence that involves encoding human-generated text or speech into a form which can be interpreted by computers to perform a variety of tasks. Natural language processing gathers, processes, and organizes large amounts of free-text data more efficiently than humans. In orthopaedics, it has been utilized for retrospective chart review, automated reporting of electronic health record data, analyzing operative notes and radiology reports, and patient reviews of physicians and practices.

Summary Although still in its infancy, natural language processing promises to be a valuable tool in the future of orthopaedic research. It will not eliminate the need for the essential human component of questioning involved in research, but natural language processing can improve the quality, efficiency, and thoroughness of research, thus improving patient care.

Keywords Natural language processing · Artificial intelligence · Machine learning · Free-text data

Introduction

What Is NLP?

Natural languages are spoken and written by humans. Although they have syntactic and grammatical rules, they are incredibly nuanced and have varying interpretations depending on the context. Sarcasm, puns, tone, dialects, and several other linguistic complexities continue evolving since humans first began communicating with one another.

Natural language processing (NLP) refers to the techniques used by computers to extract meaning from natural languages. It represents elements of several fields, including computer science, artificial intelligence, statistics, and linguistics [1]. NLP techniques vary widely in complexity. Simple applications may do little beyond counting occurrences of words or themes. For example, an NLP algorithm may count the occurrence of certain words within a set of social media posts, which may help characterize trends in what a group is discussing. Another may perform sentiment analysis on shopping reviews by summing each word's positive and negative connotations, indicating how consumers feel about specific products.

More complex NLP algorithms perform additional processing, including machine learning techniques to categorize

Investigation performed at Naval Medical Center Portsmouth, Portsmouth, VA

This article is part of the Topical Collection on *The Use of Technology in Orthopaedic Surgery—Intraoperative and Post Operative Management*

✉ Ashton H. Goldman
Ashton.goldman@gmail.com

John M. Wyatt
Jmwyatt1353@gmail.com

Gregory J. Booth
gjbooth2@gmail.com

[1] Department of Orthopaedic Surgery, 620 John Paul Jones Circle, Portsmouth 23708, VA, USA

[2] Department of Anesthesiology and Pain Medicine, 620 John Paul Jones Circle, Portsmouth, VA 23708, USA

[3] Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA

[4] Naval Biotechnology Group, 620 John Paul Jones Circle, Portsmouth, VA 23708, USA

content, extract a deeper meaning, or even generate speech. Your email may have a filter algorithm that was trained on millions of emails to detect spam. A search engine may infer the intent of your search by considering your IP address, demographics, browsing history, and other users' behavior rather than relying only on the specific words you type. An even more complex application may be a chatbot or assistant device that responds to your queries much as humans do.

A computer program takes human language in each of these examples and breaks it down into basic components, such as individual words or groups of words. It typically then removes common words which convey little meaning, such as "do," "for," and "it," and processes the components to find root meanings (e.g., "am," "is," "was," and "being" all have the same root or lemma, "be"). More complex algorithms then perform additional processing, such as using large lexicon databases to disambiguate words for the specific context in which they appear. For example, in the sentences "The instrument was used to complete the surgery" and "The instrument sounded like a guitar," an NLP algorithm should recognize that the word "instrument" refers to different things. A variety of machine learning techniques, including deep learning, can work towards understanding the meaning or generate text or speech.

NLP and Orthopaedic Research

NLP has garnered increased interest in orthopaedic surgery literature in recent years. A simple PubMed search in June 2021 for title and abstracts using terms such as "orthopedics" OR "fracture" OR "arthroplasty" AND "Natural language processing" produced 41 studies. Approximately 90% of these were in 2019–2021. NLP was used to organize the orthopaedic NLP literature produced by this search and demonstrate how simple yet powerful it can be.

Each orthopaedic title underwent processing. Each title was split into individual words with punctuation removed, and all letters were made lowercase (computers encode upper- and lowercase letters as separate characters). Words were defined as a verb, adverb, adjective, or noun and were then converted to their root form or lemma. Words such as "of," "a," and "in" (termed "stopwords") and the phrase "natural language processing," which add little value to categorizing the content of our search results, were removed. The remaining text was a concise collection of the core concepts represented by study titles.

As an example of these pre-processing techniques, the study title, "Modern Internet Search Analytics and Total Joint Arthroplasty: What Are Patients Asking and Reading Online?" was reduced to ["modern," "internet," "search," "analytics," "total," "joint," "arthroplasty," "patient", "ask," "read," "online"].

Using Latent Dirichlet Allocation (LDA), which categorizes the text, four topics were produced [2]

(1) ["classification," "fracture," "report," "text," "automate," "bone," "health," "record," "electronic," "radiology"]
(2) ["data," "report," "arthroplasty," "infection," "element," "identify," "bone," "note," "common," "operative"]
(3) ["fracture," "radiology," "report," "identification," "hip," "detection," "knee," "learn," "support," "tool']
(4) ["patient," "total," "arthroplasty," "study," "orthopaedic," "joint," "knee," "review," "online," "comment"]

These results highlight that NLP is far from perfect as the topics have substantial overlap. However, four categories emerged: [1] automated reporting using electronic health record (EHR) text, [2] identifying common data elements such as infection using operative reports, [3] identifying fractures from radiology reports patient, and [4] online reviews and comments after total joint arthroplasty.

Automated Reporting of EHR Data

The sheer volume of EHR data in existence places significant limits on how much information humans can extract manually. NLP techniques allow rapid analysis of enormous databases for research, quality, and even real-time clinical management purposes. Thirukumaran et al. noted that surgical site infection (SSI) is a costly problem in hospitals and is mostly tracked through manual record review of charts or administrative and claims data. They concluded that NLP was comparable to manual chart review and significantly improved SSI detection than models using administrative data alone [3]. Karhade et al. also used NLP to detect reoperation rates secondary to SSI within 90 days of lumbar discectomy at two academic and three community centers. Manual chart review of International Classification of Diseases (ICD) and Current Procedural Terminology (CPT) codes was used as the comparison standard. NLP had a higher sensitivity than manual review, a positive predictive value of 0.83, and an F-1 score of 0.88 at a 0.05 threshold [4]. Shah et al. compared NLP to manual chart review in collecting 19 different variables in unstructured, randomly selected arthroplasty notes. Their results showed a greater than 90% accuracy in all 19 variables [5]. NLP was noted to be more accurate in objective data such as range of motion. Accuracy of detection of hip dislocation after total hip arthroplasty by NLP compared to retrospective records review using ICD and CPT codes in unstructured telephone notes, radiology notes, and free-text medical narratives was performed by Borjali et al. All of their NLP models were found to be more accurate than record review using ICD and CPT codes [6•].

Operative Notes

NLP excels at extracting useful information from unstructured text (e.g., text in a paragraph rather than in a specific field or structure). Sagheb et al. tested the accuracy of NLP in extracting five major data elements from 20,000 total knee arthroplasty operative notes. They found greater than 98% accuracy on all their test data sets. An *F*-1 score of 99.9% was achieved on their implant model algorithm in the same study [7]. Karhade et al. used NLP as a source for automatic detection of incidental durotomy using free-text operative notes. In their independent testing set, NLP algorithms far exceeded the performance of standard review using ICD and CPT codes with a sensitivity of 0.89 compared to 0.28 [8]. Fu et al. compared NLP to manual review of operative notes, pathology reports, consultation notes, and microbiology results for detecting prosthetic joint infection (PJI) according to the Musculoskeletal Infection Society (MSIS) criteria. Other algorithms to detect the presence of variables associated with PJI included the growth of cultured organisms, documentation of inflammation, presence of sinus tract, and purulence had *F*-1 scores of 0.771–0.982, and sensitivities and specificities ranging from 0.730 to 1.000 and from 0.947 to 1.000 respectively [9]. Wyles et al. used NLP to extract three data points (approach, fixation method, and bearing surface) in total hip arthroplasty operative notes and then compared their results to outside facility electronic health records for external validity. Results of 99.2%, 95.8%, and 90.7% for their three data points were externally validated, concluding that NLP is an efficient and reliable alternative to manual chart review [10•].

Radiology Reports

Several studies have described NLP applications to radiology reports. Tibbo et al. used NLP algorithms in a two-part study. First, they used it to identify the presence of periprosthetic femur fracture. Second, they challenged the accuracy of their NLP algorithms in correctly determining the corresponding Vancouver classification of each periprosthetic femur fractures. The authors used chart and radiographic review by experienced orthopaedic surgeons as their comparison standard. Their algorithm was 100% sensitive and 99.8% specific in identifying a periprosthetic fracture. It was 78.6% and 94.8% sensitive and specific, respectively, in determining the correct Vancouver classification [11•]. NLP was used to identify radiology findings in free-text radiology reports associated with lumbar back pain by Tan et al. Their algorithm searched for 26 specific findings and these results were compared to medical expert review. At least two experts reviewed every report. Their findings show that the machine-based learning algorithm can identify the 26 specified findings with higher sensitivity and greater AUC with comparable specificity to expert medical review [12]. Wang et al. examined the utility of NLP algorithms to automatically extract the data of six major osteoporotic fracture types from radiology reports. The algorithm achieved a sensitivity and specificity of 0.796 and 0.978, a positive predictive value of 0.972, a negative predictive value of 0.831, and an *F*-1 score of 0.874 [13]. Jungmann et al. used NLP to evaluate the incidence of fracture identification during the recent COVID-19 pandemic. They used a commercially available NLP engine to analyze radiograph reports over a 6-week interval in March and April of 2015–2019 to the pandemic to 2020. The average number of confirmed fractures in non-pandemic years was 295, and in 2020 was only 233. It was concluded that NLP could be used for real-time automated monitoring of selected data points [14].

Patient Reviews of Physicians and Practices

In the growing trend of patient-centered medicine and the increase in the importance of patient satisfaction, NLP has been used to analyze free-text reviews, a previously untapped source of information in research, of physicians and surgical practices. Patawut et al. analyzed over 1000 patient comments from patients who underwent TKA between August 2016 and August 2019 and categorized them into positive and negative themes. They found that negative comments were most related to poor communication and room condition. There was no significant difference in comment theme compared to more traditional outcome measures such as length of stay, pain at 6-week follow-up, peak pain intensity, and the Knee Injury and Osteoarthritis Outcome Score. They concluded that utilizing NLP to analyze patient reviews might not be a good outcome measure, but it can help improve patient satisfaction [15]. Langerhuizen et al. analyzed over 11,000 patient reviews on yelp.com associated with a one-to-five-star rating. Shorter reviews were associated with more positive tones and positive reviews with confidence and joy. Lower ratings were associated with sadness and tentativeness. They concluded that care and compassion are vital elements when it comes to patient satisfaction. Patient-clinician relationships and patient experiences seem to be related to communication and a lack of perceived empathy by the surgeon [16•]. NLP and sentiment analysis were used to analyze free-text comments compared with a national inpatient survey by Greaves et al. Over 6000 free-text comments posted online during the same period as the survey were analyzed. The comments encompassed interactions at 161 hospital trusts in England. Results showed an agreement of 89%, 84%, and 81% of recommendations of a hospital, being treated with dignity, and hospital cleanliness, respectively, between their algorithm and quantitative ratings from the inpatient survey. They concluded that using NLP to analyze free-text patient reviews is an accurate means of assessing patient opinions [17].

Other Medical Applications of NLP

NLP usage in other medical fields demonstrates the potential it has within orthopaedics. Sriram et al. use it to discuss the possibility of gathering data efficiently and tackle the immense problem of enhancing care for diabetic patients. It has also been used to identify significant bleeding events in the charts of critically ill patients accurately and quickly [18]. Keimeyer et al. analyzed the ability of NLP systems to analyze unstructured data in a clinical context. They found most of the current NLP programs were simple rules-based algorithms with a few machine learning algorithms. Their review showed NLP was applicable for retrieving and structuring clinical data; however, given the probable future direction of NLP, more complicated algorithms capable of capturing a broader scope are needed [19]. Sheikhalishahi et al. analyzed the capability of NLP to be used in large-scale electronic health record search to elicit information on chronic diseases [20]. Their goal was to use NLP to further help conduct clinical and longitudinal research to further clarify chronic disease processes. Their review focused mainly on diseases of the circulatory system, which tends to have larger amounts of information in unstructured notes. This could have some value related to chronic orthopaedic diseases such as osteoarthritis, rheumatoid arthritis, and avascular necrosis in identifying risk factors and progression patterns, thus hopefully improving treatment.

Conclusions

NLP shows promise in orthopaedic surgery for research and quality applications, and there is growing evidence that it may be useful for real-time clinical management. It takes unstructured human language and distills it into meaningful data. Several advantages of NLP are its speed resistance to fatigue, scalability, and, in some instances, the ability to supplement machine learning techniques to achieve insights that are difficult or impossible for humans.

While machine learning and NLP have shown promise, the algorithms require significant manpower. Each algorithm is limited by the human-derived elements coded before the algorithm can go to work [21]. Physicians will need to understand and adapt to using NLP. This change requires physician insight into the AI, an understanding that NLP is a computer program programmed by humans, and a realization that machine learning requires validation before its application [22]. While it may one day become the gold standard for data mining, we are still in its infancy. Currently, the never-ending resource of unstructured text and electronic health records may provide the future of evidence-based medicine. While there is currently a trend to more complex algorithms, understanding the limitations and abilities of NLP will be required to effectively utilize this powerful resource.

Disclaimer The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the United States Government.

I am a military service member. This work was prepared as part of my official duties. Title 17 U.S.C. 105 provides that "Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. 101 defines a United States Government work as a work prepared by a military service member or employee of the United States Government as part of that person's official duties.

Declarations

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Conflict of Interest John M. Wyatt, Gregory J. Booth, and Ashton H. Goldman declare that they have no conflict of interest.

References

Papers of particular interest, published recently, have been highlighted as:
- Of importance

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011;18(5):544–51.
2. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
3. Thirukumaran CP, Zaman A, Rubery PT, Calabria C, Li Y, Ricciardi BF, Bakhsh WR, Kautz H. Natural language processing for the identification of surgical site infections in orthopaedics. J Bone Joint Surg Am. 2019;101(24):2167–74.
4. Karhade AV, Bongers MER, Groot OQ, Cha TD, Doorly TP, Fogel HA, Hershman SH, Tobert DG, Schoenfeld AJ, Kang JD, Harris MB, Bono CM, Schwab JH. Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? Spine J. 2020;20(10):1602–9.
5. Shah RF, Bini S, Vail T. Data for registry and quality review can be retrospectively collected using natural language processing from unstructured charts of arthroplasty patients. Bone Joint J. 2020;102-B(7_Supple_B):99–104.
6.• Borjali A, Magnéli M, Shin D, Malchau H, Muratoglu OK, Varadarajan KM. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: a case study of detecting total hip replacement dislocation. Comput Biol Med. 2021;129:104140. Most retrospective reviews, and thus most research in the field of orthopaedic surgery, heavily rely on chart mining with ICD and CPT code searches to supply their data set. Borjali et al specifically compared to NLP algorithms to manual chart review searching via ICD and CPT codes. They found NLP searches of free-text notes were far more accurate in identifying the desired condition than manual search of ICD and CPT codes. This will allow researches to have larger and more thorough data sets/populations in their future research endeavors and therefore strengthening their studies
7. Sagheb E, Ramazanian T, Tafti AP, Fu S, Kremers WK, Berry DJ, Lewallen DG, Sohn S, Maradit Kremers H. Use of natural language processing algorithms to identify common data elements in operative notes for knee arthroplasty. J Arthroplast. 2021;36(3):922–6.

8. Karhade AV, Bongers MER, Groot OQ, Kazarian ER, Cha TD, Fogel HA, Hershman SH, Tobert DG, Schoenfeld AJ, Bono CM, Kang JD, Harris MB, Schwab JH. Natural language processing for automated detection of incidental durotomy. Spine J. 2020;20(5): 695–700.

9. Fu S, Wyles CC, Osmon DR, Carvour ML, Sagheb E, Ramazanian T, Kremers WK, Lewallen DG, Berry DJ, Sohn S, Kremers HM. Automated detection of periprosthetic joint infections and data elements using natural language processing. J Arthroplast. 2021;36(2): 688–92.

10.• Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. J Bone Joint Surg Am. 2019;101(21):1931–8. Wyles et al. demonstrated NLP was accurate in detecting multiple desired variables from operative reports. They also took the next step and externally validated their algorithms by testing on a separate set of operative notes from physicians at outside facilities with similar accurate results. This displays the ability of well-designed NLP algorithms to accurately detect and categorize information from many differing sources allowing efficient gathering of extremely large data sets. The research could then be more generalizable to a larger population group

11.• Tibbo ME, Wyles CC, Fu S, Sohn S, Lewallen DG, Berry DJ, et al. Use of natural language processing tools to identify and classify periprosthetic femur fractures. J Arthroplast. 2019;34(10):2216–9. This reference is important in two aspects. First, Tibbo et al. demonstrated extremely high accuracy in their initial algorithms searching/identifying periprosthetic fractures out of free-text radiology reports. It shows NLP has the ability to extract information from free text accurately, thoroughly, and efficiently. Second, it also demonstrates the ability of NLP to perform more complex data categorization accurately. It took the second step from simply identifying the fracture to classifying it according the Vancouver classification and did so with great accuracy

12. Tan WK, Hassanpour S, Heagerty PJ, Rundell SD, Suri P, Huhdanpaa HT, James K, Carrell DS, Langlotz CP, Organ NL, Meier EN, Sherman KJ, Kallmes DF, Luetmer PH, Griffith B, Nerenz DR, Jarvik JG. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. Acad Radiol. 2018;25(11):1422–32.

13. Wang Y, Meehrabi S, Sohn S, Atkinson A, Amin S, Liu H. Automatic extraction of major osteoporotic fractures from radiology reports using natural language processing. International Conference on Healthcare Informatics Workshop (ICHI-W): IEEE; 2018. p. 64-5.

14. Jungmann F, Kämpgen B, Hahn F, Wagner D, Mildenberger P, Düber C, Kloeckner R. Natural language processing of radiology reports to investigate the effects of the COVID-19 pandemic on the incidence and age distribution of fractures. Skelet Radiol. 2021;

15. Patawut Bovonratwet MTSS, MD; Wasif Islam, BS; Michael P. Ast, MD. Natural language processing of patient-experience comments after primary total knee arthroplasty. J Arthroplast 2020;26(3):927-934.

16.• Langerhuizen D, Brown L, Doornberg J, Ring D, Kerkhoffs G, Janssen S. Analysis of online reviews of orthopaedic surgeons and orthopaedic practices using natural language processing. J Am Acad Orthop Surg. 2021;29(8):337–44. In today's growing trend of patient centered medicine, patient satisfaction is an increasingly important factor in practices. This article demonstrated NLP's ability to synthesize free-text information on websites the general public accesses to find information. Tapping into this resource can allow physicians to tailor and focus on ideals and conditions patients consider important to improve their practice

17. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res. 2013;15(11):e239.

18. Sriram RD, Reddy SSK. Artificial intelligence and digital tools: future of diabetes care. Clin Geriatr Med. 2020;36(3):513–25.

19. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, Forshee R, Walderhaug M, Botsis T. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform. 2017;73:14–29.

20. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform. 2019;7(2):e12239.

21. Gwo-Chin Lee M. More data please! The evolution of orthopaedic research: commentary on an artile by Cody C. Wyles, MD, et al.: "Use of natural language processing algoritms to identify common data elements in operative notes for total hip arthroplasy". J Bone Joint Surg. 2019;101(21):e118.

22. Cote M, Lubowitz J, Brand J, Rossi M. Artificial intelligence, machine learning, and medicine. A little background goes a long way towards understanding. Arthrosc - J Arthrosc Relat Surg. 2021;37(6):1699–702.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.