Artificial Intelligence In Medicine 108 (2020) 101949

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Position Paper

EBM+: Advancing Evidence-Based Medicine via two level automatic identification of Populations, Interventions, Outcomes in medical literature

Nikolaos Stylianou [a,*], Gerasimos Razis [c], Dimitrios G. Goulis [b], Ioannis Vlahavas [a]

[a] School of Informatics, Aristotle University of Thessaloniki, Greece
[b] School of Medicine, Aristotle University of Thessaloniki, Greece
[c] Atypon Systems, Inc., United States

ARTICLE INFO

Keywords:
Evidence Based Medicine
PICO
Machine learning
Neural networks
Natural Language Processing

ABSTRACT

Evidence-Based Medicine (EBM) has been an important practice for medical practitioners. However, as the number of medical publications increases dramatically, it is becoming extremely difficult for medical experts to review all the contents available and make an informative treatment plan for their patients. A variety of frameworks, including the PICO framework which is named after its elements (Population, Intervention, Comparison, Outcome), have been developed to enable fine-grained searches, as the first step to faster decision making.

In this work, we propose a novel entity recognition system that identifies PICO entities within medical publications and achieves state-of-the-art performance in the task. This is achieved by the combination of four 2D Convolutional Neural Networks (CNNs) for character feature extraction, and a Highway Residual connection to facilitate deep Neural Network architectures. We further introduce a PICO Statement classifier, that identifies sentences that not only contain all PICO entities but also answer questions stated in PICO. To facilitate this task we also introduce a high quality dataset, manually annotated by medical practitioners. With the combination of our proposed PICO Entity Recognizer and PICO Statement classifier we aim to advance EBM and enable its faster and more accurate practice.

1. Introduction

Evidence-Based Medicine (EBM) enables medical professionals to apply treatments based on the complete available evidence. Evidence is considered to be all previously performed research, usually in the forms of Randomized Control Trials (RCTs) or Clinical Trials (CTs) that investigate the effects of a treatment on a specific group of patients and present the outcomes. Other forms of studies, such as systematic reviews, can also be considered as evidence when they present findings of treatments for a specific target group.

However, as the number of medical publications has increased dramatically, it is becoming more challenging for health and medical practitioners to identify the best treatment [1]. Specifically, PubMed alone had over 1 million new publications on a yearly basis during the last 4 years.[1] What is more, recent research suggests that there is an infrastructure gap to support EBM practice [2].

RCTs and CTs usually set the research questions in the form of PICO framework. The framework is named after the elements that comprise it, which are Population, Intervention, Comparison and Outcome respectively. While the framework enables refined queries [3], the results are massive and reviewing the contents is a time-consuming process [4]. The PICO framework is not the only one to support EBM [5,6], but is the one predominantly used to this day by medical practitioners.

Research on advancing EBM with Artificial Intelligence (AI) has been slow, due to the lack of high quality and substantial size benchmark datasets, as well as the instability in the frameworks that have been used. Furthermore, while significant strides have been made, researchers confront the task in a variety of approaches, ranging from sentence classification and entity recognition [7,8] to question answering and information retrieval [9,10]. What is more, advancements in entity recognition could support sentence classification and question answering approaches. The publication of a high quality corpus

* Corresponding author at: School of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece.
E-mail address: nstylia@csd.auth.gr (N. Stylianou).
URL: http://www.atypon.com (G. Razis).
[1] https://www.ncbi.nlm.nih.gov/pubmed/.

https://doi.org/10.1016/j.artmed.2020.101949
Received 18 February 2020; Received in revised form 11 August 2020; Accepted 12 August 2020
Available online 13 August 2020
0933-3657/

with PICO entities facilitates the ability to use PICO entities and sets a benchmark for the entity recognition task.

1.1. Related work

Evidence-Based Medicine is of great importance to patient care and as a result has been tackled in a variety of ways. During the many years of EBM practice, a variety of frameworks have been proposed to replace PICO, like PICOS and SPIDER [11,12], as well as more complicated ones like PIBOSO [5]. However, a comparative study in terms of sensitivity and generalization [6] concludes that PICO should be used due to its simple and direct definition of the clinical questions.

From a Natural Language Processing (NLP) standpoint, a plethora of approaches have also been deployed. Early research has focused on the extraction of evidence from medical publications using document structures and headings [13,14]. These extracted information were, in succeeding research, used to identify sentences that describe PICO elements [7,15,16].

More detailed extraction systems have also been proposed, such as ExaCT [17] which applies rules to extract clinical aspects and ACRES [18] which summarizes findings from clinical trials. Alamri and Stevenson [19] describe a Question Answering system that identifies contradiction in similar studies.

All the aforementioned approaches are based on small, hand-annotated datasets, some of which are not publicly available. To tackle this, methods that use automatically created annotations have been utilized in information retrieval [20,14] and question answering tasks [10]. A novel approach for sentence classification of PICO elements is proposed in Wallace et al. [21], which uses a small amount of manually annotated data to create a large training corpora using distant supervision.

The focus or previous research has been on identifying sentences or paragraphs that contain clinical evidence, without explicitly identifying the individual PICO entities. Moreover, the PICO framework is not used by all preceding research. Due to the lack of entity level annotation data, Ferracane et al. [22] are identifying medical entities in text, without explicitly annotating them to entity types, by exploiting coreference in text. This is solved with the publication of a 5000-abstracts dataset with two-level hierarchical entity annotation [23]. The entities are first mapped to PICO terms, which are expanded to more detailed labels depending on the PICO term (four different labels for Populations, seven for Interventions and six for Outcome). The authors have also published an approach to use syntactic patterns to improve performance [24], as well as a strong baseline for a Named Entity Recognition (NER) task on both levels. The baseline, as described in [23], is a combination of a Bidirectional Long Short-Term Memory (BiLSTM) Network and a Conditional Random Field (CRF) layer. The baseline model uses both pre-trained word embeddings and word representations that are created with character embeddings passed through a Convolutional Neural Network (CNN) and a Long Short-Term Memory Network (LSTM) for text representation.

Due to the difficulty and lack of the required qualitative annotation data, evidence extraction has been ignored by some tasks, moving to higher level extraction tasks. These include question answering systems [25,26,9,10], summarization systems [18] and bias measuring systems [27,28]. However, these tasks are directly related to the evidence extraction ones, the findings of which can be used to increase their performance.

1.2. Our approach

In this research we propose a neural network approach which achieves state-of-the-art results in the task of PICO Entity Recognition. We showcase an overall performance increase of 6% over the baseline, with improvements up to 11% in recall in some cases — without sacrificing precision. We further expand our scope by identifying PICO Statements, i.e. sentences that not only contain all the PICO elements but also answer clinical questions. To facilitate our PICO statement sentence classification we also introduce a publicly available, expert annotated, corpus. Additionally, we explicitly identify the improvement over the sentence classification to PICO Statements, using PICO entities identified from our PICO Entity Recognition system.

2. Materials and methods

Our contribution is two fold. We firstly introduce a state-of-the-art PICO Entity Recognizer which improves the strong baseline, and secondly a PICO sentence classifier which identifies sentences that not only contain all the PICO elements but also provide answers to PICO questions, called PICO Statements. Additionally, we demonstrate that the performance of the PICO sentence classifier is greatly enhanced by the identification of the PICO entities.

2.1. PICO Entity Recognizer

The PICO Entity Recognizer is tackled as a sequence to sequence problem in which a label is assigned to each token of the input sentence. We approach the task with an end-to-end neural network with highway stacked Recursive Neural Networks (RNNs), using word and character embeddings as inputs and a CRF layer for the output. We opt for a CRF layer instead of Softmax as the task has inherited relations between neighbouring labels and it allows to jointly decode the best chain of labels for the input sequence given. The full architecture of our PICO Entity Recognizer is presented in Fig. 1.

Let $X = (x_1, x_2, \ldots, x_n)$ be the sequence of tokens where each token is made of a sequence of character $[c_1, \ldots, c_m]$, n is the number of tokens in the sequence and m is the number of characters in a token. We use deep contextualized representations, commonly referred to as ELMo embeddings [29], for the input sequence and CNN-based character-level representations of a span of characters based on [30].

From the pre-trained ELMo embeddings, we formulate a vector $V_x \in \mathbb{R}^{n_{\max} \times d_{\text{ELMo}}}$ containing the contextualized representations of the sequence X. We randomly initiliaze a trainable character embedding vector $C_x \in \mathbb{R}^{n_{\max} \times |m_{\max}| \times d_{\text{char}}}$, where d_{char} is the dimension of the character embeddings.

We introduce a 2D Convolution over the character embeddings, instead of the more common 1D approach [31], with four variable size filters and kernels. With the use of a 2D Convolution, we capture character information over the word sequences, as entities usually span over a series of tokens. In comparison, 1D approaches create word representations based on the characters of each word, while our approach additionally uses the characters of adjacent words to infuse context. As a result, we remove the need of an RNN layer to obtain meaningful representations for unknown words from the character embeddings and avoid issues such as vanishing and exploding gradient [32]. The resulting vector of each convolution is $C_{\text{xk}} \in \mathbb{R}^{d_{\text{char}} \times d_{\text{filter}} \times d_{\text{kernel}}}$, where d_{filter} and d_{kernel} are the filter and kernel size respectively. Similarly to Kim et al. [30], we use a non-linearity function and a max-over-time pooling operation to obtain a fixed-dimensional representation of the word in $C_{\text{xp}} \in \mathbb{R}^{d_{\text{char}} \times d_{\text{kernel}}}$. We concatenate the results, such that we get $C_x \in \mathbb{R}^{d_{\text{char}} \times d_{\text{filters}}}$, where $d_{filters}$ is the sum of all the filter sizes. Fig. 2 illustrates the use of 2D CNNs over a span of character based word representations. The concatenated vector is passed through a highway layer [33] that applies two non linear transformations on C_x based on a *transform gate* T and a *carry gate* C resulting in a similar shape vector C_{xH}.

Consequently, the input sequence is represented by a vector $\tilde{x}$, comprised by a token-wise concatenation of the contextualized word representation with the output of the highway CNNs, to be used in the following layers. As a result, at time-step t, the word representation will be:

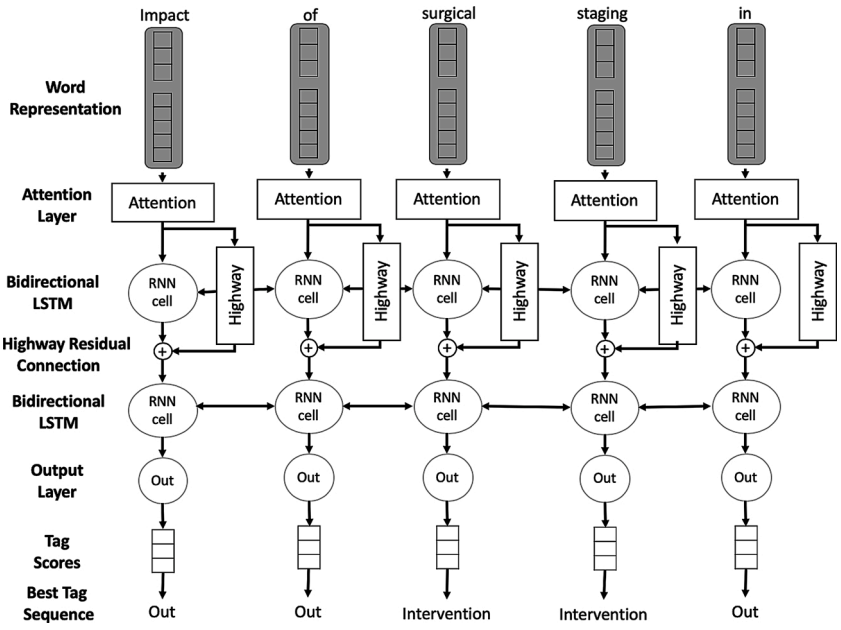

Fig. 1. PICO Entity Recognizer system architecture. "Out" is used to indicate that the word is not a PICO entity, while "Intervention" is use to classify the word in the Intervention entity type.

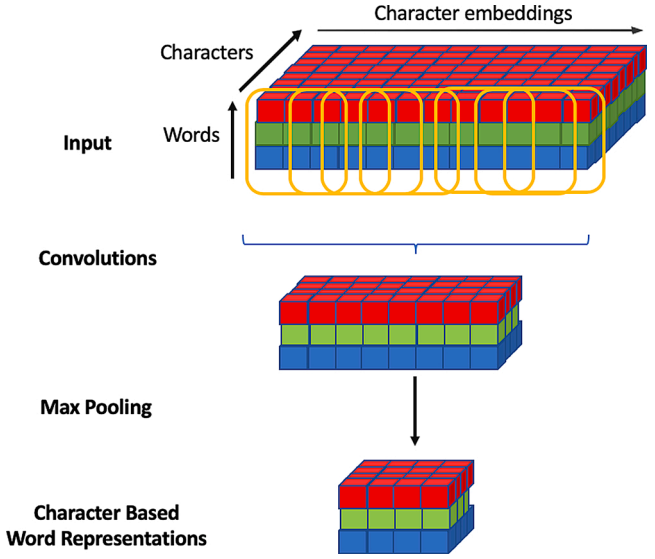

Fig. 2. 2D CNN architecture for character-based word representations.

$$\tilde{x_t} = [V_{x_t}; C_{\mathrm{xH}_t}] \tag{1}$$

We apply a self-attention mechanism over the input sequence, following the architecture proposed in Bahdanau et al. [34] over the input representation in $\tilde{x}$. The attention mechanism is used to enhance the ability of the model to extract features from the word representations vector. The resulting vector x is the input to the following RNN layers.

Our proposed model consists of two Bidirectional Long Short-Term Memory networks (BiLSTMs) over the representation vector, with a highway residual connection in between. Each LSTM cell is defined by the following formula:

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{2}$$

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{3}$$

$$\tilde{c}_t = tanh(W_c h_{t-1} + U_c x_t + b_c) \tag{4}$$

$$c_t = \mathrm{ft} \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{5}$$

$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o) \tag{6}$$

$$h_t = o_t \odot tanh(c_t) \tag{7}$$

where x_t is the input vector and h_t is the hidden state at time t. U_i, U_f, U_c, U_o denote the weight matrices of the different input gates, W_i, W_f, W_c, W_o are the weight matrices for the hidden state h_t and b_i, b_f, b_c, b_o are the bias vectors. The element-wise sigmoid function is denoted with σ and the element-wise product with $\odot$. The LSTM cells do not use a peephole connection. The final output vector of the BiLSTM is defined as the concatenation of the hidden states of a forward LSTM cell ($\overrightarrow{h_t}$) and a backward LSTM cell ($\overleftarrow{h_t}$) over the same sequence of inputs,

$$h_t^i = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{8}$$

resulting in a vector $h_t^i \in \mathbb{R}^{n_{\max} \times 2 \times d_{\mathrm{rnn}}}$ where d_{rnn} is the hidden size of each LSTM cell and i is the BiLSTM layer number.

The highway residual connection is implemented by applying a highway layer addition between the BiLSTMs as follows:

$$h_t^i = \mathrm{BiLSTM}_i(h_{t-1}^i, x_t^{i-1}, W^i) \tag{9}$$

$$x_t^i = h_t^i + z_t^i \tag{10}$$

$$h_t^{i+1} = \mathrm{BiLSTM}_{i+1}(h_{t-1}^{i+1}, x_t^i, W^{i+1}) \tag{11}$$

where z_t^i is the output of the vector x_t^{i-1} passing through a highway layer and i the BiLSTM layer number. As a result, we implement a highway connection every two RNN layers. The final vector h_x is passed through a projection layer to a CRF layer that handles the label prediction. The CRF layer uses the Viterbi algorithm to search for the chain of labels efficiently.

2.2. PICO Statement identification task

Intuitively, a statement that contains all the PICO entities gives an overview of the study. However, past approaches focus on sentences in

abstracts that describe a single PICO entity. Through the work presented thus far, it becomes obvious that sentences can contain more than one PICO elements. As a result, this information, while present was never annotated hence not allowing for a more advanced sentence identification system.

To bridge the gap, we created the "PICO Statements dataset", which identifies sentences that are PICO Statements based on two criteria; containing all of the PICO entities and, providing an answer to a PICO stated question. Each sentence is annotated to reflect on the PICO elements that it contains and whether it is a PICO Statement or not.

In order to emphasise the importance of identifying PICO Statements, both sentences (Statement A and B) in the following examples (Table 1), which are part of the "PICO Statements dataset", contain Population, Intervention/Comparator and Outcome entities. However, only Statement B provides an answer to a clinical stated question.

Subsequently, Statement B is more important than Statement A to a medical practitioner, and we strongly believe highlighting it in a study will be beneficial and time-saving. This task and the dataset described in Section 2.2.1, provide the foundation towards effective PICO Statement identification.

2.2.1. PICO Statements dataset

For our dataset annotation, we used a double blind standard approach as proposed in [35]. After consulting with medical experts, we engineered PICO queries to reflect six distinct categories of interest in modern medicine – Endocrinology, Gynecology, Neurology, Orthopedic Surgery, Pediatrics and Thoracic Surgery – and extracted abstracts by querying MEDLINE's PubMed Online PICO interface. More specifically, we identified distinct Populations for each category such as "Graves' disease" for Endocrinology, "Premature delivery" for Gynecology and "Multiple sclerosis" for Neurology, to be used towards document collection. The collected documents had to be published in the last five years, have only English contents and describe a CT or an RCT study. For each publication, we collected both the abstract and title.

Our annotators consisted of 26 MSc students from the medical department of the Aristotle University of Thessaloniki, who already held medical degrees and the annotation process was evaluated using Inter Annotator Agreement (IAA) [35] and Cohen's Kappa [36]. We assigned scores on each annotator based on their IAA, and selected the annotations from the annotators with the higher IAA in cases of disagreement. Table 2 presents the IAA and Cohen's Kappa scores of the dataset.

A total of 130 abstracts have been successfully annotated and were used for this task. The dataset is available in two formats [37], an XML format, for PICO sentence classification and a pseudo-IOB format [38] to enable entity predictions within the sentences.

2.2.2. Statement classifier

As the size of the dataset is prohibiting the use of neural approaches, we used machine learning techniques and algorithms to set our PICO Statement classifier baseline. We approach the task as a binary classification of the sentences to PICO Statements and we experimented with Multinomial Naive Bayes (MNB), Linear SVC (SVC), Gradient Boosting (GB), Extreme Gradient Boosting (XGB), Decision Trees (DT) and Random Forest (RF).

We employed Natural Language Processing techniques to extract syntactic and grammatical features from the abstracts as well as the PICO annotations from our PICO Entity Recognizer. Specifically, for each sentence, we extracted information about the location of the sentence in relation to the structure of the abstract if any exists (e.g. "Background", "Conclusions", etc.) and the number of PICO entities types found in the sentence using the PICO Entity Recognizer. We also used the word2vec embeddings [39] for the word representations and applied average pooling to create the sentence representations.

Table 1
Example statements that contain all PICO elements.

Statement A	These patients may particularly benefit from post-interventional exercise programs, but no randomized study has evaluated the safety and efficacy of exercise in this population.
Statement B	In patients after TAVI, exercise training appears safe and highly effective with respect to improvements in exercise capacity, muscular strength, and quality of life.

Table 2
PICO IAA and Cohen Kappa agreement statistics.

Labels	IAA	Kappa
Population	81.25%	60.53%
Intervention/comparison	78.69%	54.41%
Outcome	81.36%	61.26%
PICO	95.74%	84.54%

2.3. Experimental setup

We trained our PICO Entity Recognizer on the 5000 abstracts dataset, on the top level entities (Population, Intervention/Comparator and Outcome), using the predefined train and test splits. We experimented with a variety of different batch sizes, kernel sizes and filters as well as hidden sizes for the LSTM cells. After hyperparameter tuning, for our final PICO Entity Recognizer we used a batch size of 16. For the CNNs, we used kernels with sizes 1, 2, 3 and 4 and filters 40, 80, 120 and 160 respectively. We implemented two BiLSTMs, with a Highway Residual connection between, with hidden sizes of each LSTM layers, d_{rnn}, 712 for the first and 365 for the second. The character embeddings size d_{char} was set to 300. The ELMo embeddings dimension d_{ELMo} is 1024 and the weights are pre-trained on PubMed [40], commonly referred to as BioELMo. We use sentence padding to the maximum sentence length in the training set (n_{max}) and character padding to the maximum word length in the training set (m_{max}). Adam [41] was used for weight optimization through back-propagation with a learning rate of 0.001 and a decay of 0.90 per epoch. We employed a 0.5 Dropout [42] between layers to avoid overfitting. The model was trained for 20 epochs with early stopping.

For our sentence classifier, given the limited dataset, we used a 10-fold cross validation methodology to measure performance. We opted for 10-fold instead of a leave-one-out approach, to maintain moderate variance in our test sets. The word representations used are pre-trained on PubMed [43].

The experiments were run on a computer with a single Titan V 12GB graphics card, 32GB of memory and an Intel i7-8700 processor. The PICO Entity Recognizer and the PICO Statement classifier are both available at https://github.com/nstylia/pico_entities/.

3. Results

We compare the performance of our PICO Entity Regonizer to the baseline approach presented in Nye et al. [23], and the state-of-the-art approaches of BioFlair [44], BERT [45] and Cho et al. [46], referred to as CombBiLSTM in this publication, which achieve competitive performances in a plethora of biomedical corpora for sequence tagging. Additionally, we highlight the performance of our sentence classifier with the use of PICO entity tags by our PICO Entity Recognizer. In all the compared methodologies we used the recommended hyper-parameters as described in their respective publications.

The results in Table 3 display two different results as baseline. This is due to a disagreeing performance between the reported scores in Nye et al. [23] and the results we were able to achieve through the code provided from the cited repository.[2] As a result we report both scores for reasons of transparency. Our model, labeled "Att-HR-BiLSTMs",

[2] https://github.com/bepnye/EBM-NLP.

Table 3
PICO Entity Recognizer performance comparison.

Models	Population			Intervention			Outcome			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline – code	73	82	77	50	60	54	77	56	65	68	61	64
Baseline – published	78	66	71	61	70	65	69	58	63	–	–	–
Att-HR-BiLSTMs	82	89	80	55	71	62	80	59	68	71	70	70
BERT	78	57	68	53	62	57	72	85	78	69	66	68
BioFlair + BioELMo	77	87	79	58	67	60	84	53	62	71	67	68
CombBiLSTM	73	67	79	54	60	57	76	52	62	67	61	64

Underlined values represent the best scores, per metric per entity.

achieved the best overall performance and the best individual entity performance in terms of Population. We notice that, compared to our approach, BERT performs better in Outcome entities due to a very high Recall and BioFlair + BioElmo performs slightly better in terms of Precision in Intervention and Outcome entities, however the overall performance of both suffers. What is more, BERT needs at least 40 epochs of training before it achieves these results with 25% slower training time per epoch, BioFlair+BioElmo required a training of 50 epochs but is 43% faster and CombBiLSTM requires 40 epochs of training with a 38% faster epoch training. Consequently, our approach requires less overall time to train.

The results of the different algorithms presented in Table 4 present the performance of different machine learning algorithms over the PICO Statement dataset, with the same set of features. We report two different scores for each algorithm, depending on the use of PICO entity predictions to highlight what approaches can benefit from the use of predictions. We notice that while the majority of algorithms are able to depend on the existence of PICO entities, the Naive Bayes is unable to due to the feature independence assumption and therefore maintains the same performance.

3.1. Ablations

While the gains in performance are clear, we demonstrate that our approach has no redundant components and that the changes have contributed to the improved performance of the final model.

Table 5 exhibits the slightly reduced performance transitioning from the 1D CNN with a BiLSTM network approach [31,47] to a 2D CNN approach on the baseline model. We notice that while the transitioning step shows no performance gains, it allows for faster computations. Specifically, we averaged 6 and 10 min training time reduction, per epoch, on the baseline and full model respectively, making our approach 20% faster. In addition, we display an increase in precision and recall with the use of the highway residual connection to stacked BiLSTMs (HR) and the self-attention mechanism (Att). Furthermore, the addition of a highway residual connection every two RNN layers allows for effective information passing that in turn boosts the performance by 2%.

As we notice in Table 4, there is a significant performance boost on some algorithms, towards the prediction of PICO Statements using the PICO Entity Recognizer to predict PICO entities in the sentences. In Table 6 we compare the difference in performance in these algorithms when trained using the gold labels instead of the predictions from our PICO Entity Recognizer system. Moreover, we show that the performance of the statement classifier using the PICO entity predictions over the sentences has varying differences over using the gold annotations that already exist in the dataset. Specifically, we notice that Gradient Boosting is able to handle incorrect PICO entity predictions and use the rest of the feature space to make a correct decision while the Decision Tree algorithms are heavily depended on the PICO Entity Recognizer. These results also suggest that a perfect PICO Entity Recognizer could only boost the classifiers performance to the scores in Table 6, with the used features.

Table 4
PICO Statement classifier with and without the use of PICO Entity Recognizer predictions.

Models	Without PICO entities			With PICO entities		
	Precision	Recall	F1	Precision	Recall	F1
MNB	17%	89%	29%	17%	89%	29%
SVC	8%	7%	7%	57%	82%	65%
GB	2%	10%	1%	79%	96%	86%
XGB	25%	2%	4%	79%	90%	84%
DT	22%	24%	23%	77%	76%	76%
RF	24%	9%	12%	77%	79%	77%

Underlined values represent the per metric per method of acquired entities used in the experiments.

4. Discussion and conclusions

In this study (a) we presented a novel end-to-end PICO Entity Recognizer, (b) we introduced a PICO Statement identification task and (c) we provided a PICO Statement classifier, and showcased the importance of a good PICO entities recognizer to this task. To enable the statement classification task we also created a high quality manually annotated dataset and provided benchmark results on a series of machine learning sentence classification algorithms.

In our Entity Recognizer, we introduced a 2D CNN approach that negates the need to have RNN layers to extract features from character embeddings by providing similar results and being less computationally expensive. We also introduced a residual highway connection to avoid issues such as vanishing gradient and allow for deeper neural architectures when the number of data is not extensive. From the ablation study, it is evident that both our 2D approach and the Highway Residual connection have contributed to the performance of the final model.

Furthermore, our Entity Recognizer is an end-to-end system and does not use external information to provide annotations. By avoiding the use of syntactical information, as proposed in previous research, we are not restricted to annotate limited sections of the documents only (e.g. abstract), but text of varying length. In an extension of our system, the use of pre-existed ontologies (e.g. UMLS) could benefit the performance of the system in specific entity types that are hard to match without pre-existing knowledge.

We also identified a need for easier parsing of the available medical publications by medical practitioners that are using EBM to ascertain the best possible treatment. To this end, we introduced a PICO Statement classifier which can extract the sentences of medical publications containing such statements. While the system maintains a high performance, it is heavily based upon identifying the PICO entities in the abstracts. As a result, a powerful PICO Entity Recognizer, such as the one we proposed, can strengthen the performance of the statement classifier.

Conclusively, we showcased how a robust PICO Entity Recognizer can benefit the PICO Statement classification task, which is directly related to extracting evidence from medical publications. With the ability to identify PICO entities in full medical publications, we have also enabled other researchers to efficiently use PICO entities in their respective tasks (e.g. question answering, information retrieval, etc.).

Table 5
PICO Entity Recognizer ablations.

Models	Population			Intervention			Outcome			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1D CNN + BiLSTM	73	82	77	50	60	54	77	56	65	68	61	64
2D CNN	52	77	72	52	54	53	68	77	72	66	61	63
Att-HR-BiLSTMs	82	89	80	55	71	62	80	59	68	71	70	70
-HR	72	86	78	50	71	58	84	57	68	69	68	69
-Att	73	84	78	50	67	57	79	59	67	69	68	68
−2D + 1D-BiLSTM	72	89	78	50	66	57	77	55	65	68	67	67

Underlined values represent the best scores, per metric per entity.

Table 6
PICO Statement ablations.

Models	PICO Predicted entities			PICO Gold entities		
	Precision	Recall	F1	Precision	Recall	F1
SVC	57%	82%	65%	58%	84%	68%
GB	79%	96%	86%	82%	98%	89%
XGB	79%	90%	84%	79%	89%	86%
DT	77%	76%	76%	96%	97%	96%
RF	77%	79%	77%	97%	98%	97%

Underlined values represent the per metric per method of acquired entities used in the experiments.

With our contributions we aim to enhance the practice of EBM and assist medical practitioners to faster decisions during treatment procedures.

Conflict of interest

None declared.

Acknowledgements

This research was partially funded by Atypon Systems, Inc.[3]

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.artmed.2020.101949.

References

[1] Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7(9):e1000326.
[2] Hung S-Y, Ku Y-C, Chien J-C. Understanding physicians' acceptance of the medline system for practicing evidence-based medicine: a decomposed tpb model. Int J Med Inform 2012;81(2):130–42. http://www.sciencedirect.com/science/article/pii/S1386505611002012.
[3] Huang X, Lin J, Demner-Fushman D. Evaluation of pico as a knowledge representation for clinical questions. AMIA annual symposium proceedings, vol. 2006 2006:359.
[4] Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA 1999;282(7):634–5.
[5] Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinform 2011;12(March (2)):S5.
[6] Methley AM, Campbell S, Chew-Graham C, McNally R, Cheraghi-Sohi S. Pico, picos and spider: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. BMC Health Serv Res 2014;14(Nov (1)):579.
[7] Chung GY. Sentence retrieval for abstracts of randomized controlled trials. BMC Med Inform Decis Mak 2009;9(1):10.
[8] Jin D, Szolovits P. Advancing pico element detection in medical text via deep neural networks. 2018 (arXiv iv preprint), arXiv iv:1810.12780.
[9] Abacha AB, Zweigenbaum P. Means: a medical question-answering system combining nlp techniques and semantic web technologies. Inf Process Manag 2015;51(5):570–94.
[10] Gulden C, Kirchner M, Schüttler C, Hinderer M, Kampf M, Prokosch H-U, et al. Extractive summarization of clinical trial descriptions. Int J Med Inform 2019.
[11] Tacconelli E. Systematic reviews: crd's guidance for undertaking reviews in health care. Lancet Infect Dis 2010;10(4):226.
[12] Cooke A, Smith D, Booth A. Beyond pico: the spider tool for qualitative evidence synthesis. Qual Health Res 2012;22(10):1435–43.
[13] Xu R, Supekar K, Huang Y, Das A, Garber A. Combining text classification and hidden markov modeling techniques for structuring randomized clinical trial abstracts. AMIA annual symposium proceedings 2006:824–8.
[14] Boudin F, Nie J-Y, Bartlett JC, Grad R, Pluye P, Dawes M. Combining classifiers for robust pico element detection. BMC Med Inform Decis Mak 2010;10(May (1)):29.
[15] Huang K-C, Liu CC-H, Yang S-S, Xiao F, Wong J-M, Liao C-C, et al. Classification of pico elements by text features systematically extracted from pubmed abstracts. In: 2011 IEEE international conference on granular computing; 2011. p. 279–83.
[16] Verbeke M, Van Asch V, Morante R, Frasconi P, Daelemans W, De Raedt L. A statistical relational learning approach to identifying evidence based medicine categories. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. EMNLP-CoNLL'12. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 579–89. http://dl.acm.org/citation.cfm?id=2390948.2391014.
[17] Kiritchenko S, De Bruijn B, Carini S, Martin J, Sim I. Exact: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak 2010;10(1):56.
[18] Summerscales RL, Argamon S, Bai S, Hupert J, Schwartz A. Automatic summarization of results from clinical trials. In: 2011 IEEE international conference on bioinformatics and biomedicine; 2011. p. 372–7.
[19] Alamri A, Stevenson M. Automatic detection of answers to research questions from medline abstracts. Proceedings of BioNLP 15 2015:141–6.
[20] Scells H, Zuccon G, Koopman B, Deacon A, Azzopardi L, Geva S. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval. SIGIR'17. New York, NY, USA: ACM; 2017. p. 1237–40. https://doi.org/10.1145/3077136.3080707.
[21] Wallace BC, Kuiper J, Sharma A, Zhu MB, Marshall IJ. Extracting pico sentences from clinical trial reports using supervised distant supervision. J Mach Learn Res 2016;17(132):1–25.
[22] Ferracane E, Marshall I, Wallace BC, Erk K. Leveraging coreference to identify arms in medical abstracts: an experimental study. Proceedings of the seventh international workshop on health text mining and information analysis 2016: 86–95.
[23] Nye B, Li JJ, Patel R, Yang Y, Marshall IJ, Nenkova A, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Proceedings of the conference. Association for computational linguistics. Meeting, vol. 2018 2018:197.
[24] Huang K-C, Liu CC-H, Yang S-S, Xiao F, Wong J-M, Liao C-C, et al. Classification of pico elements by text features systematically extracted from pubmed abstracts. In: 2011 IEEE international conference on granular computing; 2011. p. 279–83.
[25] Demner-Fushman D, Lin J. Knowledge extraction for clinical question answering: preliminary results. Proceedings of the AAAI-05 workshop on question answering in restricted domains. Pittsburgh, PA: AAAI Press (American Association for Artificial Intelligence); 2005. p. 9–13.
[26] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Comput Linguist 2007;33(1):63–103.
[27] Sarker A, Mollá D, Paris C. Automatic evidence quality prediction to support evidence-based decision making. Artif Intell Med 2015;64(2):89–103. http://www.sciencedirect.com/science/article/pii/S093336571500041X.
[28] Marshall IJ, Kuiper J, Banner E, Wallace BC. Automating biomedical evidence synthesis: robotreviewer. Proceedings of the conference. Association for computational linguistics. Meeting, vol. 2017 2017:7.
[29] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. 2018 (arXiv iv preprint), arXiv iv: 1802.05365.
[30] Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models. In: Proceedings of the thirteenth AAAI conference on 450 artificial intelligence. AAAI'16; 2016. p. 2741–9.
[31] Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. 2016 (arXiv iv preprint), arXiv iv:1603.01354.
[32] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. International conference on machine learning 2013:1310–8.

[3] https://www.atypon.com/.

[33] Srivastava RK, Greff K, Schmidhuber J. Highway networks. 2015 (arXiv iv preprint), arXiv iv:1505.00387.
[34] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2014 (arXiv preprint), arXiv:1409.0473.
[35] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Roberts I, et al. Building a semantically annotated corpus of clinical texts. J Biomed Inform 2009;42(5): 950–66. Biomedical Natural Language Processing.
[36] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20 (1):37–46.
[37] Stylianou N, Razis G, Goulis DG, Vlahavas I. Pico statements dataset. 2020. https://doi.org/10.17632/p5rbn8mygp.1.
[38] Ramshaw L, Marcus M. Text chunking using transformation-based learning. Third workshop on very large corpora 1995:82–94. https://www.aclweb.org/anthology/W95-0107.
[39] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th international conference on neural information processing systems – vol. 2. NIPS'13. Red Hook, NY, USA: Curran Associates Inc.; 2013. p. 3111–9.
[40] Jin Q, Dhingra B, Cohen WW, Lu X. Probing biomedical embeddings from language models. 2019 (arXiv iv preprint), arXiv iv:1904.02181.
[41] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014 (arXiv preprint), arXiv:1412.6980.
[42] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15 (1):1929–58.
[43] Moen S, Ananiadou TSS. Distributional semantics resources for biomedical text processing. Proceedings of LBM 2013:39–44.
[44] Sharma S, Daniel Jr R. Bioflair: pretrained pooled contextualized embeddings for biomedical sequence labeling tasks. 2019 (arXiv iv preprint), arXiv iv:1908.05760.
[45] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018 (arXiv preprint), arXiv iv: 1810.04805.
[46] Cho M, Ha J, Park C, Park S. Combinatorial feature embedding based on cnn and lstm for biomedical named entity recognition. J Biomed Inform 2020;103:103381.
[47] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. 2016 (arXiv iv preprint), arXiv iv: 1603.01360.