Article

Not peer-reviewed version

Natural Language Processing Accurately Categorizes Indications, Findings, and Pathology Reports from Multicenter Colonoscopy

Shashank reddy Vadyala *

Posted Date: 4 January 2023

doi: 10.20944/preprints202301.0061.v1

Keywords: Neural Network; Machine Learning; Natural Language Processing (NLP); Text Mining; Sentence Classification; Colorectal Cancer; Clinical Information.

Article

Natural Language Processing Accurately Categorizes Indications, Findings, and Pathology Reports from Multicenter Colonoscopy

Shashank Reddy Vadyala

Department of Computational Analysis and Modeling, Louisiana Tech University, Ruston, LA United States; srv009@latech.edu

Abstract: Colonoscopy is used for colorectal cancer (CRC) screening. Extracting details of the colonoscopy findings from free text in electronic health records (EHRs) can be used to determine patient risk for CRC and colorectal screening strategies. In this study, we developed and evaluated the accuracy of a deep learning model framework to extract information for the clinical decision support system to analyze relevant free-text colonoscopy reports, including indications, pathology, and findings notes. The Bio-Bi-LSTM-CRF framework was developed using Bidirectional Long Short-term Memory (Bi-LSTM) and Conditional Random Fields (CRF) to extract several clinical features from these free-text reports, including indications for the colonoscopy, findings during the colonoscopy, and pathology of the resected material. We then trained the Bio-Bi-LSTM-CRF and existing Bi-LSTM-CRF models on 80% of 4,000 manually annotated notes obtained from the colonoscopy reports of 3,867 patients. The clinical notes were from a group of patients aged over 40 years old enrolled in four Veterans Affairs Medical Centers. A total of 10% of the remaining annotated notes were used to train hyperparameter, while the remaining 10% were used to evaluate the accuracy of our model (Bio-Bi-LSTM-CRF) and to compare the results with the outcomes obtained using Bi-LSTM-CRF. The results of our experiments showed that the bidirectional encoder representations by integrating dictionary function vector from Bio-Bi-LSTM-CRF and strategies character sequence embedding approach is an effective way to identify colonoscopy features from EHR-extracted clinical notes. Therefore, the Bio-Bi-LSTM-CRF model is concluded to be capable of creating new opportunities to identify patients at risk for colon cancer and to study their health outcomes.

Keywords: neural network; machine learning; natural language processing (NLP); text mining; sentence classification; colorectal cancer; clinical information

Introduction

In the last several decades, the amount of complex and varied information (e.g., structured, semi-structured, and unstructured) has dramatically grown in various sectors, including education [1], media [2], energy [3], and healthcare [4]. In healthcare, electronic health records (EHRs) contain information in both unstructured and structured formats on patient health, disease status, and care received, which can be useful for epidemiologic and clinical research [5]. In EHRs, unstructured data are documented without standard content specifications, often recorded as free text, while structured data are generally entered into discrete data fields with standardized responses or parameters (e.g., age, weight) [6]. Compared to the structured data within the EHR, free-text (unstructured) data often contain more granular and contextual descriptions of clinical events and thus enhance communication between clinical teams [7]. Therefore, extracting information from these free-text document resources has a considerable potential value for a variety of purposes, ranging from clinical decision support and quality care improvement to the secondary use of clinical data for research, public health, and pharmacovigilance activities [8]. Natural language processing (NLP) can be an

 doi:10.20944/preprints202301.0061.v1

efficient way of automatically extracting and arranging this information, making NLP a potentially pivotal technology for enabling quality measurement from EHR data[9–14].

Background

Colorectal cancer (CRC), which accounts for almost 10% of all cancer-related deaths in the United States [15], is the third most common cancer type in both sexes. Colonoscopy is a well-established procedure used to detect and prevent CRC. In recent years, owing to the growing efficacy of colonoscopy screening rates, the use of colonoscopy has significantly increased [16]. Understanding the dynamics of changes in colorectal polyps and cancer involving sizes, shapes, and the number of polyps is critical because they are associated with the efficacy of colorectal screening strategies. However, manual extraction of detailed findings of a colonoscopy (e.g., number of polyps, polyp sizes, and polyp locations) from a colonoscopy procedure report and connecting with the pathology report (e.g., the histology of polyps found during colonoscopy) is quite time-consuming [17]. Extracting information entity identification from the clinical free text requires scalable techniques such as NLP, machine learning, and neural networks to convert clinical free text into a structured variable for text mining and further information extraction [18–20]. Colonoscopy named entity identification has drawn a considerable amount of research in recent years, and various techniques for identifying entities have been proposed [21]. However, a limitation of conventional methods—such as Decision Tree, Support Vector Machines (SVM), Hidden Markov Model (HMM), and Conditional Random Field (CRF)—is that these methods are either rule-based or dictionary-based, which entails manual formulation and applicability to exclusively current datasets. When the dataset is updated or replaced, updating the rules contributes to high machine costs. This causes a low recall if the original rules or dictionary are used. Accordingly, the latest advanced techniques are data-driven, including the methods of machine learning and deep learning. In particular, the models of Bi-LSTM-CRF (Bi-Long Short-Term Memory and Conditional Random Field) have been successfully applied to biomedical texts, demonstrating better results [22]. Existing work regarding NLP in connection with CRC and colonoscopy is limited in scope and quantity. For instance, D'Avolio et al. adapted a clinical information retrieval system to identify pathology reports consistent with CRC from an electronic health record [23]. An application was proposed to automatically identify patients in need of CRC screening by detecting the timing and status of colorectal screening tests mentioned in the electronic clinical narrative documentation [2,17,23,24]. Yet, due to the problem of entity boundary extraction with different entity lengths and named entity recognition, none of the previous studies assessed detailed findings of the colonoscopy procedure, such as the number, shape, and size of polyps. Such an understanding of the dynamics of changes in colorectal polyps and cancer is critical because this information is associated with the efficacy of colorectal screening strategies [25].

Named entity recognition (NER), also known as entity identification, entity chunking, and entity extraction, is a process of information extraction that helps to locate and classify named entities mentioned in unstructured and structured text into pre-defined categories such as medical codes, person names, and so forth. Applications of NER in the medical domain are complicated by the following two key reasons [26]. First, because of non-standard abbreviations, acronyms, terminology, and several variations of the same entity, certain entities may fail to appear in the training dataset. Second, free text clinical notes are noisy due to the frequent use of short or incomplete sentences by medical personnel, as well as the presence of grammatical and typographical errors. The third challenge of clinical NER for multiple facilities is potential inter-facility variation compared with single-facility data. Fourthly and finally, clinical reports involve a paragraph border and incomplete sentences, which makes it impossible to identify complex characters and complicates sentence boundary detection [27].

A dependency of NLP is the sentence boundary detection (SBD) task, necessary to identify the sentence segments within a text [28]. Good segmentation of sentences will not only improve translation quality, but also reduce latency in processing of colonoscopy reports. Most of the existing systems—such as SBD systems SATZ—An Adaptive Sentence Segmentation System [29], proposed

 doi:10.20944/preprints202301.0061.v1

by Palmer and Hearst (1997), Disambiguation of Punctuation Marks (DPM), and Automated Speech Recognition (ASR) system—are highly accurate on standard and high-quality text. However, extracting sentences or detecting the boundary of sentences from clinical text (e.g. colonoscopy reports) is challenging, because such texts are characterized by an abundance of acronyms and abbreviations, and often contain misspellings, punctuation errors, and incomplete sentences [30].

Neural embedding methods such as word2vec [31] and GloVE [32] have been widely explored over the past five years to construct low-dimensional vector representations of words and passages of text and entity boundary extraction. Neural embeddings are a common group of low-dimensional vector representation methods, phrases, or text (50-500 dimensions) whose values are assigned by a neural network functioning as a black box. However, it is not easy to view these dimensions in a meaningful way in clinical notes. Constructing neural embeddings requires extensive training and tuning of several parameters and hyperparameters.

In this study, we describe and evaluate a method for identifying clinical entities and entity boundary extraction from CRC clinical reports. Given a sequence of words, our model represented each word, phrase, or text using a concatenated low-dimensional vector of its corresponding word and character sequence embedding. Furthermore, a near-comprehensive vector representation was built of words and selected bigrams, trigrams, and abbreviations. The word vector and dictionary features were then projected into dense representations of the vectors. These were then fed into the Bidirectional Long Short-Term Memory (Bi-LSTM) to capture contextual features. Finally, a Conditional Random Field (CRF) was used to capture dependencies between adjacent tags. With the ability of Bio-Bi-LSTM-CRF, we captured detailed colonoscopy-related information within colonoscopy records, such as indicators, the number, size, and location of polyps, as well as the procedure and polyp histology (adenoma, tubulovillous, and adenocarcinoma).

The remainder of this paper is structured as follows. The Methods section explains the proposed Bio-Bi-LSTM-CRF in depth. Section Model Evaluation explains about strategies for model evaluation Section Results presents the experiments and results. Section Discussion discusses several important issues of the proposed model. Section Strengths and Limitations explains about strengths and limitations of the model and finally, Section conclusion concludes the current paper and points out potential future work.

Methods

Dataset and Annotation

A total of 44,405 de-identified colonoscopy reports, which represented all colonoscopies performed on 36,537 patients between 2002 and 2012, were extracted from four Veterans Affair Medical Centers (see Figure 1). These colonoscopy reports contained various sections, such as indications and findings, with corresponding textual input by a colonoscopist. The exact sections varied among the four facilities. Only sections containing text with information on the indication of polyps found during the colonoscopy were retained for further analysis. Furthermore, pathology reports included sections such as impressions and findings with text input by a pathologist. Only those sections that contained text with information on polyp pathology were extracted for further analysis.

Manual annotation of colonoscopy reports provided a gold standard to benchmark the automated methods. A total of 4,000 colonoscopy reports (1,000 from each of the four facilities) were randomly selected to annotate sentences. Of these reports, 80% were used as training data, 10 % as data validation set for hyperparameter tuning, and the remaining 10 % as test data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

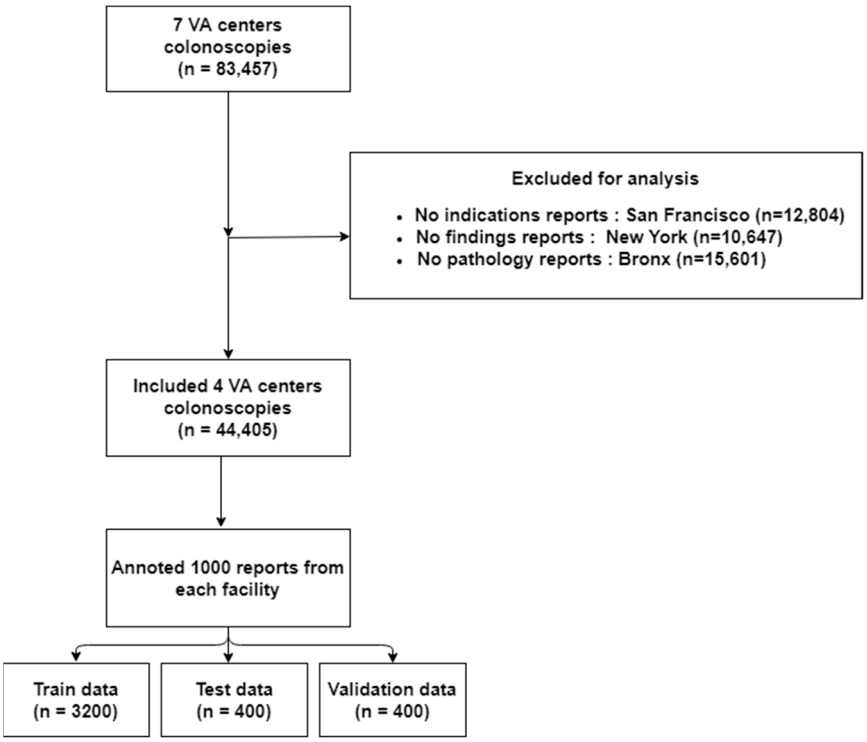

Figure 1. Use of colonoscopy reports for training, validating, and testing automated text extraction methods.

The annotation process used in this study was previously described in detail elsewhere [33]. Briefly, two annotators manually marked the sentences devised a list of lexical units (trigger phrases) and corresponding elements (attributes) for each sentence. Based on the existing literature, three common phenotyping tasks were selected. The final list of the variables is shown in Table 1.

Table 1. List of the variables of interest from colonoscopy reports.

Lexical Units	Description
Indicators	Reason or indications for colonoscopy examination: Abnormal computed tomography; Anemia; Diarrhea; Dyspepsia; Endo Foreign body removal; Hematochezia; History of colon cancer; M Modification of bowel habits; Neoplasia; Personal history of colonic p polypectomy site; Presence of fecal occult blood; Screening; Screenin surveillance for colon neoplasia; Sigmoidoscopy; Surveillance of patient neoplastic polyps; Treatment of bleeding from such lesions as va malformation; Ulceration; Weight loss
Polyp Descriptions	The diameter of the polyp: Small (1-4mm), medium (5-9mm), or large (>10mm) Morphology of the polyp: sessile, pedunculated, mass Number of polyps: either a numerical value or qualitative description such as many, several, numerous, etc.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

Location	Proximal: Ileocecal valve, Cecum, Hepatic flexure, Ascending, Transverse, and Splenic flexure Distal location: Descending, Sigmoid, Rectosigmoid, and Rectum
Polyp removal procedure	Type of procedure for removing the polyp
Polyp histology	Adenoma, Tubulovillous, Villous, High-grade dysplasia, Adenocarcinoma

The main components of the entity extraction pipeline are shown in Figure 2. The pipeline includes (1) data cleaning and pre-processing; (2) sentence boundary detection; (3) vector representation; and (4) the entity extraction model. We used Python 3.7 and the open-source packages NLKT, sci-kit–learn, and Tensor Flow that provide various NLP methods, machine learning methods, and neural network modules.

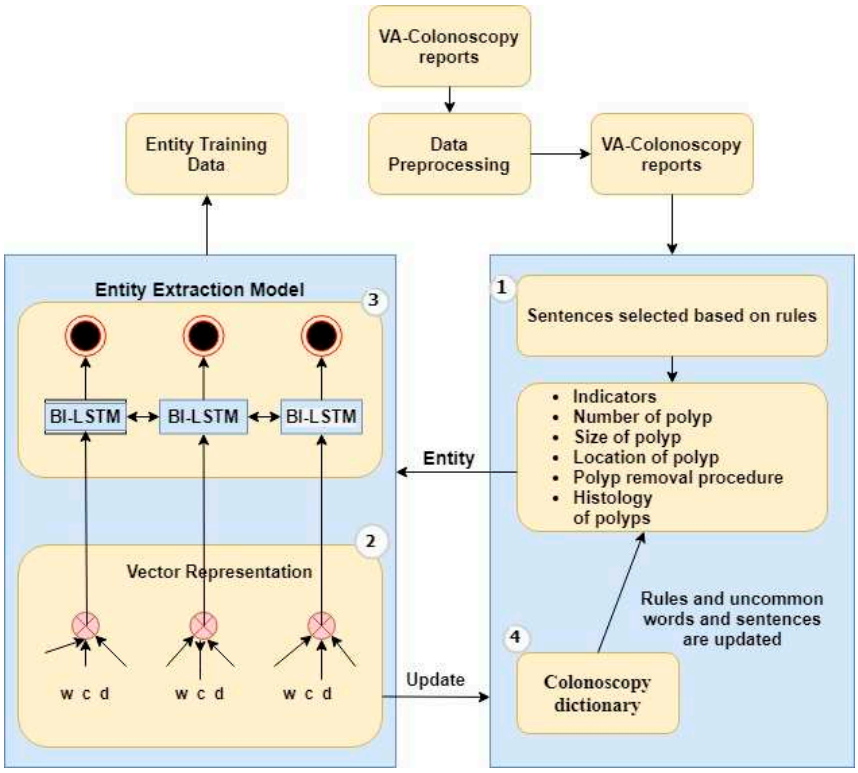

Figure 2. Model architecture. Our model had the following four parts: (1) sentence boundary detection; (2) vector representation; (3) entity extraction model; and (4) rule update.

Data cleaning and Preprocessing

We pre-processed the text reports to remove all special characters (e.g., commas in a list). Misspellings in our data set included typographical errors (e.g., "cacel" instead of "cecal"), phonetic errors that could be associated with lack of familiarity with medical terms (e.g., byopsi and methastasis), and everyday language errors (e.g., hoooooootsnare). The correction of spelling errors from clinical entries using the MetaMap and auto spell checker by Google's query suggestion service.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

Sentence Boundary Detection

Sentence segmentation is not a simple task in clinical data. Automatically segmenting clinical free text is an essential first step to information extraction. There are many existing sentence boundary detection (SBD) packages that primarily rely on punctuation marks and wide spaces. In most text, sentence boundaries are indicated by a period; however, colonoscopy reports frequently contain inter-sentence periods (e.g., "Dr. Patil" or "size of 3.5 cm" or "No. of polyps: "), which makes these reports different from typical biomedical texts. Theoretically, boundary failures can result from standard medical terminologies. To solve this problem, we applied a machine learning approach to identify the sentence boundaries.

To perform sentence boundary detection, it is essential to represent the input sentences with suitable tags. In the present study the SF format (where S=beginning of a sentence, F= ending of the sentence) was used for tagging in sentence boundary detection. The features and keywords listed in Table 4 were used to identify and tag the sentence boundaries. We used a decision tree for sentence boundary detection classification. This approach was constructed by applying the C4.5 algorithm [34]. An example is provided in Textbox 1.

Textbox 1. Example of a clinical note.

[S] Six sessile polyps were found in the sigmoid colon and in the proximal descending colon [F]. [S] The polyps were 4 to 15 mm in size. All polyps were completely removed by cold snare polypectomy [F].[S] Resection and retrieval were complete[F]. [S]Multiple small and large- mouthed diverticulitis were found in the sigmoid colon [F].

The clinical note extracted sentences containing essential lexical units. To avoid duplicate sentences and improve sentence diversity, the sentences were sorted by the TF-IDF cosine distance. The sentences were manually de-identified and checked by automatic sentence boundary detection.

Vector Representation

Clinical notes are characterized by lexical and morphological richness. Therefore, using the recognition of a named entity to capture morphological information from complex medical terms can help to avoid vocabulary problems and compensate for traditional embedding words. The name entity recognition in clinical colonoscopy notes is usually handled using a sequence labeling task. Given a sequence of words in a clinical notes sentence, we labelled each word in the sentence using the BIEOS (Begin, Inside, End, Outside, Single) tag scheme.

For example, in Textbox 2 below, "The mass was circumferential" and "Estimated blood loss was minimal" sentences are not significant. The widely used annotation modes for named entity recognition training data include BIO, BIEO, and BIEOS, where B is the beginning of an entity, I is the center of an entity, E is the end of an entity, S is a single entity, and O is not an entity.

Textbox 2. Example output after the sequence labeling task.

[S]A sessile polyp was found in the descending colon [F]. [S] The polyp was 8 mm in size [F]. [S] The polyp was removed with a cold snare [F]. [S] Resection and retrieval were complete [F]. [S] [O] Estimated blood loss was minimal [O] [F] . [S] A fungating partially obstructing mass that could be easily traversed was found in the sigmoid colon [F] .[S] [O] The mass was circumferential [O] [F] .[S] The mass measured five cm in length from 25 to 30 cm [F].

In this study, due to the uncertainty in the boundary of medical vocabulary terms and the complexity of representing uncommon and unknown entities, we created a new vector representation by integrating dictionary function vector. Each word's vector representation e_i consisted of word vector w_i and dictionary feature vector d_i as $e_i = w_i + d_i$ where + was the concatenation operator, as shown by the diagram. The word vector was constructed as a combination

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

of word embedding and character embedding. To represent character embedding, we used bidirectional LSTM to extract character features and implement certain character features.

Dictionary feature to vector

We used the n-gram models to divide the original sentence into text segments based on the context of the words. Next, combining with the domain dictionary (colonoscopy) C, we produced a binary value dependent on whether or not the text segments were present in C. Dictionary feature vectors were built in various dimensions based on the number of entity categories in the dictionary. Bi-LSTM converted the final dictionary function vector from the dictionary element.

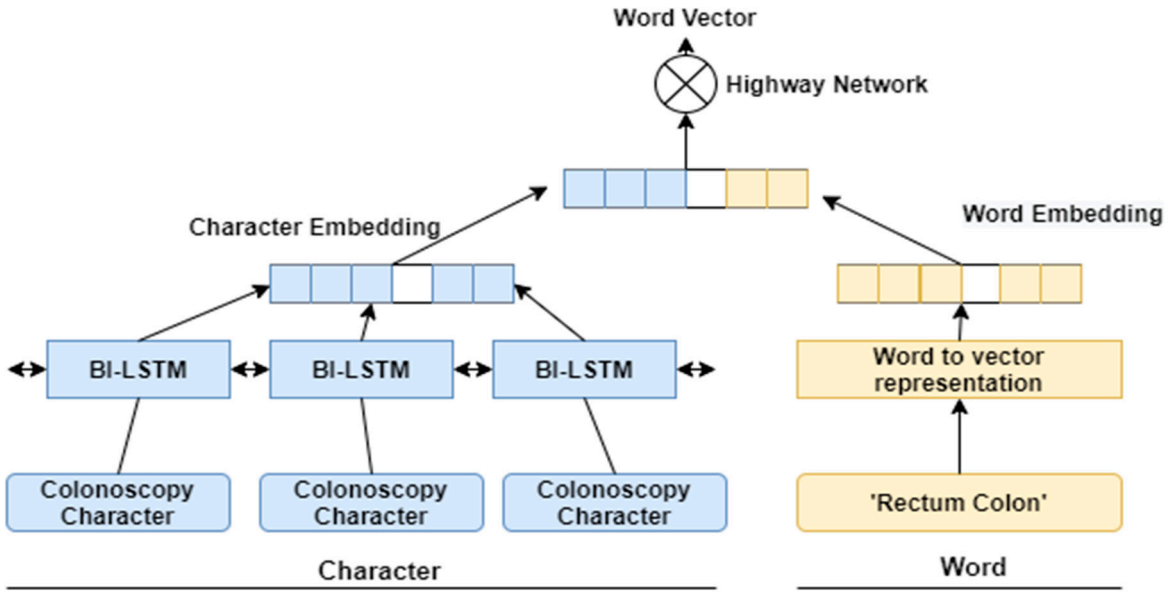

Figure 3. Word vector as a combination of character embedding and word embedding. Each word trains a word vector to the model.

Entity Extraction Model

The entity extraction using Bi-LSTM-CRF[35] uses word embedding as inputs; in the present study, we employed embedding character and word embedding. It allowed the model to be trained with more feature information. The combined character words and dictionary feature vectors were fed into the model. This step has the following advantages: (1) more details about the features and (2) better identification of the entity boundary problem. Next, we fed this into a Bi-LSTM model. Instead of splitting the two feature vectors as inputs, three separate Bi-LSTM versions of the proposed model are shown in Figure 3. To obtain a vector of the final character representation, we implemented a Bi-LSTM neural network and added the forward and backward sequences' output vectors. Our key aim of using Bi- LSTM was that it functions best to take on long-term dependencies in a phrase, and the bidirectional sequential design provides further advantages by remembering both a word's previous and potential meaning.

CRF Layer: Using the colonoscopy reports with the sentences are input for BI-LSTM neural network obtained a word representation, ignoring neighboring marker dependencies. For example, the tag "I" (Inside) cannot be followed with the tag "B"" (Begin) or "E" (End). We used conditional random fields (CRF) to determine labels from a sequence of background representations. Contrary to using CRF independently for tagging decisions, this approach may produce a greater tagging accuracy in general.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

Rules for Specific Task

We used manual rules to address the differences between unstructured colonoscopy clinical notes from multiple facilities. The aim of defining rules is to identify sentences that contain certain entities from the text. These rules were dynamically updated for each facility. The main five rules are explained in Table 2.

Table 2. Description of the five major rules.

Task	Rules
Sentence boundary detection	Keywords (e.g., Number of polyps, jar, etc.), period, semicolon, numPreBlanklines, and numPostBlanklines
Polyps size	Keywords (e.g., small, medium, large, etc.), number preceeding "CM", or number preceding "MM"
Location of polyps	Keywords (e.g., rectal, sigmoid, etc.) or distance from rectum given in "CM".
Word processing	Abbreviations (e.g., "SIG" for sigmoid, "REC" for rectal, etc.) and special word processing (e.g., "Recto-sigmoid" is a combination of "rectal" and "sigmoid", "Tubulovillous" is a combination of "Tubular" and "Villous", etc.)
Number of polyps	Keywords (polyps, number, numerics, and sessile)

Dataset and Experimental Settings

As previously mentioned, 4,000 colonoscopy reports were manually annotated to identify named entities of interest. The five medical named entities of interest are shown in Table 3. There were several sentences to every instance. These sentences were further separated into clauses by using tags "S" and "F" to mark sentence beginning and end (see Section "Sentence Boundary Detection"). We calculated micro-averaged precision, recall, and F1-score for exact matching words and classify sentences accordingly to conceptual meaning from colonoscopy reports. We then compared our model Bi-LSTM-CRF with improved NER and sentence boundary detection with the Bi-LSTM-CRF.

Table 3. Statistics of entity and sentence.

Type	Number of Entities	Number of Sentences
Indication	4,215	-
Number of polyps	9,542	4,687
Location of polyps	12,647	5,321
Size of polyps	8,654	3,549
Polyp removal procedure	3,657	2,648
Polyp histology	6,548	4,598

We used Tensor Flow to complete the CRF module. To this end, we first ran a bidirectional LSTM-CRF model forward pass for each batch, including the forward pass for both the forward and backward states of LSTM. We then divided the learning rate by five if the validation accuracy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

decreased and used mini-batches of size 100; the training was stopped when the learning rate went under the threshold of 10^{-8}. For classifiers, we set hidden layers size to 300; because neural network architectures are usually over-parameterized and vulnerable to overfitting to solve it, we used regularization strategies. Specifically, Elastic Net regularization was added to each LSTM hidden layer. The parameters predominantly included tag indices and batch size, where tag indices indicated the number of actual tags, and the batch size reflected the number of batch samples. The parameter settings for the proposed model are shown in Table 4.

Table 4. Parameter setting of the proposed method.

Parameter	Value
Word vector embedding size	200
Dictionary feature vector embedding size	100
hidden neurons for each hidden layer	300
Batch size	100
Tag Indices	4
Learning Rate	0.005
Number of epochs	10
Optimizer	Adam optimizer

Model evaluation

The proposed Bio-Bi-LSTM-CRF method and the state-of-the-art model Bi-LSTM-CRF were trained on the training subset, and the validation subset was used for tuning the hyper-parameters. Ultimately, the trained models were tested on the test subset. We used precision, recall, accuracy, and F1 score to evaluate the performance of the Bio-Bi-LSTM-CRF and Bi-LSTM-CRF models in terms of overall accuracy.

Evaluations of precision, recall, accuracy, and F1 score were performed for each report. For example, all of the polyp numbers, sizes, and locations in a colonoscopy report had to be correct for it to be considered accurate regardless of the number of polyps. Also, the precision, recall, accuracy, and F1 score were performed on the entity-level (e.g., each polyp size, number, and location) to find the models' performance for each facility. An exact match with the gold standard of manually annotated records was required for each entity versus the record-level where each record was considered separately.

Results

The proposed Bio-Bi-LSTM-CRF outperformed Bi-LSTM-CRF by achieving higher accuracy by 5.6%, 6.6%, and 19.1% percentage points in the indication, findings, and pathology reports, respectively (see Table 5). The Bio-Bi-LSTM-CRF model's absolute accuracy was relatively high for all three reports: 88.0% for findings reports, 93.5% for indication reports, and 96.5% for pathology reports. The proposed Bio-Bi-LSTM-CRF thus was found to have a 15.3% increase in precision, a 14.8% increase in recall, and a 10.3% increase in F1over the state-of-the-art model Bi-LSTM-CRF. The hierarchical influence structure of Bio-Bi-LSTM-CRF aided in reducing noisy text from clinical colonoscopy notes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

Table 5. Comparative results of the Bi-LSTM-CRF and Bio-Bi-LSTM-CRF models.

Reports	Method	Accuracy	Recall	Precision	F1
Indication Reports	Bi-LSTM-CRF	88.5% (354/400)	82.5%	82.5%	83.2%
	Bio-Bi-LSTM-CRF	93.5% (374/400)	93.0%	93.0%	93.4%
Findings Report	Bi-LSTM-CRF	82.5% (300/400)	83.5%	84.0%	82.7%
	Bio-Bi-LSTM-CRF	88.0% (352/400)	88.7%	89.0%	88.1%
Pathology Report	Bi-LSTM-CRF	81.0% (324/400)	77.6%	75.0%	79.7%
	Bio-Bi-LSTM-CRF	96.5% (386/400)	96.0%	96.0%	96.4%

Both the Bi-LSTM-CRF and Bio-Bi-LSTM-CRF models performed better in entity identification indications reports, and average accuracy of pathology reports amounted to 84.75% and 95%, respectively (see Table 6). However, the two models struggled in organizing the findings reports that mentioned number, location, and removal procedure of polyps (e.g., "two diminutive mass polyps"), as well as the reports where the status of the location of the polyp was defined by long-range dependencies and words (e.g., "a single small 3-5mm polyp away from rectum and sigmoid colon"). In contrast to findings reports that contained multiple sentences with information about polyps, indication reports and pathology reports had no long dependency sentences and multiple sentences. To understand the characteristics of polyps, the model had to reconsider the sentence and to relate "polyp" or "sessile."

Table 6. Comparative results by facility between Bi-LSTM-CRF and Bio-Bi-LSTM-CRF model.

Facility	Methods	Accuracy	Recall	Precision	F1
Albany	Bi-LSTM-CRF	83.0% (332/400)	79.0%	81.0%	78.0%
	Bio-Bi-LSTM-CRF	92.0% (368/400)	91.5%	91.0%	90.0%
Ann Arbor	Bi-LSTM-CRF	86.0% (344/400)	84.0%	82.0%	85.0%
	Bio-Bi-LSTM-CRF	94.0% (376/400)	92.0%	93.0%	93.5%
Detroit	Bi-LSTM-CRF	83.0% (332/400)	83.0%	78.0%	81.0%
	Bio-Bi-LSTM-CRF	91.0% (364/400)	91.0%	90.5%	91.0%
Indianapolis	Bi-LSTM-CRF	80.0% (320/400)	81.0%	82.0%	84.0%
	Bio-Bi-LSTM-CRF	93.0% (372/400)	95.5%	95.5%	95.5%

On the other hand, Bio-Bi-LSTM-CRF models entity identification on findings reports correctly. Bi-LSTM-CRF models achieved an average of 92.6% recall, as shown in Table 8. The Bio-Bi-LSTM-CRF model only made a few misclassifications of entities where the polyp location was represented in the form of distance from the rectum (130 cm). The number of polyps was described in an atypical fashion, including uncommon terms such as "semi-mass of large 10mm polyp," which only appeared in one and a few notes respectively in the dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 January 2023 doi:10.20944/preprints202301.0061.v1

The Bio-Bi-LSTM-CRF model's overall accuracy was between 8.0% and 13.0% points higher than the Bi-LSTM-CRF method in entity identification. The accuracy of the Bio-Bi-LSTM-CRF methods between facilities ranged from 92.0% to 94.0%, as shown in Table 7. The F1-scores of Albany facility, Ann arbor facility, Detr oit facility, and Indianapolis facility are 93.7%, 96.3%, and 87.1%; as shown in Table 8. The Bio-Bi-LSTM model achieved the highest overall results for all three facilities. The proposed Bio-Bi-LSTM-CRF thus reports a 12.8% increase in the F1 score. Likely, our Bio-Bi-LSTM-CRF model adapted well to other physicians' reporting style and language to achieve comparable performance in other facilities. For the polyp removal procedure, the performance of the two models of about the same level (at around 90.0%).

In comparison, the Bio-Bi-LSTM-CRF model shows an improvement of 20.9% F1 score over the Bi-LSTM-CRF in a type of CRC information from pathology reports. Bi-LSTM-CRF is struggling to extract entities from long dependences terms and multiple sentences. The weakest performance area was for location information of polyp. There is a large variety of location representation in the dataset. These are also relatively specific and unlikely to occur in the other facility data.

Table 7. Comparative results by performance evaluation between Bi-LSTM-CRF and Bio-Bi-LSTM-CRF model.

Category	Methods	Accuracy	Recall	Precision	F1
Indicator	Bi-LSTM-CRF	83.5% (334/400)	82.5%	82.5%	83.2%
	Bio-Bi-LSTM-CRF	93.5% (374/400)	93.1%	93.0%	93.4%
Number of polyps	Bi-LSTM-CRF	80.0% (320/400)	76.0%	86.0%	77.0%
	Bio-Bi-LSTM-CRF	85.0% (340/400)	88.0%	89.0%	89.0%
Size	Bi-LSTM-CRF	78.0% (312/400)	78.0%	80.5%	79.0%
	Bio-Bi-LSTM-CRF	81.0% (324/400)	88.0%	89.0%	90.0%
Location of polyp	Bi-LSTM-CRF	82.0% (328/400)	83.0%	78.0%	82.0%
	Bio-Bi-LSTM-CRF	94.0% (376/400)	86.0%	87.0%	85.0%
Polyp removal procedure	Bi-LSTM-CRF	89.0% (356/400)	94.0%	93.0%	84.0%
	Bio-Bi-LSTM-CRF	92.0% (368/400)	94.0%	93.0%	92.0%
Type of CRC	Bi-LSTM-CRF	81.0% (324/400)	77.6%	75.0%	79.7%
	Bio-Bi-LSTM-CRF	96.5% (386/400)	96.0%	96.0%	96.4%

Discussion

In this study, we proposed a Bio-Bi-LSTM-CRF model for an efficient and accurate entity extraction from colonoscopy report notes, which could have broader applications to other free-text medical narratives. The proposed Bio-Bi-LSTM-CRF was found to analyze a large sample of colonoscopy reports accurately. Our findings demonstrated a clear improvement in accuracy, precision, and recall within an academic healthcare system. The Bio-LSTM-CRF model consistently obtained better entity extraction accuracy than the Bi-LSTM-CRF model with identical feature sets. Our results also demonstrated that the addition of sentence boundary detection, rules, and domain dictionary in the tasks could significantly boosted accuracy, thus demonstrating the efficacy of merging practices and domain dictionary in entity extraction tasks.

Our results highlight several advantages of using Bio-Bi-LSTM-CRF in routine quality measurement using data in EHRs. The critical advantage of Bio-Bi-LSTM-CRF is that it is efficient and economically feasible as compared with manual data review. Otherwise, it would be expensive and time-consuming to review tens of thousands of colonoscopies reports manually. Another advantage is of using Bio-Bi-LSTM-CRF is that it allows providers to continue to use natural narrative when describing patient care. There has been criticism that structured note systems in current EHRs force providers to create unnatural and overly structured notes, which takes extra time to develop and impedes communication because these notes are difficult to read [36].

Strengths and Limitations

The present study has several limitations. First, compared to the direct and implicit term metrics, the neural embeddings were found to have inferior performance on the biomedical term similarity/relatedness benchmarks, as assessed both by our tests (against several different implementations of word2vec that other groups have developed for similar tasks) and those reported by others (see the Results sections). However, word2vec performance can be optimized for specific tasks by adjusting different choices of parameters and hyper-parameters, different neural network architectures, and other ways of handling multi-word phrases [37,38]. Second, we did not test all possible implementations, which limits generalizability of our results neural embeddings in general. We decided to include only the most essential, informative words and phrases in our vocabulary in contrast to the word2vec based embeddings that encompass a much larger vocabulary. We feel that the relatively prevalent words and phrases are likely to be the most valuable for representing biomedical articles. However, it is worth exploring the effects of different vocabulary restrictions, especially when applying our vectors to other corpora, domains, and tasks.

In many aspects, our research is novel. First, to our knowledge, this is only the research to examine the validity of NLP for multiple center data to investigate indications, findings, and pathology laboratory results of dictated consultation notes [39,40]. Secondly, previous NLP research in CRC clinical notes concentrates exclusively on a single clinical condition, such as tumor presence[23]. Our analysis is more comprehensive than other more widely reported studies, looking at four different medical centers[41]. There are several limitations to our research. In our initial training dataset (n=4,000), our choice to randomly sample resulted in under-sampling of CRC cases. As a result, the model was never trained on this features like stool testing. stool testing has several valuable features, and subsequently performed poorly during the final test set for this feature, possibly underestimating a properly trained model's effectiveness. This suggests that a real-world application of this technology may require a more purposive sampling strategy than our random sampling approach. Although the inclusion of many bigrams, trigrams, and abbreviations should reduce the problem of term ambiguity to some extent, the vector representations of words represent an overall average across word instances and different word senses. The implicit term metrics have relatively few parameters, which we have attempted to set at near-optimal values. However, the choice of 300-dimensions for the vector representations, and the particular weighting schemes for the similarity of vectors, may be regarded as best guesses and might be tuned further for optimal performance in specific tasks. However, this study's goal was to establish the feasibility of using a deep learning model to extract clinical features from dictated consult notes and inform the approach to more extensive future studies.

Conclusion

Our proposed Bio-BI-LSTM-CRF improved the perception of unknown entities, which means that the model should have a better capability to deal with emerging medical concepts and multicenter data without extra training resources. The pre-processing of the word-level input of the model with external word embeddings allowable to improve performance further and achieve state-of-the-art for the colonoscopy clinical notes entity extraction. This idea could not only be applied in medical entity extraction but also other medical named entity recognition applications such as drug names and adverse drug reactions, as well as named entity recognition tasks in different fields.

References

1. Alblawi, A.S. and A.A. Alhamed. *Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, NLP and analytics*. in *2017 IEEE Conference on Big Data and Analytics (ICBDA)*. 2017. IEEE.
2. Denecke, K. *Extracting medical concepts from medical social media with clinical NLP tools: a qualitative study*. in *Proceedings of the fourth workshop on building and evaluation resources for health and biomedical text processing*. 2014. Citeseer.
3. Strubell, E., A. Ganesh, and A. McCallum, *Energy and policy considerations for deep learning in NLP.* arXiv preprint arXiv:1906.02243, 2019.
4. Jiang, F., et al., *Artificial intelligence in healthcare: past, present and future.* Stroke and vascular neurology, 2017. **2**(4).
5. Dinov, I.D., *Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data.* Gigascience, 2016. **5**(1): p. s13742-016-0117-6.
6. Mishuris, R.G. and J.A. Linder, *Electronic health records and the increasing complexity of medical practice:"it never gets easier, you just go faster".* Journal of general internal medicine, 2013. **28**(4): p. 490-492.
7. Polnaszek, B., et al., *Overcoming the Challenges of Unstructured Data in Multi-site, Electronic Medical Record-based Abstraction.* Medical care, 2016. **54**(10): p. e65.
8. Meystre, S., *Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress (2017).*
9. Osborne, J.D., et al., *Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning.* Journal of the American Medical Informatics Association, 2016. **23**(6): p. 1077-1084.
10. Schwartz, J.T., et al., *Applications of machine learning using electronic medical records in spine surgery.* Neurospine, 2019. **16**(4): p. 643.
11. Holzinger, A., et al. *Combining HCI, natural language processing, and knowledge discovery-potential of IBM content analytics as an assistive technology in the biomedical field*. in *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. 2013. Springer.
12. UmaMaheswaran, S., et al., *Empirical Analysis for Improving Food Quality Using Artificial Intelligence Technology for Enhancing Healthcare Sector.* Journal of Food Quality, 2022. **2022**.
13. Kumar Attar, R., *The Emergence of Natural Language Processing (NLP) Techniques in Healthcare AI*, in *Artificial Intelligence for Innovative Healthcare Informatics*. 2022, Springer. p. 285-307.
14. Park, E.H., et al., *Evaluating the impact on clinical task efficiency of a natural language processing algorithm for searching medical documents: Prospective crossover study.* medRxiv, 2022.
15. Jemal, A., et al., *Cancer statistics, 2010.* CA: a cancer journal for clinicians, 2010. **60**(5): p. 277-300.
16. Center, M.M., A. Jemal, and E. Ward, *International trends in colorectal cancer incidence rates.* Cancer Epidemiology and Prevention Biomarkers, 2009. **18**(6): p. 1688-1694.
17. Denny, J.C., et al., *Extracting timing and status descriptors for colonoscopy testing from electronic medical records.* Journal of the American Medical Informatics Association, 2010. **17**(4): p. 383-388.
18. Goldberg, Y., *Neural network methods for natural language processing.* Synthesis lectures on human language technologies, 2017. **10**(1): p. 1-309.
19. Jagannatha, A.N. and H. Yu. *Structured prediction models for RNN based sequence labeling in clinical text*. in *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing*. 2016. NIH Public Access.
20. Pons, E., et al., *Natural language processing in radiology: a systematic review.* Radiology, 2016. **279**(2): p. 329-343.
21. Imler, T.D., et al., *Natural language processing accurately categorizes findings from colonoscopy and pathology reports.* Clinical Gastroenterology and Hepatology, 2013. **11**(6): p. 689-694.
22. Li, F., et al., *Recognizing irregular entities in biomedical text* via *deep neural networks.* Pattern Recognition Letters, 2018. **105**: p. 105-113.
23. Denny, J.C., et al., *Natural language processing improves identification of colorectal cancer testing in the electronic medical record.* Medical Decision Making, 2012. **32**(1): p. 188-197.
24. D'Avolio, L.W., et al., *Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC).* Journal of the American Medical Informatics Association, 2010. **17**(4): p. 375-382.
25. Halligan, S., et al., *CT colonography in the detection of colorectal polyps and cancer: systematic review, meta-analysis, and proposed minimum data set for study level reporting.* Radiology, 2005. **237**(3): p. 893-904.

26. Ratinov, L. and D. Roth. *Design challenges and misconceptions in named entity recognition*. in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. 2009.
27. Zhang, J., et al., *Enhancing HMM-based biomedical named entity recognition by studying special phenomena.* Journal of biomedical informatics, 2004. **37**(6): p. 411-422.
28. Wong, D.F., L.S. Chao, and X. Zeng, *isentenizer-: Multilingual sentence boundary detection model.* The Scientific World Journal, 2014. **2014**.
29. Palmer, D.D., *SATZ-an adaptive sentence segmentation system.* arXiv preprint cmp-lg/9503019, 1995.
30. Kreuzthaler, M. and S. Schulz. *Detection of sentence boundaries and abbreviations in clinical narratives*. in *BMC medical informatics and decision making*. 2015. BioMed Central.
31. Goldberg, Y. and O. Levy, *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method.* arXiv preprint arXiv:1402.3722, 2014.
32. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.
33. Si, Y. and K. Roberts. *A frame-based NLP system for cancer-related information extraction*. in *AMIA Annual Symposium Proceedings*. 2018. American Medical Informatics Association.
34. Ruggieri, S., *Efficient C4. 5 [classification algorithm].* IEEE transactions on knowledge and data engineering, 2002. **14**(2): p. 438-444.
35. Huang, Z., W. Xu, and K. Yu, *Bidirectional LSTM-CRF models for sequence tagging.* arXiv preprint arXiv:1508.01991, 2015.
36. Walsh, S.H., *The clinician's perspective on electronic health records and how they can affect patient care.* Bmj, 2004. **328**(7449): p. 1184-1187.
37. Al-Sadi, A.M., et al., *Genetic analysis reveals diversity and genetic relationship among Trichoderma isolates from potting media, cultivated soil and uncultivated soil.* BMC microbiology, 2015. **15**(1): p. 1-11.
38. Henry, S., C. Cuffy, and B.T. McInnes, *Vector representations of multi-word terms for semantic relatedness.* Journal of biomedical informatics, 2018. **77**: p. 111-119.
39. Imler, T.D., et al., *Multi-center colonoscopy quality measurement utilizing natural language processing.* American Journal of Gastroenterology, 2015. **110**(4): p. 543-552.
40. Lee, J.K., et al., *Accurate identification of colonoscopy quality and polyp findings using natural language processing.* Journal of clinical gastroenterology, 2019. **53**(1): p. e25.
41. Raju, G.S., et al., *Natural language processing as an alternative to manual reporting of colonoscopy quality metrics.* Gastrointestinal endoscopy, 2015. **82**(3): p. 512-519.