Practical Guide to Natural Language Processing for Radiology

Ali Mozayan, MD
Alexander R. Fabbri, BA
Michelle Maneevese, MD
Irena Tocino, MD
Sophie Chheang, MD

Abbreviations: AI = artificial intelligence, BERT = Bidirectional Encoder Representations from Transformers, CBOW = contiguous bag-of-words, NLP = natural language processing, TF-IDF = term frequency–inverse document frequency

RadioGraphics 2021; 41:1446–1453

https://doi.org/10.1148/rg.2021200113

Content Codes: AI IN

From the Department of Radiology and Biomedical Imaging, Yale School of Medicine, PO Box 208042, Tompkins East 2, New Haven, CT 06520 (A.M., M.M., I.T., S.C.); and Department of Computer Science, Yale University, New Haven, Conn (A.R.F.). Presented as an education exhibit at the 2019 RSNA Annual Meeting. Received May 21, 2020; revision requested November 25 and received December 23; accepted January 12, 2021. A.R.F. and S.C. have provided disclosures (see end of article); all other authors, the editor, and the reviewers have disclosed no relevant relationships. **Address correspondence to** A.M. (e-mail: *ali.mozayan@yale.edu*).

Natural language processing (NLP) is the subset of artificial intelligence focused on the computer interpretation of human language. It is an invaluable tool in the analysis, aggregation, and simplification of free text. It has already demonstrated significant potential in the analysis of radiology reports. There are abundant open-source libraries and tools available that facilitate its application to the benefit of radiology. Radiologists who understand its limitations and potential will be better positioned to evaluate NLP models, understand how they can improve clinical workflow, and facilitate research endeavors involving large amounts of human language. The advent of increasingly affordable and powerful computer processing, the large quantities of medical and radiologic data, and advances in machine learning algorithms have contributed to the large potential of NLP. In turn, radiology has significant potential to benefit from the ability of NLP to convert relatively standardized radiology reports to machine-readable data. NLP benefits from standardized reporting, but because of its ability to interpret free text by using context clues, NLP does not necessarily depend on it. An overview and practical approach to NLP is featured, with specific emphasis on its applications to radiology. A brief history of NLP, the strengths and challenges inherent to its use, and freely available resources and tools are covered to guide further exploration and study within the field. Particular attention is devoted to the recent development of the Word2Vec and BERT (Bidirectional Encoder Representations from Transformers) language models, which have exponentially increased the power and utility of NLP for a variety of applications.

Online supplemental material is available for this article.

Introduction

Natural language processing (NLP) is a subfield of artificial intelligence (AI) focusing on the utilization of computers to analyze and manipulate human language. Perhaps one of the earliest historical references to NLP is Alan Turing's publication of "Computing Machinery and Intelligence" in 1950. This article proposed engaging a human and an NLP machine in conversation, in what is now known as the Turing test (1). The earliest efforts in developing NLP systems involved complex explicitly defined rules. This was the standard until the late 1980s, when the advent of new machine learning algorithms gave rise to statistical NLP (2).

NLP has recently undergone a renaissance that was primarily fueled by researchers working at Google. In 2013, the Word2Vec algorithm (Google, https://code.google.com/archive/p/word2vec/) was developed, which employed neural networks to learn word associations from free text without additional input from the user. This

TEACHING POINTS

- Radiology is particularly suited to benefit from applications of NLP, given that the primary mode of interphysician and physician-to-patient communication is by way of the radiology report.
- Word2Vec is a group of neural network models that converts words into a list of numbers called a vector.
- With BERT, the contextualized embedding for a given word changes depending on the sentence the word is found in.
- Word embedding models are evaluated in either an intrinsic or an extrinsic manner.
- As AI becomes integrated into the workflow, it is important for practicing radiologists to understand the basics of its limitations and its underlying processes and vocabulary.

was further developed and refined in 2018, with the advent of the Bidirectional Encoder Representations from Transformers (BERT) language model, which builds on the framework of Word2Vec to learn from not only the text itself but the context in which it is used. This article provides an overview of NLP within the context of AI and radiology, the challenges unique to NLP, and the Word2Vec and BERT NLP techniques. Furthermore, a Jupyter Notebook (Project Jupyter, https://jupyter.org/) using these NLP techniques is offered, enabling amateur enthusiasts to get started with NLP.

NLP and Radiology

Radiology is particularly suited to benefit from applications of NLP, given that the primary mode of interphysician and physician-to-patient communication is by way of the radiology report. The language of radiology uses unique terms and conditions depending on the modality being used or the body part being studied. Each radiologic examination uses a smaller subset of generalized medical vocabulary, allowing a more targeted approach for the utilization of NLP techniques. This feature lends itself well to using pretrained models that have already been created by using a reference dataset and modifying them as necessary to fulfill the desired purpose.

Typically, radiology reports have a standardized format, including subsections for findings and impression, which further facilitates the ability to parse radiology reports into machine-readable data. Radiology reports typically have limited protected health information that is more easily separated from the text compared with other types of medical reports. Freely accessible databases have already been established to assist with the procurement and organization of clinical data. The Medical Information Mart for Intensive Care (MIMIC-III) database (Lab for Computational Physiology, https://mimic.physionet.org/) features abundant information relating to patients receiving intensive care (3). For applications more specific to radiology and NLP, the MIMIC Chest X-ray (MIMIC-CXR) database (Lab for Computational Physiology, https://mimic-cxr.mit.edu/) features a large set of imaging studies and radiology reports with the aim of being used to enhance clinical data mining, NLP, and computer vision, the digital image counterpart of NLP (4).

Potential Applications of NLP to Radiology

Information Extraction

NLP can be used to detect specific diagnoses within the context of a radiology report. NLP algorithms have already been created with the ability to identify specific osteoporotic skeletal fractures from aggregate radiology reports (5). The ability to identify specific studies containing a desired diagnosis can be employed to automatically generate potential teaching cases or compile images featuring the diagnosis for computer vision algorithms to parse further.

Answering Clinical Questions

The use of NLP has the potential to analyze the typically millions of reports at large medical institutions to answer clinical and research questions. As an example, an NLP algorithm has been created with the ability to detect the presence of a thromboembolic disease diagnosis as well as incidental clinically relevant findings, establishing its utility in retrospective study (6).

Summarization and Simplification

Radiology reports typically employ heavy use of jargon that would be difficult to comprehend for patients, who on average have an eighth-grade reading level. NLP could be used to automatically generate a common-language or patient-level report to assist clinicians in explaining imaging findings to patients. This in turn would promote patient understanding and the likelihood of follow-through on clinician recommendations. Semiautomated medical text simplification by using synonym replacement has been successfully demonstrated to improve the readability of health information (7).

Fine-grained Sentiment Analysis

Sentiment analysis refers to the ability to identify subjective opinions from a body of text. The earliest NLP algorithms using sentiment analysis were largely focused on identifying whether consumer reviews of a product were overall positive or negative, now referred to as coarse-grained sentiment analysis. However, a body of text can contain multiple different sentiments regarding the targeted topic. Identifying and ranking each specific sentiment is significantly more complex and is referred to as fine-grained sentiment analysis (8).

The same techniques that were used to identify consumer opinion are being employed to extract physician opinion of a patient's health status (9). Critical results reporting typically depends on the radiologist gauging the severity of imaging findings and taking action to contact the referring clinician. NLP shows promise in assessing the severity and acuity of radiology findings for triaging communication and results reporting accordingly. NLP has been used to create algorithms that can extract stroke acuity from CT and MRI reports (10).

Text Classification and Topic Modeling

Text classification and topic modeling are characterized by identifying and organizing reports by their main concepts, applying clinical labels to each. In text classification, the concepts to be applied as labels are predefined. In topic modeling, the concepts to be applied as labels are determined automatically. The potential exists to cluster thousands of reports into a finite number of groups, without requiring any additional input from the user. The use of topic modeling to label radiology reports in this manner has been attempted but has proved to be nontrivial and less accurate than desired, but it maintains the advantage of separating reports into easily interpretable labels (11). Through text classification and topic modeling, electronic health records can be analyzed to identify cohorts for clinical trials, radiology reports can be organized by diagnosis, and large labeled datasets can be created for deep learning.

Potential Challenges

Radiology confers many advantages in the application of NLP. However, there are challenges both unique to radiology and inherent to NLP that can obfuscate NLP algorithms.

Ambiguity, Syntax, and Named Entity Recognition

NLP algorithms often break sentences down into individual words before further processing. The same word could have different meanings depending on context. For example, "drain" could refer to a catheter or to the process of drainage itself, such as that of a sinus tract. Syntax, or the order and structure of words, is also key for language comprehension, as variances in word order can confer different meanings. "The mass in the right upper lobe has increased in size, although its standardized uptake value (SUV) has decreased" conveys a different meaning than "Although the mass in the right upper lobe has decreased in size, its SUV has increased," despite containing the same exact words. Furthermore, an NLP algorithm must be able to identify named entities or specific noun phrases. Examples of this would be "frontal lobe" and "lateral ventricle," which specifically refer to brain structures.

The challenges posed by ambiguity and syntax are addressed through context analysis. *N*-grams, which refer to a set of a specified number (*n*) of words, incorporate word order and context. *N*-grams can come in the form of bigrams (n = 2) or trigrams (n = 3), for example. By examining how often one word is followed by another, the relationships between words can be better elucidated.

Synonymy

Different words and phrases are often used to convey the same information. For example, "nonechogenic" and "anechoic" similarly describe areas that cannot be reflected at US. "Enlarging" and "growing" are both used to describe lesions that demonstrate interval increase from a prior study. These synonyms are typically differentially used because of radiologist preference and have no real clinical significance. The challenge of synonymy is often addressed by synonym substitution. By reducing the vocabulary used while preserving the original meaning, simplification of NLP as well as the necessary resources to do so is achieved. The Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT, SNOMED International, https://www.snomed.org/) includes a collection of medical terms with their synonyms and can be used for this purpose.

Medical Abbreviations

Medical abbreviations are frequently used but have different meanings within different contexts. "PAP" can refer to pulmonary artery pressure or pulmonary alveolar proteinosis. The use of machine learning to facilitate NLP can employ supervised models, which incorporate prelabeled and cleaned data with programmed outputs, or unsupervised methods for discovering patterns in unlabeled data (12).

Unsupervised approaches have been developed for medical abbreviation disambiguation in discharge summaries, which can be used for other text genres without model retraining (13).

Coreference

Many reports use references to previously mentioned concepts or words that are not repeated in subsequent sentences or paragraphs. As an example, a report may include this wording: "A 2-cm right hepatic lesion demonstrates portal venous enhancement. It is likely related to the patient's underlying untreated hepatitis C infection." In this case, "it" refers to the liver lesion. Coreference resolution refers to identifying all words or phrases in a body of text that refer to the same concept. Algorithms have been developed identifying such uses of "it," with results comparable to human interpretation (14).

Developer Environments and Tool Kits

Currently, most research software for NLP is written in Python (Python Software Foundation, https://www.python.org/), although Java (Pluralsight, https://www.pluralsight.com/paths/java) and C++ (https://www.stroustrup.com/C++.html) are used occasionally. Python is a high-level programming language with a comprehensive standard library that offers relative simplicity and an abundance of freely available open-source environments and tools. Python has a wide abundance of software bundles available, known as packages, including many with data science and NLP applications. A large repository of these packages can be found on the official Python website's Python Package Index (PyPI, Python Software Foundation, https://pypi.org/). The recommended package installer is known as "pip" (PyPA, https://pip.pypa.io/en/stable/), which downloads and installs packages from PyPI.

Virtual Environments

Python virtual environments provide an isolated location to install packages for a specific application. While virtual environments are not strictly necessary to dive into data science and NLP with Python, they keep track of package dependencies and facilitate the process for others to reproduce the same work.

Commonly employed virtual environments include Virtualenv (PyPA, https://virtualenv.pypa.io/en/stable/), Poetry (Poetry, https://python-poetry.org/), and Anaconda (Anaconda, https://www.anaconda.com/distribution). Many packages used for data analysis are easily installed by the package manager Conda (Anaconda, https://docs.conda.io/projects/conda/en/latest/). Conda builds on pip and can install the Python package dependencies offered by pip as well as those outside the main Python library. Anaconda has the most utility for NLP, as it comes preinstalled with many useful Python libraries for data science. If one wishes to save the time and disk space required to install the over 1500 packages that Anaconda offers, Miniconda (Anaconda, https://docs.conda.io/en/latest/miniconda.html) is another option for more advanced users that requires the user to choose each package to install.

NLP Libraries

There are various libraries that employ some of the more common NLP techniques for use within Python. The Natural Language Toolkit (NLTK, NLTK Project, http://www.nltk.org), SpaCy (Explosion, https://spacy.io), and the Stanford CoreNLP (Stanford NLP Group, https://stanfordnlp.github.io/CoreNLP) are among the most comprehensive text processing libraries with the largest coverage in terms of functionalities for processing text. They each offer tools for performing tasks common to NLP and differ in their speed and performance advantages in completing certain tasks.

NLTK covers all components of a typical NLP pipeline. It offers extensive documentation, including an online book (http://www.nltk.org/book/) to guide users through the basics of using Python for NLP.

SpaCy is less complete than NLTK but offers all major components of NLP while focusing on speed and production. State-of-the-art NLP models are incorporated directly into spaCy's tools. Integrating models, including deep learning models, into spaCy's framework is straightforward. As a result, spaCy has become popular for production as well as for research projects. Also, projects like scispaCy (Explosion, https://spacy.io/universe/project/scispacy) build on spaCy to analyze scientific texts. Despite the popularity of spaCy, depending on the specific task, developers may want to use NLTK.

The Stanford CoreNLP is an NLP library written in Java. It can be used from the command line, a web server, or with third-party application programming interfaces (APIs) for other programming languages, including Python.

In addition to these tools, packages such as scikit-learn (or sklearn, https://scikit-learn.org/stable/) and NumPy (NumPy, https://numpy.org/), which are generally used in machine learning and data analysis, are also used in the NLP workflow. For deep learning, common frameworks include PyTorch (https://pytorch.org/), Tensorflow (Google, https://www.tensorflow.org/), and DyNet (DyNet, http://dynet.io/).

Word2Vec

Word2Vec is a group of neural network models that converts words into a list of numbers called a vector. Words that appear in similar contexts have similar vectors. Word2vec is an example of word embedding, the collective name for mapping words and phrases to vectors. Through word embedding, words and phrases are converted to

	Spherical	Soft
Apple	7	2
Banana	2	8
Orange	9	7
Pear	4	3

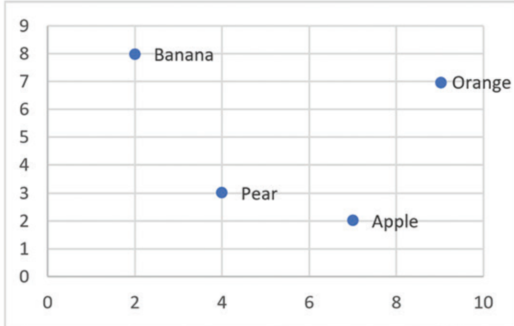

Figure 1. Through word embedding models such as Word2Vec, a word can be converted to an array of numbers on the basis of its characteristics. In this simplified example, the word "apple" is converted to the vector [7, 2] on the basis of its relative sphericity and softness. Words that represent similar concepts have similar vectors.

an array of numbers on the basis of their inherent characteristics (Fig 1). By converting words to numerical vectors in this manner, mathematical operations between them are made possible. Through Word2Vec, words can be added together as vectors would to yield a result. For example, vector word embeddings for "opacity" and "linear" in the setting of a chest radiograph report could be combined in this manner to generate the desired vector representing "atelectasis."

The vector for a single word can have hundreds of dimensions. Word2Vec employs two methods for vectorizing words, which are known as the contiguous bag-of-words (CBOW) model and the skip-gram model.

The simple bag-of-words model creates vectors from text by counting the number of occurrences of unique words. As an example, the phrase "he is good and she is good" would be scored as follows:

"he" = 1
"is" = 2
"good" = 2
"and" = 1
"she" = 1.

The sentence in this example would hence be converted to the vector [1, 2, 2, 1, 1]. The order and structure of words in the original document are lost to the "bag" in this model.

The CBOW model works by looking at the words immediately preceding and following a target word in a body of text, generating smaller bags to predict the target word. The difference between the predicted targeted word and actual targeted word is calculated as an error vector, which in turn is used to modify the model and refine each word embedding vector.

The skip-gram model can be thought of as the inverse of the CBOW model, as it predicts the context of a target word from the word itself (Fig 2). Skip-grams are much more computationally intensive than the CBOW model but have more

Contiguous Bag-of-Words

the quick brown fox jumps over the lazy dog

Skip-Grams

the quick brown fox jumps over the lazy dog

Figure 2. The contiguous bag-of-words (CBOW) model utilizes the words immediately preceding and following a target word to predict it. Conversely, the skip-gram model predicts the context of a target word from the word itself.

utility in analyzing words that appear infrequently in a body of text. The open-source library Gensim (https://radimrehurek.com/gensim/) offers an implementation of Word2Vec utilizing the skip-gram model.

Word2Vec vectors are often incorporated into NLP models as a way of creating machine-interpretable word representations. One limitation of Word2Vec is that it does not assign different vectors for different meanings of the same word. Over the last few years, work has moved from pretraining word-fixed representations like Word-2Vec to contextual representations, so that the representation of each word is dependent on the context in which it appears.

Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT, Google, https://arxiv.org/abs/1810.04805) is an open-source unsupervised framework published by Google AI in 2018 for language representation (15). It offers the ability to generate different vector representations for a word dependent on its context in a sentence. A single word in a body of text will ultimately have the same Word2Vec vector irrespective of context.

Figure 3. Bidirectional Encoder Representations from Transformers (BERT) masks several words out of a series of tokens and trains itself to predict the masked tokens by utilizing context clues. Unlike with Word2Vec, this is accomplished in both left-to-right and right-to-left contexts, and different word embeddings are generated for a word depending on the context.

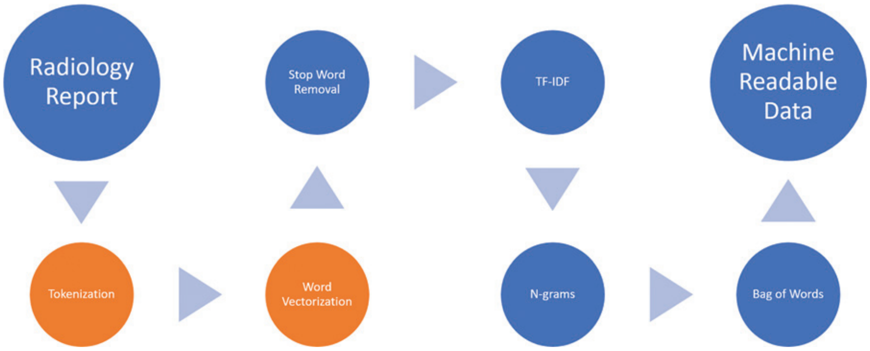

Figure 4. A sample natural language processing (NLP) pipeline converts a radiology report to machine-readable data. Tokenization and word vectorization are highlighted in orange, as they are nearly ubiquitous to all NLP tasks. *TF-IDF* = term frequency–inverse document frequency.

With BERT, the contextualized embedding for a given word changes depending on the sentence the word is found in.

As an example, the vector for "playing" has the same vector representation for Word2Vec if it is used in the sentence "She is playing with her friends" or "She is playing guitar." BERT, on the other hand, generates embeddings that are unique to each sentence.

BERT is trained by taking a body of text and randomly selecting about 15% of the words to replace with a "[mask]" token. It then attempts to predict the masked tokens (Fig 3).

The "bidirectionality" of BERT derives from the fact that it incorporates context for a given word in both a left-to-right and right-to-left manner. "Transformers" refers to the neural network architecture that BERT uses to learn contextual relationships between words, which enables BERT's ability to analyze immense general language datasets (16). Transformers have become the default neural network architecture in NLP after showing state-of-the-art performance in machine translation.

BERT and Word2Vec also differ in their application to downstream tasks. Word2Vec can be viewed as a shallow neural network consisting of a single layer. An extracted vector can be used as input to a larger downstream model. On the other hand, BERT is a much deeper network, consisting of up to 24 transformer network layers. The larger network allows a deeper understanding of language representations during training, as each layer may itself learn a specific aspect of language, such as syntax or word order (17). The BERT model gives a head start for downstream tasks, as it does not need to learn aspects of language such as syntax from the beginning. Rather than training another model on top of BERT's representation, the BERT model can be applied to new tasks such as sentiment analysis by merely adding another layer to the BERT architecture.

NLP Pipeline

Converting free text to a structured numerical output through NLP is typically a stepwise approach (Fig 4). With the advent of neural network models, frequently the only tools employed for NLP are tokenization and word vectorization. Without neural networks, additional methods are employed, including removal of stop words, the calculation of word frequency in a document and across documents (term frequency–inverse document frequency [TF-IDF]), *n*-grams, and bag-of-words.

Tokenization

The task of tokenization involves separating free text, typically by splitting words by white space into tokens (18) (Fig 5). Punctuation characters can also be removed to separate tokens. However, tokenizing text by punctuation requires special consideration because splitting words by hyphens or apostrophes can obscure the meaning of words ("eye-opening," as an example).

Removing Stop Words

A stop word is a word that is typically filtered out through an NLP pipeline. Many NLP methods use frequency to determine the significance of a word or token. Stop words are typically among the most common words in language but are

Findings: Mediastinal lymphadenopathy is stable when compared to prior CT chest dated 10/22/2009. For example, a subcarinal node, measuring 2.2 cm at image 26 previously measured 2 cm and a peritracheal lymph node measuring 1 cm on image 12 previously measured 1 cm….

→

Findings : Mediastinal lymphadenopathy is stable when compared to prior CT chest dated 10/22/2009 . For example , a subcarinal node , measuring 2.2 cm at image 26 previously measured 2 cm and a peritracheal lymph node measuring 1 cm on image 12 previously measured 1 cm ….

Figure 5. In this example, the difference is subtle. "Findings:" is tokenized into "Findings" and ":". In the context of a radiology report, "Findings" on its own has a meaning separate from "findings:".

Findings : Mediastinal lymphadenopathy is stable when compared to prior CT chest dated 10/22/2009 . For example , **a** subcarinal node , measuring 2.2 cm at image 26 previously measured 2 cm **and a** peritracheal lymph node measuring 1 cm on image 12 previously measured 1 cm …

→

Findings : Mediastinal lymphadenopathy stable compared prior CT chest dated 10/22/2009 . example , subcarinal node , measuring 2.2 cm image 26 previously measured 2 cm peritracheal lymph node measuring 1 cm image 12 previously measured 1 cm . ..

Figure 6. Words like "and" and "a" are stop words without enough significance to merit being analyzed.

among the least significant (19). Each NLP library typically has a predefined set of stop words. An example of stop word removal is demonstrated in Figure 6.

Term Frequency–Inverse Document Frequency

TF-IDF is a measure of how often a term is found within a specific document, with the typical exclusion of stop words, and how frequently a word is used across documents. The most characteristic words within a document typically have the highest TF-IDF score, with the exclusion of stop words (19).

Words analyzed and simplified in this manner are characteristic of the bag-of-words model. In bag-of-words, the order of the words does not matter. Instead, only the frequency (or TF-IDF) of the individual words matters. Each report then has a number frequency assigned to each word (Fig 7). These numerical data are then used to train the model.

Now Your Turn

For those with a modest background or significant inclination for getting hands-on with data science, we offer a Jupyter notebook on the Google Colaboratory platform (Google, https://research.google.com/colaboratory/). This notebook allows the user to upload text to follow along with an NLP pipeline involving tokenization, stop word removal, TF-IDF, bag-of-words, and *n*-grams. Furthermore, it generates vector representations of the uploaded text using both Word2Vec and BERT and predicts words using these algorithms. It can be found at *https://colab.research.google.com/drive/1DbA44SEdqtj2N815lJmU7xbyjmCIiol0*.

Figure 7. Each word is given a weight on the basis of its frequency within the document. In our example, words that appear in every report are not considered as important as words that appear infrequently.

	tf_idf_weights
findings	1.000000
right	1.048790
impression	1.048790
unremarkable	1.100083
mediastinal	1.154151
left	1.154151
axillary	1.154151
hilar	1.211309
lobe	1.271934
pericardial	1.271934

To get started using the Jupyter notebook within Google Colaboratory, select the Runtime menu option and then Run All. Under the "Load data" section of the notebook, upload a file with name "report.txt" containing the free text you wish to analyze. Appendix E1 offers sample radiology report text that can be used for analysis.

Evaluating Word Embedding Models

Word embedding models are evaluated in either an intrinsic or an extrinsic manner (20). In intrinsic evaluation, the word embedding quality is examined by manipulating the representations themselves without a particular end task in mind. In extrinsic evaluation, the word embeddings are input to downstream NLP tasks to compare the resulting performance according to the downstream task's metric, such as classification accuracy.

A simple method to evaluate word vectors is asking humans to assess the similarity between pairs of words, ranked from least similar to most similar, and compare this similarity with the cosine similarity, the Word2Vec numeric measure of similarity between vectors. Previous work has collected datasets with human judgments for such intrinsic evaluation (21). An example of extrinsic evaluation is comparing the performance of using the word vectors as input to tasks, such as sentiment analysis, and the model takes these representations as input and optimizes for the downstream task.

Word2Vec may be evaluated in either an intrinsic or extrinsic manner. In contrast, BERT is typically evaluated using extrinsic techniques. BERT's significant contribution is creating contextual word representations that differ for a word depending on the sentence in which it appears, which lends itself to more contextual sentence-based downstream tasks such as sentiment analysis.

Conclusion

NLP is a growing field within radiology, employing state-of-the art machine learning techniques with many potential use cases. As AI becomes integrated into the workflow, it is important for practicing radiologists to understand the basics of its limitations and its underlying processes and vocabulary. Radiologists who have a working knowledge of NLP are empowered to better assess novel technology products and facilitate improved communication and collaboration with computer and data science colleagues.

***Disclosures of Conflicts of Interest.*—A.R.F.** *Activities related to the present article:* disclosed no relevant relationships. *Activities not related to the present article:* employed by Sol Wellness. *Other activities:* disclosed no relevant relationships. **S.C.** *Activities related to the present article:* disclosed no relevant relationships. *Activities not related to the present article:* medical advisor to Agamon; holds stock in Agamon. *Other activities:* disclosed no relevant relationships.

References

1. Turing AM. Computing Machinery and Intelligence. In: Epstein R, Roberts G, Beber G, eds. Parsing the Turing Test. Dordrecht, the Netherlands: Springer, 2009; 23–65.
2. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18(5):544–551.
3. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3(1):160035.
4. Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019;6(1):317.
5. Wang Y, Mehrabi S, Sohn S, Atkinson EJ, Amin S, Liu H. Natural language processing of radiology reports for identification of skeletal site-specific fractures. BMC Med Inform Decis Mak 2019;19(Suppl 3):73.
6. Pham AD, Névéol A, Lavergne T, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 2014;15(1):266.
7. Leroy G, Endicott JE, Mouradi O, Kauchak D, Just ML. Improving perceived and actual text difficulty for health information consumers using semi-automated methods. AMIA Annu Symp Proc 2012;2012:522–531.
8. Fink CR, Chou DS, Kopecky JJ, Llorens AJ. Coarse- and Fine-Grained Sentiment Analysis of Social Media Text. Johns Hopkins APL Tech Dig 2011;30(1):22–30. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.2919&rep=rep1&type=pdf.
9. Deng Y, Stoehr M, Denecke K. Retrieving Attitudes: Sentiment Analysis from Clinical Narratives. Paper presented at MedIR@ SIGIR2014.
10. Ong CJ, Orfanoudaki A, Zhang R, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS One 2020;15(6):e0234908.
11. Zech J, Pain M, Titano J, et al. Natural Language-based Machine Learning Models for the Annotation of Clinical Radiology Reports. Radiology 2018;287(2):570–580.
12. Hinton GE, Sejnowski TJ, Poggio TA. Unsupervised learning: foundations of neural computation. Cambridge, Mass: MIT Press, 1999.
13. Kreuzthaler M, Oleynik M, Avian A, Schulz S. Unsupervised abbreviation detection in clinical narratives. Paper presented at Proceedings of the clinical natural language processing workshop (ClinicalNLP), 2016.
14. Li Y, Musilek P, Reformat M, Wyard-Scott L. Identification of pleonastic it using the web. J Artif Intell Res 2009;34:339–389.
15. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805. Published 2018. Accessed March 28, 2020.
16. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Paper presented at Advances in Neural Information Processing Systems, 2017.
17. Tenney I, Das D, Pavlick E. BERT Rediscovers the Classical NLP Pipeline. Paper presented at Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019.
18. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge, England: Cambridge University Press, 2008.
19. Rajaraman A, Ullman JD. Mining of massive datasets. Cambridge, England: Cambridge University Press, 2011.
20. Wang B, Wang A, Chen F, Wang Y, Kuo CCJ. Evaluating word embedding models: Methods and experimental results. APSIPA Trans Signal Inf Process 2019;8:e19.
21. Luong MT, Socher R, Manning CD. Better word representations with recursive neural networks for morphology. Paper presented at Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 2013.