The American Journal of Pathology, Vol. ■, No. ■, ■ 2022

ELSEVIER

The American Journal of PATHOLOGY

ajp.amjpathol.org

REVIEW

Natural Language Processing in Pathology

Current Trends and Future Insights

Pilar López-Úbeda,* Teodoro Martín-Noguerol,† José Aneiros-Fernández,‡ and Antonio Luna†

From the R+D+I Department and the MRI Unit,† Radiology Department, HT Medica, Jaén; and the Department of Pathology,‡ HT Medica, Granada, Spain*

Accepted for publication July 29, 2022.

Address correspondence to Pilar López-Úbeda, Ph.D., R+D+I Department, HT Medica, Carmelo Torres nº2, 23007 Jaén, Spain. E-mail: p.lopez@htmedica.com.

Natural language processing (NLP) has been shown to play a main role in advancing health care, being key to extracting structured information from electronic health reports. In the last decade, several advances in the field of pathology have been derived from the application of NLP to pathology reports. A comprehensive review of the most used NLP methods for extracting, coding, and organizing information from pathology reports is presented, including how the development of tools is used to improve workflow. In addition, this article discusses, from a practical point of view, the steps necessary to extract data and encode natural language information for its analytical processing, ranging from preprocessing of text to its inclusion in complex algorithms. Finally, we highlight the potential of NLP-based automatic solutions for improving workflow in pathology and their further applications in the near future. *(Am J Pathol 2022, ■: 1–10; https://doi.org/10.1016/j.ajpath.2022.07.012)*

Pathologists must ensure the accuracy, completeness, and usefulness of the information that is transmitted in the patient's electronic health record using pathology reports.[1] These reports are often composed of unstructured information (ie, free-text reports without a specific format). Pathology reports include valuable information about tissue sample or whole lesion location, size, as well as other macroscopic and microscopic features, including shape, morphologic characteristics, or number of cells, among some of the most common hallmarks. Besides, data regarding pathology sample behavior at different stains, including immunohistochemistry markers, are also detailed. Finally, a diagnostic impression or conclusion is included by the pathologist in which the main findings and an attempt to provide an answer to a prior clinical query are summarized. The final pathology diagnosis is usually encoded into standard codes for nomenclature of clinical terms, such as SNOMED-CT or International Statistical Classification of Diseases and Related Health Problems for Oncology 10th Revision (ICD-10). In addition to purely pathologic information, pathology reports may also include patient information provided by referring clinicians, usually related to prior patient's history, data concerning biopsy or surgical procedures (eg, organ, site, and location), as well as prior biochemical studies. This patient's information is commonly followed by clinical diagnostic suspicion. Other demographic data, such as sex, age, patient's name, and external identification number, are usually present in pathology reports.

Electronic health records are transforming the practice of medicine and have important implications for pathologists, including their workflow and communication of results to patients and referring clinicians. However, current health information systems are not equipped to analyze and extract this knowledge because of the time and cost involved in manual processing. This manual processing of information is time-consuming, costly, and error prone, and it imposes inherent limitations on the volume and type of information that can be extracted.[2]

The advent of artificial intelligence (AI) has positively affected industries ranging from education to medicine. Through AI, machines can gather information to learn and make measurements, judgments, and predictions in the

Disclosures: A.L. is occasional lecturer of Philips, Siemens Healthineers, Bracco, and Canon; receives royalties as book editor from Springer-Verlag; and is a member of Siemens Healthineers Digital Oncology Advisory Board.

https://doi.org/10.1016/j.ajpath.2022.07.012

context of previous knowledge. AI is based on machine learning (ML), a set of computational methods used to learn from patterns and relationships in training data to predict.[3] AI is accompanied by concepts, such as accuracy, speed, cost reduction, and improved knowledge beyond what humans can perceive. These potential facets of AI make it an attractive tool to help realize the promises of accuracy, value, and innovation in health care through technical innovations. The emergence of AI in pathology rests primarily on three supports: i) Large volumes of information and results are now instantly available in hospitals and clinics as a result of the digitization of clinical, laboratory, and imaging medical data in laboratory information management systems and specific modules of electronic health records. ii) ML techniques can now extract meaningful information from unstructured data sources, including medical images and clinical reports, at high speed. iii) Innovative ML algorithms help the specialist to predict outcomes and provide new insights.[4]

The field of AI known as natural language processing (NLP) is being applied to medical documents to build applications that can understand and analyze this huge amount of textual information automatically.[5] Language comprehension is a challenging task, and simply looking up keywords or phrases is of limited use. Understanding written language requires knowledge of concepts and their synonyms, grammar, sentence relationships, and time references. Other limitations, such as disambiguation or negation detection, are difficulties that can be encountered in NLP.[6]

Early applications of NLP in pathology include assisting clinical workflows, facilitating quality assurance practices, and improving clinical care. Previous efforts have been made to structure and code pathologic data,[7,8] which are essential for fast and reliable access to diagnostic information, especially for cancer registries.[9] As cancer is one of the leading causes of death worldwide, AI-based research has been focused on detecting precise data on numerical parameters related to cancer, such as information on tumor grade and size, and even tumor behavior.[10,11] This type of information is crucial for obtaining TNM (tumor, node, and metastasis) stages of cancer.[12] Other recently published studies have focused on the coding of pathologic information using widely known terminologies, such as International Statistical Classification of Diseases and Related Health Problems for Oncology (ICD-O) and SNOMED-CT.[8,13,14]

Previous publications have provided systematic reviews on what trends have emerged about the use of NLP in pathology.[1,15,16] However, in this review, we explore the evolution of component processes in NLP, by which natural language can be converted into logical or mathematical components, and which methods are mostly applied in practice. Use cases and applications in NLP, such as automatic classification, information extraction, and summary generation, are also discussed. Finally, future directions of this research are detailed.

From Textual Data to Model Inference

Human language is the form of communication through the information that often contains imprecision, an important problem for mathematics and, by extension, for computer science and AI. AI algorithms do not accept human language as input, so it must be recoded into a logical structure before it can be processed. In this section, we discuss the steps that any NLP system must take to achieve the desired goal. Moreover, Figure 1 summarizes the NLP process [F1] pipeline. The first step in the development process of any NLP system is the collection of data relevant to the task at hand. These data must then be preprocessed and converted into a format that is understandable for the modeling algorithms. Next comes the modeling and evaluation phase, where models are built and compared using one or more evaluation metrics. Once the best model has been chosen from among those evaluated, this model is deployed in production. Finally, the performance of the model must be regularly monitored and, if necessary, updated to maintain its performance.

Preprocessing Pathology Reports

Because most pathology reports are written in a free-text format, the first step in data preparation is preprocessing. In textual NLP tasks, this means that any raw text needs to be carefully processed before the algorithm can process it. Data preprocessing usually consists of several steps that depend on a specific task and the type of text to be handled. The main steps involved in the treatment of pathology reports are summarized in Table 1. [T1]

Feature Representation

Once the text has been processed properly according to the objective to be achieved, it would be necessary to convert the words, sentences, and documents into numerical representations to make them a valid input to the algorithm. This task is normally performed by vectorization. A vector is a list of numerical values that together represent the meaning of a unit of text.

Bag of Words

The Bag of Words model is a method used in language processing to represent reports describing the occurrence of words. In this model, each report looks like a bag containing words discarding any information about word order or structure.[17] In other words, this method is only concerned with whether the known words appear in the document, not where in the document, although it is also possible to give a weight to each word using a statistical measure, such as the term frequency−inverse document frequency.[18] Term frequency−inverse document frequency evaluates how

REV 5.6.0 DTD ■ AJPA3776_proof ■ 26 August 2022 ■ 8:50 pm ■ EO: AJPA-D-22-00252

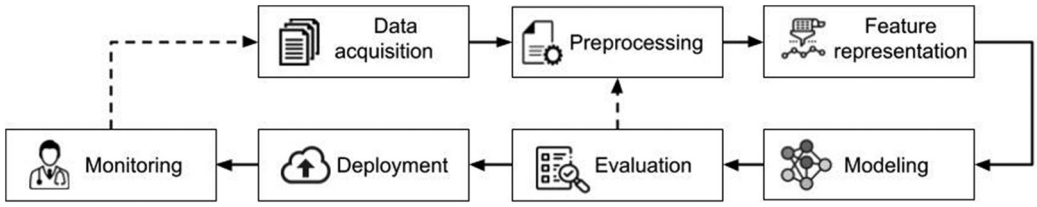

Figure 1 Natural language processing pipeline.

relevant a word is to a document in a collection of documents.

One-Hot Encoding

This technique uses a representation of categorical variables as binary vectors. The main idea is to generate a vocabulary size vector filled with all zeros except one position.[19] Then, for a word, only the corresponding column is filled with the value one, and the rest have a value of zero.

Word Embeddings

This novel technique focuses on mapping the semantic meaning of a word in a geometric space. For this purpose, a numerical vector is associated with each vocabulary word, so that the distance between any two vectors captures part of the semantic relationship between the two associated words. Word embeddings were popularized by Word2Vec in 2013.[20] Pennington et al[21] generated the algorithm GloVe, which aims to perform the meaning embeddings procedure of Word2Vec explicitly. Afterward, fastText was designed to resolve this situation by improving Word2Vec.[22] Finally, contextual word embeddings, such as Embeddings from Language Models (ELMo)[23] and BERT, have emerged.[24]

Machine Learning Models

ML is the study of computer algorithms that can improve automatically through experience and using data.[3] Performing ML involves generating a model, which is trained on some training data and then can process additional data to make predictions. Several types of models for ML systems have been used and investigated in the pathology field. For a description of these models, we have performed a division including traditional algorithms, deep learning (DL), and transformer-based models:

i) Traditional algorithms refer to things we have been doing for years and are often the basis for more advanced ML. Algorithms such as SVM and eXtreme Gradient Boosting (XGBoost) are the most used in pathology and are considered traditional ML methods.

ii) DL has revolutionized many application domains of ML. Deep neural networks are part of a broader family of ML methods based on artificial neural networks. An artificial neural network employs a hierarchy of layers in which each layer considers information from a previous layer and then passes its output to other layers.[25] Although traditional ML algorithms are usually linear, deep learning algorithms are stacked in a hierarchy of increasing complexity and abstraction. The most widely used neural networks in NLP-related research in pathology include recurrent neural networks, such as long short-term memory, multilayer perceptron, and convolutional neural network.

iii) Transformer-based model is a DL model introduced in 2017, used primarily in the field of NLP.[26] Similar to recurrent neural networks, transformers are designed to handle sequential data, such as natural language for tasks like summary generation and text classification. However, unlike recurrent neural networks, transformers process the entire input all at once, and the self-attention mechanism provides context for any position in the input sequence. Self-attention is an attention

Table 1 Text Preprocessing with a Sample Portion of a Pathology Report

Task	Example
Original text	Immunohistochemical factors of prognostic value. Estrogen receptor: tumor cells with nuclear positivity: 91%—100%
Tokenization is a way of separating a piece of text into smaller units called tokens	Immunohistochemical, factors, of, prognostic, value., Estrogen, receptor: tumor, cells, with, nuclear, positivity: 91%—100%
Word stemming is the process of eliminating a part of a word or reducing a word to its root	Immunohistochemical, factor, of, prognostic, value, estrogen, receptor: tumor, cell, with, nuclear, positivity: 91%—100%
Lemmatization is used to reduce words to a normalized form	Immunohistochemical, factor, of, prognostic, value., estrogen, receptor: tumor, cell, with, nuclear, positivity: 91%—100%
Stop word removal involves removing words that commonly appear in all documents in the corpus	Immunohistochemical, factor, prognostic, value., estrogen, receptor: tumor, cell, nuclear, positivity: 91%—100%

REV 5.6.0 DTD ■ AJPA3776_proof ■ 26 August 2022 ■ 8:50 pm ■ EO: AJPA-D-22-00252

mechanism relating different positions of a single sequence to compute a representation of the sequence. In addition, transformers allow much more parallelization than recurrent neural networks and, therefore, reduce training times. Many pretrained models are already available for reuse and serve as a starting point for different tasks in pathology, such as T5, GPT3, GPT-2, BERT, XLNet, and RoBERTa, which demonstrate the ability of transformers to perform a wide variety of such NLP-related tasks and have the potential to find real-world applications.[24,27–30]

Potential Applications of NLP to Pathology

Routine pathology and laboratory workflow produces large amounts of unstructured material that requires more robust NLP to be translated into clinical management and research information. The analysis of pathology reports using NLP has been particularly impacting in recent years, especially in the areas of information extraction, summarization, and categorization. Information extraction pipelines are among the most notable developments using regular expressions, to highlight key findings included in textual reports (eg, extraction of molecular test results) and approaches focusing on topic modeling with the aim of grouping reports under common topics.

Classification

Text classification can automatically analyze text and then assign a set of predefined tags or categories based on its content.[31,32] In the discipline of pathology, classification is the task that allows pathologists to organize the bewildering morphologic manifestations of disease into comprehensible order. Ideally, each of the diagnoses could be grouped according to patients with similar clinical manifestations and identical responses to therapy.[33]

Many earlier attempts aimed at generating rules for classifying pathology reports. The *ICD-O* and the prediction of Current Procedural Terminology (CPT) were the most used terminologies in these studies.[34] Regarding *ICD-O*, Hammami et al[13] implemented a rule-based classification algorithm to classify unstructured text reports into morphologic codes of the *ICD-O* morphology. An NLP-based classifier relying on ad hoc linguistic rules defined on a large data set of 27,239 pathology reports was developed and tested, achieving a micro-F1 score of 98.14%. After manual validation, they observed that the data set contained pathology reports related to 377 different morphologies and was characterized by an unbalanced distribution due to the presence of rare morphologies and the highly detailed ICD-O-M classification. On the other hand, assignments of CPT codes are informed by guidelines and are typically integrated into the laboratory information management systems. Levy et al[35] compared SVM, eXtreme Gradient Boosting, and BERT methods for the prediction of primary CPT codes as well as 38 ancillary CPT codes, using both the diagnostic text alone and text from all subfields of the pathology reports. They demonstrated that BERT outperformed eXtreme Gradient Boosting in predicting primary CPT codes, reaching a macro-F1 score of 82.5%.

Other research has focused on categories such as malignant and benign cancer using DL architectures.[9] Multitask learning techniques have also been applied in classification tasks to identify the site, laterality, behavior, histology, and grade.[10,11,36] In general terms, multitask learning is a subfield of machine learning that aims to solve multiple different tasks at the same time, taking advantage of the similarities between the different tasks by improving learning efficiency. In pathology, multitask learning often attempts to address information extraction from documents by learning to simultaneously extract multiple key features of cancer. In this way, as Figure 2 shows, multitask learning [F2] considers knowledge from multiple partially or fully related tasks to learn shared features.[37] Gao et al[38] proposed a hierarchical self-attention network model for cancer pathology information extraction and text classification. Inspired by the transformer architecture and its attention mechanism, the authors reported macro-F1 scores of 63.36, 49.99, 84.02, 30.23, and 74.3 for extracting useful information, such as site, laterality, behavior, histology, and tumor grade, respectively. Subsequently, the authors improved the proposed method and presented a simple modular plug-in to capture and use the sequential context.[39]

Information Extraction

Locating and extracting information in pathology represents a problem well suited to AI, as algorithms can efficiently, systematically, and comprehensively review pathology reports for a given finding of interest. The task of automatically extracting structured information from unstructured and/or semistructured machine-readable reports and other electronically represented sources is known as information extraction.[40]

Most previous studies on automatic extraction of biomarker data focused on biomedical literature.[41,42] Younesi et al[43] proposed a system for improved recognition of biomarker names in published literature, which is implemented in ProMiner[41] and presents a rule-based system for gene name normalization. Also, the BioNER[42] system uses a custom-made biomarker-specific disease dictionary to extract disease-related biomarkers from MEDLINE publications. However, the performance of these systems is insufficient for processing pathology documents because of intrinsic differences between such clinical documents and published scientific articles.

Inspired by the information extraction task, rule-based approaches, and the use of widely known tools and

REV 5.6.0 DTD ■ AJPA3776_proof ■ 26 August 2022 ■ 8:50 pm ■ EO: AJPA-D-22-00252

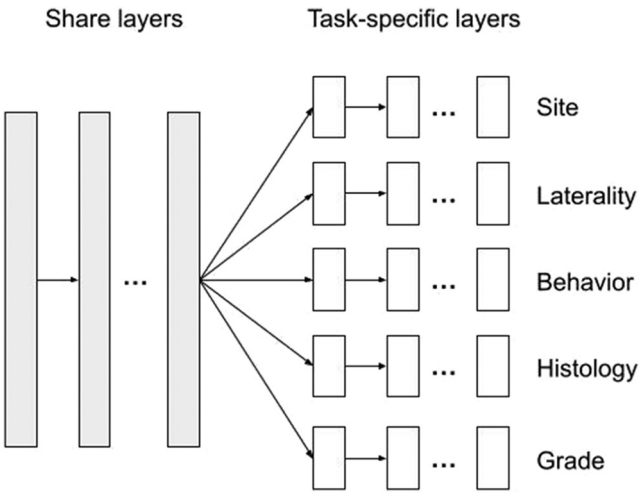

Figure 2 A general multitask learning framework to detect site, laterality, behavior, histology, and grade using shared layers.

frameworks, are employed for extracting information on specific types of cancer using pathology reports.[44–48] On the one hand, Glaser et al[49] aimed to automate the extraction of the stage, grade, and quality information from transurethral resection of bladder tumor pathology reports using a rule-based approach. They identified critical terminology for staging bladder cancer in transurethral resections, including the terms carcinoma *in situ*, lamina propria, muscularis propria, high grade, low grade, and invading. Ryu et al[50] extracted fundamental text entities from pathologic examination reports of patients with colon cancer. The authors used 12,352 pathology reports of surgical specimens to generate a processing model for rule-based text mining that identifies entities such as condition occurrence, measurement value, and specimen name. On the other hand, Napolitano et al[51] used the open-source language engineering framework GATE[52] and aimed at the prediction of chunks of the report text containing information regarding the morphology of cancer, the tumor size, its hormone receptor status, and the number of positive nodes. Other frameworks, such as cTAKES,[53] are well known for their ability to recognize most medical terms (including anatomic sites and disease names) from clinical documents and to correctly normalize these terms to their unified medical language system identifiers.[54]

Because of the special relevance of cancer as one of the main causes of death and the increasing health costs for oncological treatments, a challenge has been generated to identify entities related to oncology named Cantemist.[55] Cantemist was the first shared task that specifically focuses on the task of information extraction in pathology reports of a critical cancer-related concept type by identifying parts of the report, such as pleomorphic high-grade sarcoma, and normalizing them with *ICD-O-3* codes (8802/34: giant cell sarcoma, grade IV).

Summarizing

Text summarization consists of compressing the source text into a reduced version that retains its informative content and overall meaning. Because of the large amount of information contained in pathology reports, text summarization has become an important tool for interpreting the information in the texts. Text summarization methods can be classified into extractive and abstractive summarization. On the one hand, an extractive summarization method involves identifying relevant sentences from the report and putting them together to generate a summary. On the other hand, an abstractive summarization is used to understand the main concepts in each document and then expresses those concepts in clear natural language.[56]

Concerning extractive summarization, Oliveira et al[57] developed a pipeline-based NLP algorithm that incorporated ML and rule-based methods to extract diagnostic elements from narrative pathology reports. For the last step of the algorithm, the structured output was used to summarize each report. The performance was validated on 949 pathology reports. Qiu et al[58] implemented extractive summarization using convolutional neural network and Word2Vec learned embeddings for extracting CD-O-3 Q11 topographic codes from a corpus of breast and lung cancer pathology reports. The purpose of this study was the extraction of information from the export, and therefore, helped to a quicker understanding of a report. However, to the best of our knowledge, the research performed for report simplification and summarization is a simple task consisting of extracting words or sentences from the report and placing them as a summary without considering grammatical nuance, sequence flow, or context. Abstractive summarizations are a challenge nowadays in pathologic diagnosis as it is necessary to use a different vocabulary than the original document. Developing such summaries can be difficult as they would require natural language generation.

Topic Modeling

As previously detailed, keyword extraction involves identifying important words within reports that summarize their content. In contrast, the topic modeling allows grouping these keywords using an intelligent scheme, enabling pathologists to further focus on certain aspects of a report. Q12

This task has also been analyzed by NLP researchers in the field of health care and more specifically in pathology.[59] For instance, Levy et al[35] utilized advanced topic modeling to identify topics that characterize a cohort of 93,039 pathology reports. For this purpose, the authors deployed term frequency–inverse document frequency representation with latent dirichlet allocation algorithm, which identifies topics characterized by a set of words and then derives the distribution of topics over all clusters. In addition, the authors showed the 10 most important words for each topic for the

REV 5.6.0 DTD ■ AJPA3776_proof ■ 26 August 2022 ■ 8:50 pm ■ EO: AJPA-D-22-00252

diagnostic text and all subfields of the report. Furthermore, Kalra et al[60] provided useful biomarker information to readers by underlining words such as presence range tumor necrosis. Finally, Arnold et al[61] demonstrated the application of a topic model to discover relevant clinical concepts and structure a patient's clinical history. Also, 117 reports were used for the topic study by identifying labels, such as tumor histopathology and staining and malignant biopsy, according to the concepts found in the reports.

Machine Translation

Many online translation tools are now available for use through the web, and two of the most popular tools for online translation are Google Translate (*https://translate.google.com*, last accessed July 15, 2022) and Bing Microsoft Translator (*https://www.bing.com/translator*, last accessed July 15, 2022).[62] In the medical field, a clinician facing a language barrier and no professional interpreter might choose to use an automatic translator to help communicate with a patient. Machine translation can be used to clarify patient histories, review a clinical diagnosis, or restate the recommended treatment plan and follow-up to facilitate understanding.[63] Johnsi et al[64] used breast cancer pathology reports, translated them offline using Google Translate, and performed autocorrection of the translation done by Google Translate, using domain-specific resources generated for this purpose. Breast Cancer Pathology Lexicon was developed and is composed of 1124 terms.

In addition, the number of resources available for languages other than English is scarce, so many studies have used machine translation techniques to validate their results. For example, RadLex terminology, an ontology that aims to develop a useful vocabulary for radiologists, is not available in Spanish, so the purpose of the study by Cotik et al[65] study was to use RadLex translated into Spanish (via Google Translate) as the main source of information to detect pathologic findings.

Question Answering

Automatic question answering (QA) has been successfully applied in several domains, such as search engines and chatbots. Biomedical QA, as an emerging QA task, enables innovative applications to effectively perceive, access, and understand complex biomedical knowledge.[66] In general, QA itself is a challenging benchmark NLP task for evaluating the abilities of intelligent systems to understand a question, retrieve and utilize relevant materials, and generate their answer.

Most of these studies in pathology concerning the QA task use images for their research because pathology images play a vital role in the diagnosis and treatment of diseases. Clinical QA, such as PathVQA,[67] helps physicians to analyze many images required for medical decision-making and population screening. The PathVQA data set consists of 32,799 question-answer pairs generated from 1670 pathology images collected from two pathology textbooks, and 3328 pathology images. However, to the best of our knowledge, there is scarce literature related to QA systems in pathology using textual reports. For example, Abu Taha[68] developed a prototype for quality control in Arabic using simple rules. For this, the researcher built an ontology on the domain of pathology to be like a knowledge repository containing a sample of information about some diseases. Similarly, Albarghothi et al[69] covered the pathology domain by selecting a sample of diseases and their details to use it in a QA system. Experiment tests were conducted on 100 questions, including factoid and complicated questions. In the evaluation phase, the model achieves promising results of 81% for accuracy. In addition, Qiu et al[70] proposed a novel model that introduced domain-specific features (eg, clinical named entity information) into a pretrained language model for QA task. The authors employed the BERT model to capture contextual information using a data set consisting of 2714 question-answer pairs in Chinese pathology reports.

Report Generation

Almost all known diseases require laboratory and pathology tests for confirmation of a diagnosis and establishment of an appropriate and rapid treatment. Pathologists generate textual reports, which is often a tedious and time-consuming task. Because of the increasing number of patients, pathologists often must complete the writing of many reports in a limited amount of time. This can lead to several problems: practitioners face great pressures because of the increased workload and the growing complexity of their work, which can lead to a higher frequency of diagnostic errors, and the less experienced physician may have difficulty studying the images and performing the task more slowly.

Some recent studies in the field of AI and NLP have undoubtedly shown a promising approach to help pathologists to write the diagnosis and would also reduce their daily workload. Zhang et al[71] presented MDNet, a unified network to establish a direct multimodal mapping from medical images and diagnostic reports. Their method provided a novel perspective for diagnosing medical images by generating diagnostic reports. For evaluation, they applied MDNet to a data set of bladder cancer images with diagnostic reports. Moreover, Pahwa et al[72] proposed a MedSkip network with a convolutional neural network−transformer−based architecture. The first component of this architecture is the visual extractor, where the preprocessed image is fed into the convolutional layers of the network. The second component includes the use of a memory-driven transformer that generates the report automatically. The authors evaluate their model on two publicly available data sets, one containing

REV 5.6.0 DTD ■ AJPA3776_proof ■ 26 August 2022 ■ 8:50 pm ■ EO: AJPA-D-22-00252

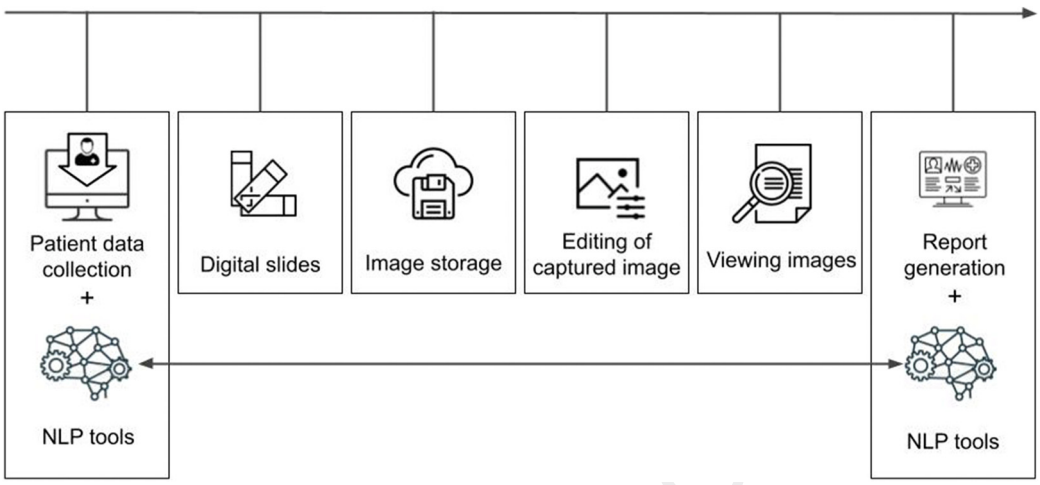

Figure 3 Example of workflow in pathology departments using natural language processing (NLP) tools and techniques.

pathologic images (PEIR gross data set)[73] and one composed of radiological images.[74]

Pathologist's Workflow Optimization with NLP and Future Directions

Current literature has shown that modern NLP-based ML approaches can achieve high accuracy in numerous pathology-related tasks, such as recognizing tumors and their location or grade. However, these studies have been performed almost exclusively on retrospective analysis of archival cases, as is the case of previous stored pathology reports. To further improve the efficiency of pathologic diagnosis of these algorithms, they must be carefully integrated and prospectively validated in real clinical workflows.[75] An example of NLP inclusion in the pathology workflow is presented in Figure 3. NLP methods and tools can be harnessed at different points through the pathologist's workflow. On the one hand, NLP applications can process a patient's clinical information before performing a diagnosis. This is of considerable benefit, as it would allow distributing patients according to certain criteria, sorted or ordered according to urgency (known as worklist prioritization), type of work, and time to be spent. On the other hand, NLP applications can be used to assist the pathologist in generating an accurate diagnostic report. Finally, this whole process aided by NLP methods could be used for data analysis and visualization, to increase interoperability between information systems, to search for reports, and to communicate alerts of findings, among others.

Most NLP applications in pathology are intended to be part of clinical decision support tools for helping pathologists to make clinical decisions, process medical data about patients, and analyze medical knowledge needed to interpret those data.[76] These applications are not intended to replace human pathologists shortly. However, the success of NLP in pathology will require the active participation of pathologists, and the most promising applications will be those that augment human expertise rather than replace it.

Another key issue that poses challenges to the adoption of NLP applications in the clinical setting is their black-box nature and the resulting trust issues. The black-box nature of general AI systems makes pathologists reluctant to trust (due in part to liability risk) something they do not fully understand. One of the reasons why traditional ML methods remain popular is their interpretability compared with DL models.[77] Understanding the features that drive a model's prediction can aid decision-making in the clinical setting, but the complex layers of which DL is composed do not easily allow for transparency. For this reason, advances in the interpretability of DL models are critical for their adoption in clinical practice, which is known as explainable AI. To address these issues, software developers have demonstrated that when solutions are integrated into the clinical decision-making process, they help the clinical team do a better job. An example of this is the LIME framework,[78] which tries to explain what ML models are doing by highlighting important words for certain categories.

On the other hand, DL methods can be computationally expensive and require large amounts of data to learn. The use of pretrained language models based on transformers and word embeddings can reduce some of this burden. Pretrained models often only require fine-tuning, reducing the computational cost. Language understanding has already been prelearned from other tasks, which means that fewer domain-specific labeled examples may be needed.[24] The use of word embeddings has increased recently,[58] which is to be expected with the application of DL approaches, but there are still many algorithms that rely on traditional count-based features (eg, Bag of Words and term frequency–inverse document frequency).[35]

The methods used by NLP in pathology are also affected by data availability. Rare tumors or morphologies are often difficult to predict with ML because of the scarcity of data. Future efforts should be made to make available the data,

linguistic resources developed, and approaches generated to increase the value of NLP in pathology. This would help to advance the field, allowing for more comparisons between studies and increasing the reproducibility of studies.

Finally, ML algorithms have also been used to develop innovative diagnostic tools using images extracted from tissue samples because of the ability to rapidly screen pathologists' slides. In the dawn of digital pathology, where information generated from digitized specimen slides is managed and interpreted, a future direction in the field of pathology could leverage the knowledge extracted by combining digital pathology images with textual reporting. This could extend the value of digital pathology far beyond what is possible today and has been quantified previously.[79]

Conclusion

NLP technologies have been driving pathology since computers were introduced into the workflow. Early algorithms were based on classic methods; however, the emergence of language models and complex neural networks have demonstrated the possibility of understanding language more adequately. NLP technologies, driven by ML methods, now dominate the field of digital pathology because of the performance and flexibility they offer.

As recent literature in pathology shows, NLP-based applications can be applied during report development to help gather and integrate the information needed for case review, but ultimately the pathologist will review the case and make a diagnosis.

Author Contributions

P.L.-Ú., T.M.-N., J.A.-F., and A.L. substantially contributed to conception and design, acquisition of data, or analysis and interpretation of data; P.L.-Ú., T.M.-N., J.A.-F., and A.L. drafted the manuscript or revised it critically for important intellectual content; P.L.-Ú., T.M.-N., J.A.-F., and A.L. approved of the final version to be published; and P.L.-Ú., T.M.-N., J.A.-F., and A.L. agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Burger G, Abu-Hanna A, de Keizer N, Cornet R: Natural language processing in pathology: a scoping review. J Clin Pathol 2016, 69: 949–955
2. Juhn Y, Liu H: Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020, 145:463–469
3. Hutter F, Kotthoff L, Vanschoren J: Automated machine learning: methods, systems, challenges. Springer Nat 2019
4. Hamet P, Tremblay J: Artificial intelligence in medicine. Metabolism 2017, 69:S36–S40
5. Friedman C, Johnson SB: Natural language and text processing in biomedicine. Biomed Inform 2006:312–343
6. Khurana D, Koli A, Khatter K, Singh S: Natural language processing: state of the art, current trends and challenges. Multimed Tools App 2022:1–32
7. Pratt AW, Thomas LB: An information processing system for pathology data. Pathol Annu 1966, 1:1–21
8. Dunham GS, Pacak MG, Pratt AW: Automatic indexing of pathology data. J Am Soc Inform Sci 1978, 29:81–90
9. Khosravi P, Lysandrou M, Eljalby M, Li Q, Kazemi E, Zisimopoulos P, Sigaras A, Brendel M, Barnes J, Ricketts C, Meleshko D, Yat A, McClure TD, Robinson BD, Sboner A, Elemento O, Chughtai B, Hajirasouliha I: A deep learning approach to diagnostic classification of prostate cancer using pathology–radiology fusion. J Magn Reson Imaging 2021, 54:462–471
10. Yoon H-J, Gounley J, Young MT, Tourassi G: Information extraction from cancer pathology reports with graph convolution networks for natural language texts 2019 IEEE International Conference on Big Data (Big Data); 2019. pp. 4561–4564
11. Alawad M, Gao S, Qiu JX, Yoon HJ, Blair Christian J, Penberthy L, Mumphrey B, Wu X-C, Coyle L, Tourassi G: Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc 2020, 27:89–98
12. Nguyen A, Moore D, McCowan I, Courage M-J: Multi-class classification of cancer stages from free-text histology reports using support vector machines 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2007. pp. 5140–5143
13. Hammami L, Paglialonga A, Pruneri G, Torresani M, Sant M, Bono C, Caiani EG, Baili P: Automated classification of cancer morphology from Italian pathology reports using natural language processing techniques: a rule-based approach. J Biomed Inform 2021, 116:103712
14. Schadow G, McDonald CJ: Extracting structured information from free text pathology reports AMIA Annual Symposium Proceedings; 2003. pp. 584
15. Yao K, Singh A, Sridhar K, Blau JL, Ohgami RS: Artificial intelligence in pathology: a simple and practical guide. Adv Anat Pathol 2020, 27:385–393
16. Santos T, Tariq A, Gichoya JW, Trivedi H, Banerjee I: Automatic classification of cancer pathology reports: a systematic review. J Pathol Inform 2022, 13:100003
17. Powell RT, Olar A, Narang S, Rao G, Sulman E, Fuller GN, Rao A: Identification of histological correlates of overall survival in lower grade gliomas using a bag-of-words paradigm: a preliminary analysis based on hematoxylin & eosin stained slides from the lower grade glioma cohort of the cancer genome atlas. J Pathol Inform 2017, 8:9
18. Ramos J: Using tf-idf to determine word relevance in document queries. Proc First Instruct ConfMachine Learning 2003:29–48
19. Karthiga R, Usha G, Raju N, Narasimhan K: Transfer learning based breast cancer classification using one-hot encoding technique 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS); 2021. pp. 115–120
20. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J: Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013:3111–3119
21. Pennington J, Socher R, Manning CD: Glove: global vectors for word representation Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. pp. 1532–1543
22. Bojanowski P, Grave E, Joulin A, Mikolov T: Enriching word vectors with subword information. Trans Assoc Comput Linguist, MIT Press 2017, 5:135–146
23. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L: Deep contextualized word representations. arXiv 2018, [Preprint]

24. Devlin J, Chang M-W, Lee K, Toutanova K: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; 2019
25. Goodfellow I, Bengio Y, Courville A, Bengio Y: Deep Learning. Cambridge, MA, MIT Press, 2016
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: Attention is all you need. Adv Neural Inf Process Syst 2017:5998−6008
27. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020, 36:1234−1240
28. Beltagy I, Lo K, Cohan A: SciBERT: a pretrained language model for scientific text. arXiv 2019, [Preprint]
29. Lample G, Conneau A: Cross-lingual Language Model Pretraining; 2019
30. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, Grave E, Ott M, Zettlemoyer L, Stoyanov V: Unsupervised cross-lingual representation learning at scale. arXiv 2019, [Preprint]
31. Kowsari K, Jafari Meimandi K, Heidarysafa M, Mendu S, Barnes L, Brown D: Text classification algorithms: a survey. Inf MDPI 2019, 10:150
32. Khan A, Baharudin B, Lee LH, Khan K: A review of machine learning algorithms for text-documents classification. J Adv Information Technol 2010, 1:4−20
33. Foucar E: Classification in anatomic pathology. Pathol Patterns Rev 2001, 116:S5−S20
34. Saib W, Chiwewe T, Singh E: Hierarchical deep learning classification of unstructured pathology reports to automate ICD-O morphology grading. arXiv 2020, [Preprint]
35. Levy J, Vattikonda N, Haudenschild C, Christensen B, Vaickus L, others: Comparison of machine-learning algorithms for the prediction of current procedural terminology (CPT) codes from pathology reports. J Pathol Informatics 2022, 13:3
36. Alawad M, Yoon H-J, Tourassi GD: Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports; 2018. pp. 218−221. 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)
37. Ruder S: An overview of multi-task learning in deep neural networks. arXiv 2017, [Preprint]
38. Gao S, Qiu JX, Alawad M, Hinkle JD, Schaefferkoetter N, Yoon H-J, Christian B, Fearn PA, Penberthy L, Wu X-C, Coyle L, Tourassi G, Ramanathan A: Classifying cancer pathology reports with hierarchical self-attention networks. Artif Intell Med 2019, 101:101726
39. Gao S, Alawad M, Schaefferkoetter N, Penberthy L, Wu X-C, Durbin EB, Coyle L, Ramanathan A, Tourassi G: Using case-level context to classify cancer pathology reports. PLoS One 2020, 15:e0232840
40. Zaccaria GM, Colella V, Colucci S, Clemente F, Pavone F, Vegliante MC, Esposito F, Opinto G, Scattone A, Loseto G, Minoia C, Rossini B, Quinto AM, Angiulli V, Grieco LA, Fama A, Ferrero S, Moia R, Di Rocco A, Quaglia FM, Tabanelli V, Guarini A, Ciavarella S: Electronic case report forms generation from pathology reports by ARGO, automatic record generator for onco-hematology. Sci Rep Nat 2021, 11:1−11
41. Hanisch D, Fundel K, Mevissen H-T, Zimmer R, Fluck J: ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics 2005;6:1−9
42. Bravo A, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI: A knowledge-driven approach to extract disease-related biomarkers from the literature. Biomed Res Int 2014, 2014:253128
43. Younesi E, Toldo L, Müller B, Friedrich CM, Novac N, Scheer A, Hofmann-Apitius M, Fluck J: Mining biomarker information in biomedical literature. BMC Med Inform Decis Mak, Biomed Cent 2012, 12:1−13
44. Xu H, Anderson K, Grann VR, Friedman C: Facilitating cancer research using natural language processing of pathology reports. Stud Health Technol Inform, 2004, 107(Pt 1):565−572
45. Nguyen AN, Lawley MJ, Hansen DP, Bowman Rv, Clarke BE, Duhig EE, Colquist S: Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010, 17:440−445
46. Napolitano G, Fox C, Middleton R, Connolly D: Pattern-based information extraction from pathology reports for cancer registration. Cancer Causes Control 2010, 21:1887−1894
47. Martinez D, Li Y: Information extraction from pathology reports in a hospital setting Proceedings of the 20th ACM International Conference on Information and Knowledge Management; 2011. pp. 1877−1882
48. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, de Groen PC: Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. J Biomed Inform 2009, 42:937−949
49. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ: Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform 2018, 2:1−8
50. Ryu B, Yoon E, Kim S, Lee S, Baek H, Yi S, Na HY, Kim J-W, Baek R-M, Hwang H: Transformation of pathology reports into the common data model with oncology module: use case for colon cancer. J Med Internet Res 2020, 22:e18526
51. Napolitano G, Marshall A, Hamilton P, Gavin AT: Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. Artif Intell Med 2016, 70:77−83
52. Cunningham H: GATE: a framework and graphical development environment for robust NLP tools and applications Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002); 2002. pp. 168−175
53. Savova GK, Masanz JJ, Ogren Pv, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010, 17:507−513
54. Bodenreider O: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004, 32:D267−D270
55. Miranda-Escalada A, Farré E, Krallinger M: Named entity recognition, concept normalization and clinical coding: overview of the cantemist track for cancer text mining in Spanish, corpus, guidelines, methods and results; 2020. Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings
56. Tas O, Kiyani F: A survey automatic text summarization. Press-Academia Proced 2007, 5:205−213
57. Oliveira CR, Niccolai P, Ortiz AM, Sheth SS, Shapiro ED, Niccolai LM, Brandt CA: Natural language processing for surveillance of cervical and anal cancer and precancer: algorithm development and split-validation study. JMIR Med Inform 2020, 8:e20826
58. Qiu JX, Yoon H-J, Fearn PA, Tourassi GD: Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform IEEE 2017, 22:244−251
59. Lloyd GR, Almond LM, Stone N, Shepherd N, Sanders S, Hutchings J, Barr H, Kendall C: Utilising non-consensus pathology measurements to improve the diagnosis of oesophageal cancer using a Raman spectroscopic probe. Analyst 2014, 139:381−388
60. Kalra S, Li L, Tizhoosh HR: Automatic classification of pathology reports using TF-IDF features. arXiv 2019, [Preprint]
61. Arnold CW, El-Saden SM, Bui AAT, Taira R: Clinical case-based retrieval using latent topic analysis AMIA Annual Symposium Proceedings; 2010. pp. 26
62. Lin Y-C, Christen V, Groß A, Kirsten T, Cardoso SD, Pruski C, da Silveira M, Rahm E: Evaluating cross-lingual semantic annotation for medical forms. HEALTHINF 2020:145−155
63. Randhawa G, Ferreyra M, Ahmed R, Ezzat O, Pottie K: Using machine translation in clinical practice. Can Fam Physician 2013, 59:382−383

64. Johnsi Rani G, D G, John Mammen J: Machine translation correction using synsets. J Appl Sci Computations 2018, 5:246–252
65. Cotik V, Filippo D, Castano J: An approach for automatic classification of radiology reports in Spanish. Stud Health Technol Inform 2015, 216:634–638
66. Jin Q, Yuan Z, Xiong G, Yu Q, Ying H, Tan C, Chen M, Huang S, Liu X, Yu S: Biomedical question answering: a survey of approaches and challenges. ACM Comput Surv (CSUR) 2022, 55:1–36
67. He X, Zhang Y, Mou L, Xing E, Xie P: Pathvqa: 30000+ questions for medical visual question answering. arXiv 2020, [Preprint]
68. Abu Taha AW: An Ontology-Based Arabic Question Answering System; 2015
69. Albarghothi A, Khater F, Shaalan K: Arabic question answering using ontology. Proced Computer Sci 2017, 117:183–191
70. Qiu J, Zhou Y, Ma Z, Ruan T, Liu J, Sun J: Question answering based clinical text structuring using pre-trained language model 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2019. pp. 1596–1600
71. Zhang Z, Xie Y, Xing F, McGough M, Yang L: MDNet: a semantically and visually interpretable medical image diagnosis network Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017
72. Pahwa E, Mehta D, Kapadia S, Jain D, Luthra A: Medskip: medical report generation using skip connections and integrated attention Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. pp. 3409–3415
73. Jing B, Xie P, Xing E: On the automatic generation of medical imaging reports. arXiv 2017, [Preprint]
74. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ: Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016, 23:304–310
75. Steiner DF, Chen P-HC, Mermel CH: Closing the translation gap: AI applications in digital pathology. Biochim Biophys Acta Rev Cancer 2021, 1875:188452
76. Demner-Fushman D, Chapman WW, McDonald CJ: What can natural language processing do for clinical decision support? J Biomed Inform 2009, 42:760–772
77. Noguerol TM, Paulano-Godino F, Martín-Valdivia MT, Menias CO, Luna A: Strengths, weaknesses, opportunities, and threats analysis of artificial intelligence and machine learning applications in radiology. J Am Coll Radiol 2019, 16:1239–1247
78. Tenney I, Wexler J, Bastings J, Bolukbasi T, Coenen A, Gehrmann S, Jiang E, Pushkarna M, Radebaugh C, Reif E, Yuan A: The language interpretability tool: extensible, interactive visualizations and analysis for NLP models Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations [Internet], Online, Association for Computational Linguistics; 2020. pp. 107–118
79. Parwani AV: Next generation diagnostic pathology: use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. Diagn Pathol 2019, 14:138