International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00433-5

REGULAR PAPER

Surgicberta: a pre-trained language model for procedural surgical language

Marco Bombieri[1] · Marco Rospocher[2] · Simone Paolo Ponzetto[3] · Paolo Fiorini[1]

Received: 3 March 2023 / Accepted: 12 July 2023

Abstract
Pre-trained language models are now ubiquitous in natural language processing, being successfully applied for many different tasks and in several real-world applications. However, even though there is a wealth of high-quality written materials on surgery, and the scientific community has shown a growing interest in the application of natural language processing techniques in surgery, a pre-trained language model specific to the surgical domain is still missing. The creation and public release of such a model would serve numerous useful clinical applications. For example, it could enhance existing surgical knowledge bases employed for task automation, or assist medical students in summarizing complex surgical descriptions. For this reason, in this paper, we introduce SURGICBERTA, a pre-trained language model specific for the English surgical language, i.e., the language used in the surgical domain. SURGICBERTA has been obtained from ROBERTA through continued pre-training with the Masked language modeling objective on 300 k sentences taken from English surgical books and papers, for a total of 7 million words. By publicly releasing SURGICBERTA, we make available a resource built from the content collected in many high-quality surgical books, online textual resources, and academic papers. We performed several assessments in order to evaluate SURGICBERTA, comparing it with the general domain ROBERTA. First, we intrinsically assessed the model in terms of perplexity, accuracy, and evaluation loss resulting from the continual training according to the masked language modeling task. Then, we extrinsically evaluated SURGICBERTA on several downstream tasks, namely (i) procedural sentence detection, (ii) procedural knowledge extraction, (iii) ontological information discovery, and (iv) surgical terminology acquisition. Finally, we conducted some qualitative analysis on SURGICBERTA, showing that it contains a lot of surgical knowledge that could be useful to enrich existing state-of-the-art surgical knowledge bases or to extract surgical knowledge. All the assessments show that SURGICBERTA better deals with surgical language than a general-purpose pre-trained language model such as ROBERTA, and therefore can be effectively exploited in many computer-assisted applications in the surgical domain.

Keywords Transformers · Language models · Natural language processing · Medicine

1 Introduction

The field of artificial intelligence known as natural language processing (NLP) allows for automated processing and analysis of everyday language. In the past two decades, NLP has rapidly expanded across all information technology domains and is now being utilized more frequently in medicine. Its applications include enhancing the use of unstructured electronic health records, aiding communication with patients, conducting consultations, and finding pertinent information in papers [21]. Most cutting-edge NLP techniques rely on statistical language modeling, which involves representing words as numerical vectors that capture their probability distribution in a sentence structure [12]. These vectors, also known as word embeddings, are numerical representations of words and are frequently generated through self-supervised machine learning methods applied to large, unlabeled textual datasets. More advanced language models create distinct representations for a word based on its context, allowing them to accurately capture polysemous terms that have multiple meanings. Contextual language models based on Transformer architectures, such as BERT [8] or RoBERTa [20], are

✉ Marco Bombieri
marco.bombieri_01@univr.it

1 Department of Computer Science, University of Verona, Verona, Italy

2 Department of Foreign Languages and Literatures, University of Verona, Verona, Italy

3 DWS Group, University of Mannheim, Mannheim, Germany

Published online: 16 August 2023

trained using a deep neural network with a masked language modeling (MLM) objective [33]. These models use a bidirectional self-attention mechanism [34] to associate each word with its context, or the words surrounding it in the sentence. These features enable contextual language models to outperform non-contextual ones in various NLP tasks [8]. Although trained on enormous digital corpora consisting of billions of words, language models trained on general text frequently do not work effectively in very specialized domains such as scientific ones. As a result, several recent NLP studies have concentrated on retraining or fine-tuning language models for very specialized domains using domain-specific text (as explained in detail in Sect. 2).

While a large number of domain-specific language models have been developed to improve the understanding of the semantic information in their field of expertise, to the best of our knowledge a specialized model for surgical language does not exist yet, even if the scientific community has shown growing interest in the application of NLP in surgery [19, 28, 38–40]. There is an abundance of high-quality resources in the surgical literature, including books, online materials, and academic papers that are adopted and utilized by universities around the globe. The vast quantity of this high-quality available information can be a valuable resource for various clinical applications, involving both humans and smart robotics systems, if automatically processed via NLP techniques. For instance, one possible application of using the content extracted from textual resources is for building or extending the knowledge bases exploited by surgical robots, which they can use to make informed decisions in real-life intervention situations. Similarly, as reported in recent studies focusing on the clinical field [30, 42], humans can also benefit from this information in question-answering applications. These systems could be useful for medical students during their early training phase, or to provide a summary or simplified version of surgical descriptions.

In this paper, we follow this line of research and introduce a new pre-trained language model trained on procedural surgical language, named SurgicBERTa. The main, novel contributions[1] presented in this paper are:

1. The development of SurgicBERTa, a pre-trained language model specific for the understanding of procedural surgical language;
2. The intrinsic evaluation of SurgicBERTa with respect to the general-purpose model RoBERTa;
3. The extrinsic evaluation of SurgicBERTa with respect to RoBERTa, that is, the comparison of their performances when employed on four different downstream tasks;
4. The public release of SurgicBERTa to the research community: https://gitlab.com/altairLab/surgicberta.

The quantitative assessments are complemented with qualitative analysis on SurgicBERTa, showing that it contains a lot of surgical domain knowledge that could be useful to enrich existing state-of-the-art surgical knowledge bases. The evaluation indicates that SurgicBERTa better deals with surgical language than a state-of-the-art yet open-domain and general-purpose model such as RoBERTa, and therefore can be effectively exploited in many computer-assisted applications, specifically in the surgical domain.

The paper is organized as follows: Sect. 2 revises relevant works in this area. Then, SurgicBERTa is presented in Sect. 3. The required textual data is collected, extracted, pre-processed and used for the continuous training of RoBERTa on the MLM task with domain-specific text. Section 4 presents the intrinsic metrics and tasks used to evaluate SurgicBERTa. In particular, metrics for the intrinsic evaluation of SurgicBERTa (i.e., perplexity, accuracy, and evaluation loss of the MLM task) are presented in Sect. 4.1, while Sects. 4.2–4.5 present the downstream tasks used to compare SurgicBERTa with RoBERTa, namely, (i) procedural sentences detection, (ii) procedural knowledge extraction, (iii) ontological information discovery, and (iv) surgical terminology acquisition. Section 4.6 reports and qualitatively discusses some examples of surgical domain knowledge contained in SurgicBERTa. Finally, Sect. 5 summarizes obtained results and proposes future works.

2 Related works

2.1 Transformers and pre-trained language models

Transformers are deep-learning models widely used in NLP [34] and computer vision [9]. In particular, they have fundamentally changed the landscape of NLP by gradually replacing recurrent neural networks across the board. The core innovative part of these architectures is the self-attention mechanism [34]. Since one word can have different meanings in different contexts, self-attention allows the model to look at other positions in the input sequence for clues that can help lead to a better encoding for the current word. Moreover, the creation of large-scale, Transformer-based pre-trained language models such as BERT or RoBERTa has revolutionized the NLP domain. These models only use the encoder part of the Transformer (in contrast, e.g., to denoising autoencoders such as BART [16]). Such pre-trained large models are pre-trained once in an unsupervised way, e.g., on a language model objective, and can be fine-tuned for a large number of NLP tasks with a modest amount of training data, achieving state-of-the-art results on many of them, such as sentiment

[1] All the materials and results presented in this paper are novel, except the experiments described in Sect. 4.3, which were first presented in [4].

analysis, textual entailment, and natural language inference, crucially also across languages [15].

2.2 Pre-trained language models in biomedicine

Transformer-based pre-trained language models have also been fine-tuned for different tasks in the biomedical domain. However, they were originally built for general English, and thus they may miss some domain words or expressions. To overcome this limit, there is the possibility to train from scratch a model specific to a given domain of interest, such as in [42] where a large model specific to the clinical domain using $>$ 90 billion words of text is proposed. Developing such a model from scratch is very expensive for the computational resources and the training time required. For this reason, domain adaptation techniques, such as the MLM described in Sect. 3, have been proposed and widely used in biomedicine with fine-tuning for various downstream tasks. In [44], domain adaptation is used to obtain a cancer domain-specific language model for effectively extracting breast cancer phenotypes from electronic health records.

In [37], the authors utilize pre-trained neural models to classify patients as either seizure-free or not, as well as to extract text from clinical notes that contains their seizure frequency and the date of their last seizure. The first step of this pipeline is the unsupervised domain adaptation, using progress notes that were not selected for annotation. The obtained model has been fine-tuned for the classification and extraction tasks. Also, [41] adopted a domain adaptation technique on clinical notes from the Medical Information Mart for Intensive Care III database [14] to extract clinically relevant information. In [18], causal precedence relations are recognized among the chemical interactions in the biomedical literature to understand the underlying biological mechanisms. However, detecting such causal relations can be challenging because annotating such causal relation detection datasets requires considerable expert knowledge and effort. To overcome this limitation, in-domain pre-training of neural models with knowledge distillation techniques has been adopted, showing that the neural models outperform previous baselines even with a small number of annotated data. In [7], a domain adaptation strategy is adopted to encourage the model to learn features from the context to curate all validated antibiotic resistance genres, i.e., the ability of bacteria to survive and propagate in the presence of antibiotics, from scientific papers. In [30], a domain adaptation technique has been used to align large language models to new medical domains, showing that, after a proper adaptation step, they encode some clinical knowledge usable in question-answering applications. Finally, a domain adaptation technique has been adopted for biomedical domain adaptation in languages different than English, such as Spanish [6] and Chinese [43] showing the same improvement trend when compared to the corresponding base models.

However, due to the syntactic, semantic, and terminological differences between domains, it is often difficult to use these models to gain benefits outside the domain they were trained on. It is generally accepted that model performance may degrade when evaluated on data with a different distribution [31]. Consequently, domain adaptation on relevant domain data is essential to improve performance in very specialized domains [1], and despite the availability of several biomedical language models, to the best of our knowledge, a pre-trained surgical language model is missing. Such a model is essential for mining surgical procedural knowledge from text and developing intelligent surgical systems.

3 A language model for the surgical domain: SurgicBERTa

This section describes the development of SurgicBERTa, the pre-trained language model for the surgical domain that we contribute. SurgicBERTa has been developed on top of RoBERTa, an already available pre-trained language model for English for the general domain. Specifically, the `roberta-base` version of the HuggingFace Transformer library has been adopted. Therefore, the evaluation (presented in Sect. 4) will compare these two models along several dimensions.

RoBERTa [20] is a Transformer model that adopts the same encoder–decoder architecture made popular by BERT [8], while being trained on a larger quantity of data, consisting in a combination of datasets totaling around 160 GB of raw text: namely, texts from BookCorpus and English Wikipedia, data from the English portion of the CommonCrawl News, from OpenWebText, and some stories from CommonCrawl data. RoBERTa has been trained via MLM with dynamic masking: i.e., each time a sequence is input to the model, a new masking pattern is created. Differently from BERT, RoBERTa was not trained also on next sentence prediction, as this training task did not contribute a significant improvement of the performance in downstream tasks [20].

Leveraging RoBERTa as a starting point, we developed a new model that is tailored to the surgical domain. This involved the continuous training of RoBERTa on a large corpus of surgical text for the MLM unsupervised task. In the MLM task, a token w_t is replaced with $\langle mask \rangle$ and predicted using all past and future tokens $\boldsymbol{W}_{\setminus t} := (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{t-1}, \boldsymbol{w}_{t+1}, \ldots, \boldsymbol{w}_{|W|})$. Figure 1 illustrates the MLM task used to derive SurgicBERTa.

In more detail, to obtain a surgical model as general as possible, we collected 300 K sentences (7 M million words) from surgery books covering several heterogeneous surgical domains, including, for instance, orthopedics, abdominal

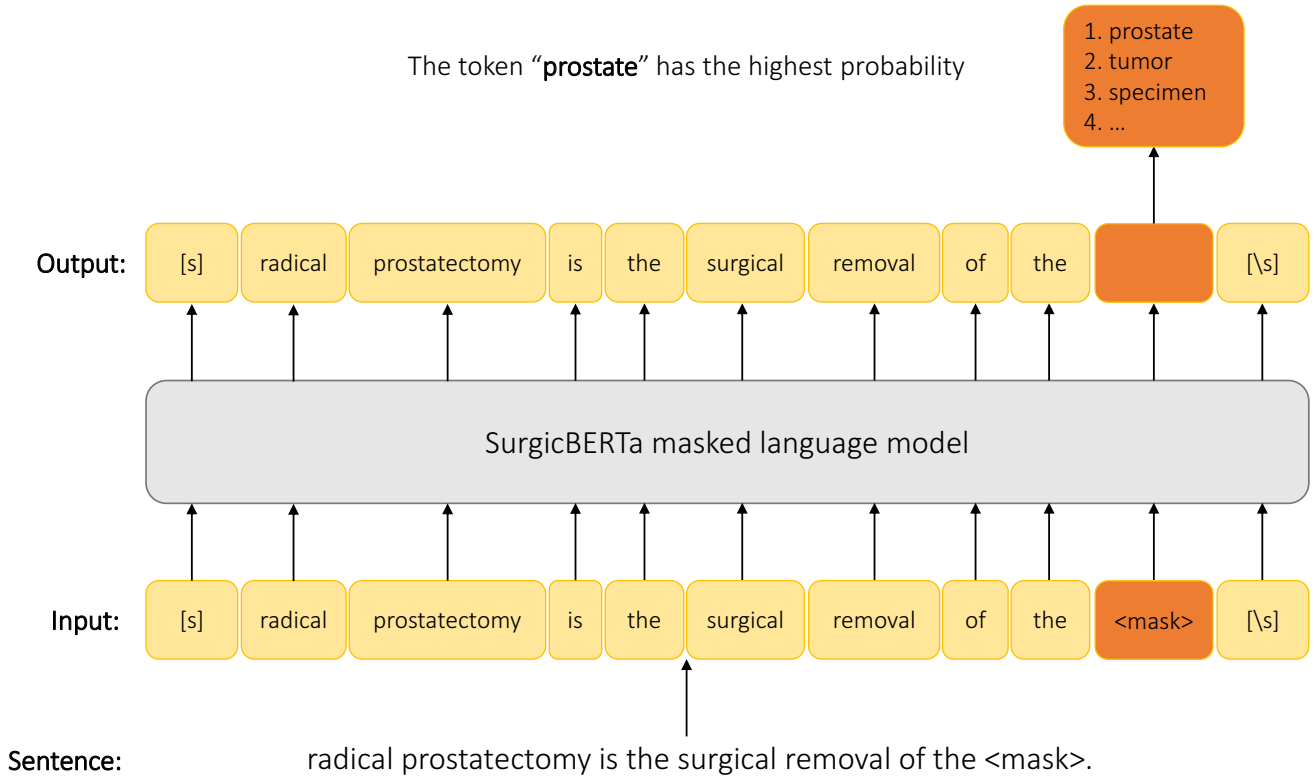

Fig. 1 MLM task used for adapting SURGICBERTA to the surgical domain. s and \s are special tokens denoting the sentence's beginning and end, respectively

surgery, and eye surgery. We searched for surgery books written in English on the web pages of several publishing houses. As keywords, we used the name of the surgical macro-areas (e.g., general surgery, abdominal surgery, gynecology surgery, eye surgery, etc.). From the results, we downloaded the digital version only of the texts to which our universities have proper free legitimate access.[2] A very minimal pre-processing of the sentences was performed, mainly to clean the text from bibliographic references and URLs. In more detail, 15% of tokens are selected for possible replacement. Among those selected tokens, 80% are replaced with the special $\langle mask \rangle$ token, 10% are left unchanged and 10% are replaced by a random token. The model is then trained to predict the initial masked tokens using cross-entropy loss. Following the ROBERTA approach, tokens are dynamically masked instead of fixing them statically for the whole dataset during pre-processing. This improves variability and makes the model more robust when training for multiple epochs. SURGICBERTA is computed using one NVIDIA RTX A6000 GPU, with 48 GB of GPU memory. We trained for 30 epochs with a learning rate of $5e{-}06$ and a batch size of 32. The Adam optimizer has been used. The implementation is based on PyTorch and Transformers libraries. The entire training required about 8 hours to be completed.

4 Evaluation

This section presents the intrinsic evaluation (Sect. 4.1) and the four downstream tasks that we use to evaluate SURGICBERTA in Sect. 4.2 through 4.5, namely: procedural/non-procedural surgical sentence classification, surgical information extraction, ontological information discovery, and surgical terminology acquisition.

4.1 Intrinsically evaluating the quality of language modeling

4.1.1 Evaluation metrics

Perplexity is one of the most common metrics for evaluating language models and measures the degree of uncertainty of a language model to generate a new token, averaged over very long sequences [27]. This means that the lower the perplexity,

[2] By choosing only the texts to which we have free access thanks to our institutions' agreement, we have probably excluded some resources which, if used in the training phase, would have allowed us to increase the training material and therefore probably also the performance. However, given the high diversity of the resources made available by our institutions and used for the training material, we do not believe that this choice has too much impact on performance.

calculated as the exponentiated average negative log likelihood of a sequence, the better the language model is able to predict a given text. While perplexity can be computed out of the box for traditional language models trained on guessing the next word given the previous context, i.e., autoregressive or causal language models, it is not well defined for language models like BERT or RoBERTa trained with the masked language modeling technique. For these models, we can compute instead the perplexity from their *pseudo-log likelihood scores (PPL)* [36], which corresponds to the sum of conditional log probabilities of each sentence token [29]. Formally, the pseudo-log likelihood scores (PPL) of a sentence $\boldsymbol{W} = (w_1, \ldots, w_{|\boldsymbol{W}|})$ under a language model with parameters Θ is defined as:

$$\mathrm{PPL}(\boldsymbol{W}) := \sum_{t=1}^{|\boldsymbol{W}|} \log P_{\mathrm{MLM}}(\boldsymbol{w}_t | \boldsymbol{W}_{\setminus t}; \Theta)$$

where $P_{\mathrm{MLM}}(\boldsymbol{w}_t | \boldsymbol{W}_{\setminus t}; \Theta)$ is the conditional probability of token $\boldsymbol{w}_t$ given all past and future tokens $\boldsymbol{W}_{\setminus t} := (w_1, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{|W|})$.

The (pseudo) *perplexity* PP of a masked language model [27] on a corpus of sentences $\mathbb{W}$ is then computed as:

$$\mathrm{PP}(\mathbb{W}) := \exp\left(-\frac{1}{N} \sum_{\boldsymbol{w} \in \mathbb{W}} PPL(\boldsymbol{W})\right)$$

where N is the number of tokens in the corpus. By computing PP on a test corpus for both RoBERTa and SurgicBERTa, we are evaluating the model's ability to predict the unseen text from the corpus and take this as an intrinsic evaluation metric of the quality of the two models.[3]

Other intrinsic metrics used to evaluate RoBERTa and SurgicBERTa on the surgical domain in this paper are the accuracy of MLM computed on the masked tokens during the evaluation step and the evaluation loss. Accuracy measures how well our model predicts the masked words by comparing the model predictions with the proper values in terms of percentage. Instead, the loss is a value that represents the summation of errors in a model. It measures how well or badly the model is performing. If the errors are high, the loss will be high, and then the model will not perform well.

Generally, the higher the accuracy in the evaluation dataset and the lower the evaluation loss, the better the model will perform.

[3] Our comparison is fair in that RoBERTa and SurgicBERTa share the same tokenizer and the same vocabulary.

Table 1 Perplexity, accuracy, and evaluation loss

Pre-trained model	Perplexity	Accuracy	Evaluation loss
RoBERTa	15.410	0.546	2.735
SurgicBERTa	**4.300**	**0.699**	**1.458**

Bold values mark the better scores for each metric

4.1.2 Results and discussion

Table 1 reports perplexity, accuracy, and loss values of RoBERTa and SurgicBERTa obtained during the evaluation of the MLM tasks as described in Sect. 4.1. SurgicBERTa has lower perplexity (−11.11), greater accuracy (+15.30%), and lower evaluation loss (−1.277) than RoBERTa. All obtained results intrinsically confirm that SurgicBERTa better deals with surgical language than RoBERTa.

4.2 Extrinsic evaluation: task A—procedural content detection

4.2.1 Task definition

The detection of procedural content consists of a binary classification task where the aim is to classify each sentence of a corpus into two different classes (*procedural* and *non-procedural*) This task is generally a preliminary and essential step for the business or robotic process automation starting from procedural content stored in textual materials because it allows models to deal with only those sentences that are important for the extraction of a workflow [26]. In the case of the surgical domain, the two classes are defined in [2]:

- *Procedural sentences* describe a specific action performed by either the robot or the human surgeon (e.g., an intervention on the body, the positioning of the robot). An example of a procedural sentence is "The colon is reflected medially over the kidney along the white line of Toldt.";
- *Non-procedural sentences* do not contain any indication of a specific surgeon action, but rather describe general, complementary information or anatomical features, not necessarily specific to perform a particular step of the intervention. An example of a non-procedural sentence is "This permits greater range of camera movement inferiorly within the retroperitoneum."

As training and testing material, we exploit the latest available version (v1.1) of the SPKS dataset,[4] containing 2250

[4] https://gitlab.com/altairLab/spks-dataset.

sentences manually annotated as procedural (approx. 68%) and non-procedural (approx. 32%).

In order to fine-tune RoBERTa and SurgicBERTa for procedural sentence classification, these pre-trained models have been extended to produce a classification output (procedural/non-procedural) by adding a softmax-activated classification layer on the pre-trained language models, and then by fine-tuning them on the SPKS dataset. A standard cross-entropy loss function has been adopted for classification. Due to the reduced size of the dataset, we utilized the classical 10-fold cross-validation protocol, which involves dividing the dataset into ten sets. In each iteration, one set is used for testing the classifier, while the remaining nine sets are used for training and hyperparameter tuning. This process is repeated ten times, and the classification performance is evaluated by computing the average of the evaluation metrics over the ten iterations.

Standard metrics for classification tasks, namely precision (P), recall (R), and F1-score, are used to compute performance. The metrics are calculated for each class (procedural/non-procedural) and we report for each of them the macro average, i.e., the mean of the considered metric on the two classes. In addition, we also compute *Accuracy* (Acc), i.e., the ratio between the correctly predicted classes, divided over the test set size, that in the case of binary classification, coincides with the micro average of P, R, and F1. For testing the statistical significance, we computed the *p value* applying the McNemar's test with significance threshold α of 0.05, as implemented in [10].

4.2.2 Results and discussion

Results of the procedural sentence detection task described in Sect. 4.2 have been reported in Table 2. SurgicBERTa improves all the performance metrics when compared to RoBERTa on both procedural and non-procedural classes. Overall, averaging the performances on both classes, SurgicBERTa improves the accuracy of 0.014, and Macro-F1 of 0.015, confirming the benefit of having a domain-specific language for surgical-related text classification. The observed performance difference of the two systems is statistically confirmed by the considered significance test.

4.3 Extrinsic evaluation: task B—procedural knowledge extraction

4.3.1 Task definition[5]

The purpose of this task is the extraction of procedural information from texts using semantic role labeling (SRL)

[5] This section summarizes findings and content previous presented in [4].

Table 2 Text classification performance of the tested methods (Extrinsic Evaluation—Task A)

Model	Procedural			Non-procedural				Macro		
	P	R	F1	P	R	F1	Acc	P	R	F1
RoBERTa	0.889 (0.019)	0.928 (0.063)	0.908 (0.039)	0.831 (0.053)	0.753 (0.017)	0.790 (0.029)	0.872 (0.021)	0.860 (0.029)	0.841 (0.021)	0.849 (0.022)
SurgicBERTa	**0.894** (0.018)	**0.945** (0.032)	**0.919** (0.016)	**0.865** (0.047)	**0.762** (0.010)	**0.810** (0.028)	**0.886** (0.015)	**0.880** (0.018)	**0.853** (0.022)	**0.864** (0.019)

The best scores are highlighted in bold. The standard deviation between the various folds is reported in brackets

techniques. Given a sentence, the SRL task aims at labeling the semantic arguments of the sentence predicates in order to extract *Who* does *What* to *Whom*, *How*, *When*, and *Where*. In this paper, we adopt the PropBank [23] approach for SRL, leveraging the catalog of semantic roles and predicate meanings codified in the Robotic-Surgery Propositional Framebank (RSPF) [3].

SRL can be organized in two complementary subtasks: (i) *predicate disambiguation*, i.e., the understanding of the correct meaning of a word describing an action (a.k.a., a predicate), and (ii) *semantic arguments identification and classification*, i.e., the detection of the argument spans of a predicate, and the assignment of them to the correct semantic role labels from RSPF. For example, given the sentence:

> *The colon is reflected medially over the kidney along the white line of Toldt.*

with task (i) the method should recognize that *reflect* has in this context the RSPF's meaning of *reflect.03*, i.e., *to bend or fold back*, and not for example the RSPF's meaning of *reflect.02*, i.e., *think about* or *reflect.01*, i.e., *cast an image back, casting back an image*. Then, given this meaning, the method has to solve the task (ii), i.e., to tokenize and classify the arguments in the sentence as follows:

> *[Arg.1: The colon] is [V: reflected] [Arg.2: over the kidney] [Arg.3: along the white line of Toldt].*

where *Arg.1*, *Arg.2* and *Arg.3* indicate (a) the *thing reflected*, (b) its *location*, and (c) *other spatial useful indications*, respectively.

Modern SRL methods rely on neural architectures that require annotated data to learn the language in a supervised way [11, 17]. To train, validate, and test the models, we used two different manually annotated textual datasets for semantic role labeling: CoNLL-2012 [25] and a smaller dataset specific to robotic surgery [5]. CoNLL-2012 is a large-scale general-English corpus with 318 k annotated predicates, covering multiple genres. We used this dataset to teach the common neural architecture the basic knowledge about the SRL task. The smaller dataset is instead domain-specific, containing 1559 SRL-annotated sentences regarding robotic surgery procedures, thus including both traditional surgical actions and specific robot operations. We used this smaller dataset to specialize the models, helping them to better understand surgical language and perform the SRL task more effectively in the given domain. The train, test, and validation splits already provided with the smaller dataset are used for the training, tuning, and evaluation of the performances. Specifically, 80% of the sentences are utilized for training (with 10% of them being set aside for validation), while the remaining 20% are dedicated to the test dataset. Moreover, for comparing the two language models on this task, the same metrics adopted for the procedural content detection task are used (cf. Sect. 4.2). For testing the statistical significance, we applied the Bootstrap test on the accuracy of the label (predicates and arguments) predictions with significance threshold α of 0.05 and using the implementation of [10].

Table 3 Performance (overall) on the SRL task (Extrinsic Evaluation—Task B). The best scores are highlighted in bold

Pre-trained model	Predicates Accuracy	Arguments Precision	Recall	F1
RoBERTa	0.907	0.771	0.752	0.762
SurgicBERTa	**0.925**	**0.778**	**0.768**	**0.773**

4.3.2 Results and discussion

Table 3 reports the performance of the procedural knowledge extraction task described in Sect. 4.3. SurgicBERTa substantially improves the predicated disambiguation task accuracy of 0.018 when compared to RoBERTa. Moreover, SurgicBERTa outperforms RoBERTa in all evaluation metrics related to the arguments disambiguation task. In particular, it improves the precision of 0.007, recall of 0.016, and F1 of 0.011. The improvement is confirmed to be statistically significant by the performed Bootstrap test. These results extrinsically demonstrate the benefit of having specialized RoBERTa in the surgical domain for the accurate extraction of actions and related information from surgical text.[6]

4.4 Extrinsic evaluation: task C—ontological information about the surgery and anatomical target

4.4.1 Task definition

The purpose of this task is to associate the name of the surgical procedure with the corresponding anatomical target or relevant feature to verify if the language models have learned this type of knowledge during training. For example, the *prostatectomy* has to be associated with *prostate*, *nephrectomy* with *kidney*, and *mastectomy* with *breast*. To evaluate our models on this task, we built a dataset consisting of the definition of 20 different surgical procedures. In particular, surgical procedures that can be performed with the aid of a robot have been chosen, together with other very frequent laparoscopic ones. The definitions are retrieved from the web or surgical manuals not used during the training of the language models. From them, the name of the corresponding

[6] A more fine-grained assessment of the application of SurgicBERTa for SRL on the surgical domain is provided in [4], where different complementary analyses and comparisons (e.g., zero-shot learning, few-shot learning) are performed.

anatomical target has been removed, and the models are asked to guess it. As evaluation metrics, we consider the ranking of the correct target word with respect to the others returned by the model, the reciprocal rank (RR), and the mean reciprocal rank (MRR) [35]. We have chosen these metrics and not others primarily with "accuracy" because we have a finite list of candidates in output that we want to be able to scroll through. MRR is a metric used to assess the performance of systems that provide a ranked list of answers in response to user queries. In the case of this task, answers are words returned to fill the $\langle mask \rangle$, i.e., the anatomical part corresponding to the procedure description, and queries are the sentences describing the procedure. In more detail, for a single query, the RR is defined as $\frac{1}{\text{rank}}$, where $rank$ is the position of the correct answer among the ones (sorted by probability, from the highest to the lowest) predicted by the model. For multiple queries $|Q|$, the MRR is the mean of the $|Q|$ RRs, i.e.,

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\text{RR}_i \tag{1}$$

The vocabulary has not been restricted, i.e., a list of possible candidates to choose from has not been used so that models can return any word belonging to the vocabulary.

To better clarify with an example, consider the following sentence (i.e., query):

a sacrocolpopexy is a surgical procedure used to treat $\langle mask \rangle$ *organ prolapse.*

Models are asked to fill in the missing word with the correct one which in the above example is *pelvic*. They will propose a list of possible candidates sorted by probability. For example, for the above sentence, RoBERTa and SurgicBERTa return the correct word *pelvic* in the third and first position, thus obtaining an RR of 0.33 and 1.0 with a log likelihood probability of 0.043 and 1.0 respectively. For testing the statistical significance, we applied the Bootstrap test on the RRs of the corrected predictions, using the same α threshold and implementation of the other tasks.

4.4.2 Results and discussion

This section summarizes the results of the above-described task, i.e., that of predicting the anatomical target given the name and a brief definition of the surgical intervention related to that anatomical target. On average, the correct target is returned by RoBERTa in position 2.35, while SurgicBERTa outperforms RoBERTa proposing the correct target in position 1.35. The MRR of RoBERTa is 0.731, while that of SurgicBERTa is 0.902. In more detail, 30% of the times SurgicBERTa performs better than RoBERTa in terms of RR. The base model performs better than SurgicBERTa only in one case (query 19), where the model is asked to predict the anatomical part related to the "endarterectomy," that is the "artery." Since RoBERTa performs only slightly better than SurgicBERTa (the first returns "artery" in 4th position while the latter in 5th), this is perhaps due to the fact that the base model may have already seen a similar sentence (or documents describing the "endarterectomy") during its training phase. The violin plots of Fig. 2 summarize the obtained RRs on each query sentence: the one for SurgicBERTa is very wide at the top and skinny in the middle and at the bottom, while the one of RoBERTa, albeit having a similar distribution, is much less wide at the top and has a median weight lower than that of SurgicBERTa. The shape of the distribution indicates that the RRs of SurgicBERTa are highly concentrated around the first quartile, meaning that the model is predicting very well the proper anatomical target very well. In contrast, the RRs of RoBERTa are more evenly distributed across the entire range, highlighting lower scores. The computed p *value* (< 0.05) confirms the statistical significance of the observed performance difference, and thus the benefit of having specialized RoBERTa for the surgical language.

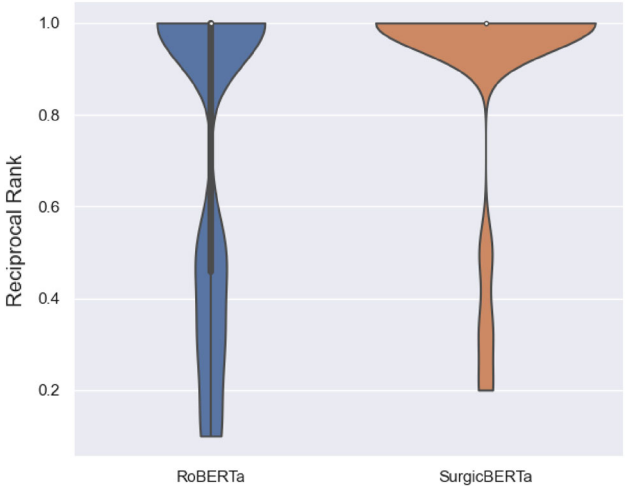

Fig. 2 Reciprocal rank of the predicted word in the task of predicting the anatomical target given the information of a surgical procedure (Extrinsic Evaluation—Task C)

4.5 Extrinsic evaluation: task D—surgical terminology acquisition

4.5.1 Task definition

This task is the same as the previous one but applied to a different dataset and therefore proposed for a different purpose: to verify whether SurgicBERTa masters the surgical language and can use it more appropriately than RoBERTa.

In particular, a dataset of 50 surgical sentences was collected from different sources, i.e., surgical books, academic papers, and web pages not used during the MLM training. The sentences were randomly chosen from those that met the following requirements:

- The sentence has not been used to train SURGICBERTA;
- One of the following holds:
  - The sentence contains an expression commonly used in surgery. To define widely used expressions, we have selected those typically abbreviated with an acronym in papers. In the sentences included in the dataset, the abbreviations have been substituted with the original expression, and the language models are asked to complete them correctly in the corresponding context;
  - The sentence contains a description of a surgical procedure. In the sentences inserted in the dataset, the verb describing the action is masked, and the language model is asked to guess it based on the context.

Since the task is the same as the previous one, we used the same metrics adopted for it, i.e., the position in which the correct solution is proposed, the RR and the MRR. We also applied the same statistical significance test.

4.5.2 Results and discussion

Table 4 summarizes the obtained results for the task described in Sect. 4.5. SURGICBERTA substantially improves all proposed metrics: the mean position at which the word filling correctly the masks is proposed by SURGICBERTA among the list of returned ones is 19.19 times better than the ROBERTA one. This means SURGICBERTA is much more familiar with surgical terminology than ROBERTA. Consequently, the MRR is improved by 0.396. 66% of the times SURGICBERTA improves the RRs when compared to ROBERTA. Only in two cases (out of 50) ROBERTA performs better than SURGICBERTA: similarly to task C, it is difficult to understand why this happens, and the same considerations may apply. The violin plots of Fig. 3 illustrate the RRs of the two language models for each query: while the one for SURGICBERTA is wide at the top, the one for ROBERTA is wide at the bottom. Furthermore, SURGICBERTA has a median weight much higher than that of ROBERTA. This highlights the best accuracy of SURGICBERTA in managing surgical terminology, also confirmed by the significance test performed (p value < 0.05). Hence, also this task confirms that SURGICBERTA better captures the surgical language.

Table 4 Mean position and MRR on the task of surgical terminology acquisition (Extrinsic Evaluation—Task D)

Pre-trained model	Mean position	MRR
ROBERTA	152.720	0.262
SURGICBERTA	**7.960**	**0.658**

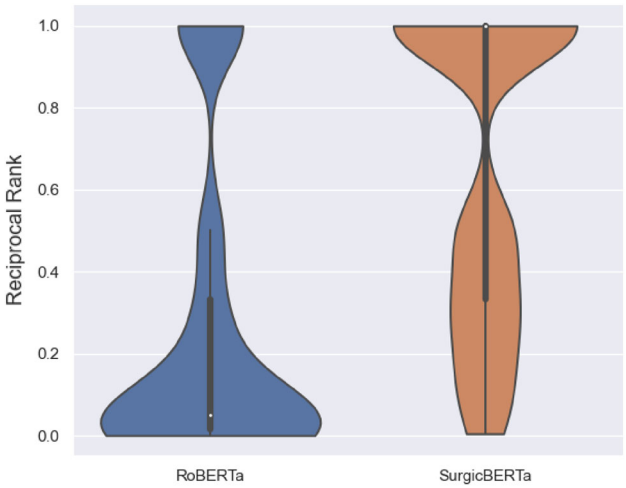

Fig. 3 Reciprocal rank of the predicted word in the task of surgical terminology acquisition (Extrinsic Evaluation—Task D)

4.6 Qualitative examples of surgical knowledge available in pre-trained language models

There is a lot of domain information implicit in pre-trained language models [24]. Adapting the domain through continual learning with MLM helps to capture this kind of knowledge. However, it is complicated to quantify this domain knowledge objectively and exhaustively due to the lack of any gold standard for the surgical domain. For this reason, this section proposes a qualitative analysis, providing examples of domain information stored in pre-trained language models.

To start with, ROBERTA and SURGICBERTA are asked to return the name of the most used surgical robot in the operating room. In particular, ROBERTA and SURGICBERTA are asked to substitute the $\langle mask \rangle$ in the following sentence with the most appropriate five words, ranking them in order of probability:

The most commonly used surgical robot is $\langle mask \rangle$.

Results are reported in Table 5. While to the best of our knowledge, none of the top five words returned by ROBERTA

Table 5 RoBERTa and SurgicBERTa most probable words for the most used surgical robots

Rank	RoBERTa Word	Probability	SurgicBERTa Word	Probability
1	Braun	0.031	Zeus	0.261
2	Juno	0.027	Xi	0.111
3	Hawk	0.017	Si	0.055
4	Orion	0.016	robotic	0.035
5	MRI	0.016	S	0.030

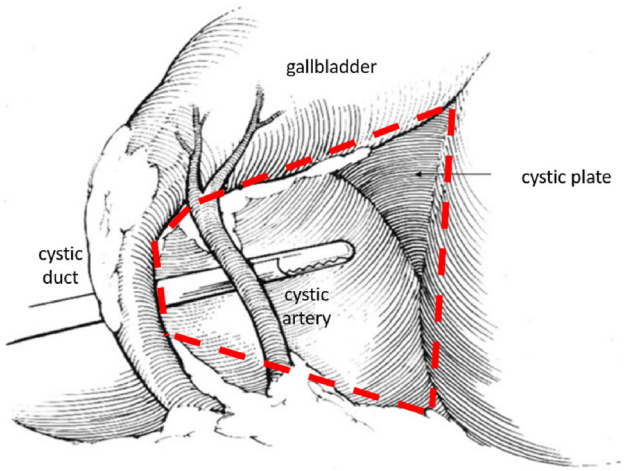

Fig. 4 Illustration of the critical view of safety method during a cholecystectomy

is the name of a surgical robot, *Zeus*,[7] *Xi*,[8] and *Si*[9] returned by SurgicBERTa are instead examples of surgical robots that have been used in operating theaters. This means that the continual MLM learning with domain text has captured this kind of information that now is available in the model. Nonetheless, it is interesting to note how some of the words returned by RoBERTa are sometimes related to the robotics field: "Hawk," "Orion," "Juno" are also examples of (non-surgical) robots. This observation may suggest that while the general model tries to be correct, it lacks specific domain knowledge.

As reported in Table 1, SurgicBERTa has a perplexity substantially lower than RoBERTa in the MLM task when applied to surgical literature. This intrinsically means that SurgicBERTa has learned the surgical language and thus also the composition of well-known surgical expressions. Consider the following example highlights how SurgicBERTa has learned specialized domain terminology. In surgery, the expression *critical view of safety* refers to a

[7] https://en.wikipedia.org/wiki/ZEUS_robotic_surgical_system.

[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6193435/.

[9] https://www.davincisurgerycommunity.com/Systems_I_A/da_Vinci_Si_Si_e.

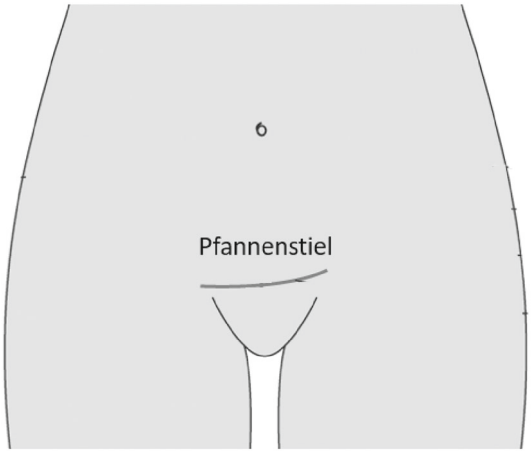

Fig. 5 Pfannenstiel incision to access the abdomen. This figure is adapted from [13]

method of secure identification in open cholecystectomy in which the cystic duct and artery are putatively identified, after which the gallbladder is taken off the cystic plate so that the gallbladder is attached only by the two cystic structures [32] as shown by Fig. 4.

To verify if RoBERTa and SurgicBERTa know this information, they are asked to complete the following sentence:

> *During cholecystectomy, it is important to achieve the critical view of* $\langle mask \rangle$.

SurgicBERTa returns the word *safety* as $1st$ result with a probability of 0.3428, while RoBERTa returns it only at $47th$ position with the probability of 0.0032.

This section ends with another example of domain knowledge available in SurgicBERTa. In surgery, a *Pfannenstiel incision* is a type of surgical incision that allows access to the abdomen (Fig. 5). The following test wants to investigate if pre-trained language models know this information:

> *The Pfannenstiel is a type of surgical incision that allows access to the* $\langle mask \rangle$.

The correct word is *abdomen* and is retrieved by SurgicBERTa at the $1st$ position with probability 0.1267 and by RoBERTa at the $5th$ position with probability 0.0478, after the words *brain* (0.1969), *heart* (0.1488), *skin* (0.0713), and *vagina* (0.0542).

These qualitative examples show that in SurgicBERTa there is a lot of surgical information that could be used, for instance, to enrich and complement the one codified in domain ontologies and knowledge bases.

Nevertheless, since the model was fine-tuned on the MLM task on surgical domain texts, SurgicBERTa could also suffer from the problems that the models thus generated typically have. Among all, we underline the frequent risk of

introducing bias into the models which in the case of a surgical model could be that of making predictions of words always considering a standard human anatomy, ignoring all possible particular cases. Also, SURGICBERTA was obtained by specializing ROBERTA on the surgical case, so some of the known biases of the latter are likely to be replicated on SURGICBERTA as well. All of these problems can be reduced by choosing better training materials or adapting de-biasing techniques to the domain. Furthermore, the relevance of the returned word could be low in domains not seen (enough) during the training: using reinforcement learning with human feedback techniques [22] could help to reduce these problems.

5 Conclusions

This paper proposed SURGICBERTA, a pre-trained language fine-tuned for capturing surgical language and knowledge, i.e., the vocabulary and expertise provided in surgical books and academic papers.

The building process has been described, and the model has been evaluated both intrinsically, by considering perplexity, accuracy, and evaluation loss during the MLM task, and extrinsically, by considering several downstream tasks, namely (i) procedural sentences detection, (ii) procedural knowledge extraction, (iii) ontological information discovery, and (iv) surgical terminology learning. All the results confirm that SURGICBERTA deals with surgical language and knowledge more adequately than ROBERTA, a language model targeting general-domain English. Moreover, the potential of SURGICBERTA has been investigated qualitatively by showing several examples of surgical domain knowledge available in the model, which could be used to complement other knowledge sources, e.g., state-of-the-art surgical knowledge bases. As future works, we will enrich SURGICBERTA by continuously training it on a larger surgical dataset and extending it in a multilingual scenario.

Acknowledgements This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 742671 "ARS").

Author Contributions "MB, MR, SPP, and PF contributed in the following way:—Conceptualization: MB, MR, SPP—Methodology: MB, MR, SPP—Supervision: MR, SPP, PF—Funding acquisition: PF—Writing—original draft: MB–Writing—review and editing: MB, MR, SPP, PF.

Funding Open access funding provided by Universitá degli Studi di Verona within the CRUI-CARE Agreement.

Data availability The SURGICBERTA language model together with the resources used in our experiments is available under an open license at https://gitlab.com/altairLab/surgicberta.

Declarations

Conflict of interest Marco Bombieri, Marco Rospocher, Simone Paolo Ponzetto, and Paolo Fiorini declare that they do not have conflicts of interest.

References

1. Bear Don't Walk, I.V.O.J., Sun, T., Perotte, A., et al.: Clinically relevant pretraining is all you need. J. Am. Med. Inform. Assoc. **28**(9), 1970–1976 (2021)
2. Bombieri, M., Rospocher, M., Dall'Alba, D., et al.: Automatic detection of procedural knowledge in robotic-assisted surgical texts. Int. J. Comput. Assist. Radiol. Surg. **16**(8), 1287–1295 (2021). https://doi.org/10.1007/s11548-021-02370-9
3. Bombieri, M., Rospocher, M., Ponzetto, S.P., et al.: The robotic surgery procedural framebank. In: Proceedings of the Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 3950–3959 (2022). https://aclanthology.org/2022.lrec-1.420
4. Bombieri, M., Rospocher, M., Ponzetto, S.P., et al.: Machine understanding surgical actions from intervention procedure textbooks. Comput. Biol. Med. (2023). https://doi.org/10.1016/j.compbiomed.2022.106415
5. Bombieri, M., Rospocher, M., Ponzetto, S.P., et al.: The robotic-surgery propositional bank. Lang. Resour. Evaluation (2023). https://doi.org/10.1007/s10579-023-09668-x
6. Carrino, C.P., Llop, J., Pàmies, M., et al.: Pretrained biomedical language models for clinical NLP in Spanish. In: Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics, Dublin, Ireland, pp. 193–199 (2022). https://doi.org/10.18653/v1/2022.bionlp-1.19
7. Chandak, S., Zhang, L., Brown, C., et al.: Towards automatic curation of antibiotic resistance genes via statement extraction from scientific papers: A benchmark dataset and models. In: Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics, Dublin, Ireland (2022)
8. Devlin, J., Chang, M., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers). Association for Computational

Linguistics, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/n19-1423,
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16 x 16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net (2021). https://openreview.net/forum?id=YicbFdNTTy
10. Dror, R., Baumer, G., Shlomov, S., et al.: The hitchhiker's guide to testing statistical significance in natural language processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 1383–1392 (2018) https://doi.org/10.18653/v1/P18-1128
11. He, L., Lee, K., Lewis, M., et al.: Deep semantic role labeling: What works and what's next. In: Barzilay, R., Kan, M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. Association for Computational Linguistics, pp. 473–483 (2017). https://doi.org/10.18653/v1/P17-1044
12. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science **349**(6245), 261–266 (2015). https://doi.org/10.1126/science.aaa8685
13. Jeelani, K.: Surgical Anatomy of the Female Pelvis and Abdominal Wall, pp. 8–14. Cambridge University Press, Cambridge (2020). https://doi.org/10.1017/9781108644396.002
14. Johnson, A., Pollard, T., Shen, L., et al.: Mimic-iii, a freely accessible critical care database. Sci. Data **3**(160), 035 (2016). https://doi.org/10.1038/sdata.2016.35
15. Lauscher, A., Ravishankar, V., Vulić, I., et al.: From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers (2020). https://doi.org/10.48550/ARXIV.2005.00633
16. Lewis, M., Liu, Y., Goyal, N., et al.: Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019). arXiv preprint arXiv:1910.13461
17. Li, T., Jawale, P.A., Palmer, M., et al.: Structured tuning for semantic role labeling. In: Jurafsky, D., Chai, J., Schluter, N., et al. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020. Association for Computational Linguistics, pp. 8402–8412 (2020) https://doi.org/10.18653/v1/2020.acl-main.744
18. Liang, Z., Noriega-Atala, E., Morrison, C., et al.: Low resource causal event detection from biomedical literature. In: Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics, Dublin, Ireland (2022)
19. Lin, C., Zheng, S., Liu, Z., et al.: SGT: scene graph-guided transformer for surgical report generation. In: Wang, L., Dou, Q., Fletcher, P.T., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII, Lecture Notes in Computer Science, vol. 13437, pp. 507–518. Springer (2022) https://doi.org/10.1007/978-3-031-16449-1_48
20. Liu, Y., Ott, M., Goyal, N., et al.: Roberta: A robustly optimized BERT pretraining approach. CoRR. arXiv: org/abs/1907.11692, (2019)
21. Locke, S., Bashall, A., Al-Adely, S., et al.: Natural language processing in medicine: a review. Trends Anaesth. Crit. Care **38**, 4–9 (2021)
22. Osborne, P., Nõmm, H., Freitas, A.: A survey of text games for reinforcement learning informed by natural language. Trans. Assoc. Comput. Linguistics **10**, 873–887 (2022)
23. Palmer, M., Kingsbury, P.R., Gildea, D.: The proposition bank: an annotated corpus of semantic roles. Comput. Linguistics **31**(1), 71–106 (2005). https://doi.org/10.1162/0891201053630264
24. Petroni, F., Rocktäschel, T., Riedel, S., et al.: Language models as knowledge bases? In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 2463–2473 (2019). https://doi.org/10.18653/v1/D19-1250
25. Pradhan, S., Moschitti, A., Xue, N., et al.: Towards robust linguistic analysis using ontonotes. In: Hockenmaier, J., Riedel, S. (eds.) Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8–9, 2013. ACL, pp. 143–152 (2013). https://aclanthology.org/W13-3516/
26. Qian, C., Wen, L., Kumar, A., et al.: An approach for process model extraction by multi-grained text classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12127 LNCS:268–282 (2020)
27. Salazar, J., Liang, D., Nguyen, T.Q., et al.: Masked language model scoring. In: Jurafsky, D., Chai, J., Schluter, N., et al. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020. Association for Computational Linguistics, pp. 2699–2712 (2020) https://doi.org/10.18653/v1/2020.acl-main.240
28. Seenivasan, L., Islam, M., Krishna, A.K., et al.: Surgical-vqa: visual question answering in surgical scenes using transformer. In: Wang, L., Dou, Q., Fletcher, P.T., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VII, Lecture Notes in Computer Science, vol. 13437, pp. 33–43 . Springer (2022). https://doi.org/10.1007/978-3-031-16449-1_4
29. Shin, J., Lee, Y., Jung, K.: Effective sentence scoring method using bert for speech recognition. In: Lee, W.S., Suzuki, T. (eds.) Proceedings of The Eleventh Asian Conference on Machine Learning. Proceedings of Machine Learning Research, PMLR, vol. 101, pp. 1081–1093 (2019). https://proceedings.mlr.press/v101/shin19a.html
30. Singhal, K., Azizi, S., Tu, T., et al.: (2022) Large language models encode clinical knowledge. https://doi.org/10.48550/ARXIV.2212.13138
31. Sohn, S., Wang, Y., Wi, C.I., et al.: Clinical documentation variations and NLP system portability: a case study in asthma birth cohorts across institutions. J. Am. Med. Inform. Assoc. **25**(3), 353–359 (2017)
32. Strasberg, S., Hertl, M., Soper, N.: An analysis of the problem of biliary injury during laparoscopic cholecystectomy. Surg. Gynecol. Obstet. **180**(1), 101–125 (1995)
33. Taylor, W.L.: Cloze procedure: a new tool for measuring readability. J. Q. **30**(4), 415–433 (1953). https://doi.org/10.1177/107769905303000401
34. Vaswani,A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp. 5998–6008 (2017) https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
35. Voorhees, E.M.: The TREC-8 question answering track report. In: Voorhees, E.M., Harman, D.K. (eds.) Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17–19, 1999, NIST Special Publication, vol. 500–246. National Institute of Standards and Technology (NIST), http://trec.nist.gov/pubs/trec8/papers/qa_report.pdf (1999)
36. Wang, A., Cho, K.: BERT has a mouth, and it must speak: BERT as a Markov random field language model. In: Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation. Association for Computational Linguis-

tics, Minneapolis, Minnesota, pp. 30–36 (2019). https://doi.org/10.18653/v1/W19-2304
37. Xie, K., Gallagher, R.S., Conrad, E.C., et al.: Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing. J. Am. Med. Inform. Assoc. **29**(5), 873–881 (2022)
38. Xu, M., Islam, M., Lim, C.M., et al.: Class-incremental domain adaptation with smoothing and calibration for surgical report generation. In: de Bruijne, M., Cattin, P.C., Cotin, S., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part IV, Lecture Notes in Computer Science, vol. 12904, pp. 269–278. Springer (2021a). https://doi.org/10.1007/978-3-030-87202-1_26
39. Xu, M., Islam, M., Ming Lim, C., et al.: Learning domain adaptation with model calibration for surgical report generation in robotic surgery. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 12350–12356 (2021b). https://doi.org/10.1109/ICRA48506.2021.9561569
40. Xu, M., Islam, M., Ren, H.: Rethinking surgical captioning: End-to-end window-based MLP transformer using patches. In: Wang, L., Dou, Q., Fletcher, P.T., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII, Lecture Notes in Computer Science, vol. 13437, pp. 376–386. Springer (2022). https://doi.org/10.1007/978-3-031-16449-1_36
41. Yang, X., Bian, J., Hogan, W.R., et al.: Clinical concept extraction using transformers. J. Am. Med. Inform. Assoc. **27**(12), 1935–1942 (2020)
42. Yang, X., Chen, A., PourNejatian, N., et al.: A large language model for electronic health records. npj Digit. Med. **5**(1), 194 (2022). https://doi.org/10.1038/s41746-022-00742-2
43. Yao, L., Jin, Z., Mao, C., et al.: Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J. Am. Med. Inform. Assoc. **26**(12), 1632–1636 (2019)
44. Zhou, S., Wang, N., Wang, L., et al.: CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. J. Am. Med. Inform. Assoc. (2022)