EasyChair Preprint

№ 2940

Extracting Outcomes from Articles Reporting Randomized Controlled Trials Using Pre-Trained Deep Language Representations

Anna Koroleva, Sanjay Kamath and Patrick Paroubek

March 11, 2020

EXTRACTING OUTCOMES FROM ARTICLES REPORTING RANDOMIZED CONTROLLED TRIALS USING PRE-TRAINED DEEP LANGUAGE REPRESENTATIONS

A PREPRINT

Anna Koroleva[*]
Institute of Applied Simulation
School of Life Sciences and Facility Management
Zurich University of Applied Sciences (ZHAW)
Waedenswil, Switzerland
aakorolyova@gmail.com

Sanjay Kamath[†]
Total
Saclay, France

Patrick Paroubek
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay, France
pap@limsi.fr

March 11, 2020

ABSTRACT

Objective: Outcomes are the variables monitored during clinical trials to assess the impact of the intervention studied on the subjects' health. Automatic extraction of trial outcomes is essential for automating systematic review process and for checking the completeness and coherence of reporting to avoid bias and spin. In this work, we provide an overview of the state-of-the art for outcome extraction, introduce a new freely available corpus with annotations for two types of outcomes – declared (primary) and reported – and present a deep learning approach to outcome extraction.

Dataset: We manually annotated a corpus of 2,000 sentences with declared (primary) outcomes and 1,940 sentences with reported outcomes.

Methods: We used deep neural word embeddings derived from the publicly available BERT (Bidirectional Encoder Representations from Transformers) pre-trained language representations to extract trial outcomes from the section defining the primary outcome and the section reporting reporting the results for an outcome. We compared a simple fine-tuning approach and an approach using CRF and Bi-LSTM. We assessed the performance of several pre-trained language models: general domain (BERT), biomedical (BioBERT) and scientific (SciBERT).

Results: Our algorithm achieved the token-level F-measure of 88.52% for primary outcomes and 79,42% for reported outcomes.

Conclusion: Fine-tuning of language models pre-trained on large domain-specific corpora show operational performance for automatic outcome extraction.

Keywords Natural Language Processing · Randomized Controlled Trials · Outcome extraction · Deep Neural Networks · Pre-trained Language Representations

[*]At the time of reported work, Anna Koroleva was a PhD student at LIMSI-CNRS in Orsay, France and at the Academic Medical Center, University of Amsterdam in Amsterdam, the Netherlands.

[†]At the time of reported work, Sanjay Kamath was a PhD student at LIMSI-CNRS and LRI Univ. Paris-Sud in Orsay, France.

1 Introduction

Outcomes of clinical trials are the dependent variables monitored during a trial in order to establish how they are influenced by independent variables such as the intervention taken, dosage, or patient characteristics. Outcomes are a key element of trial design, reflecting its main goal and determining the trial's statistical power and sample size.

Previous works have shown that outcome extraction is a difficult task because of the diversity of outcome mentions and contexts in which they occur. No common textual markers exist (e.g. capitalization, numerical symbols, cue phrase). Recent research proved that the use of deep language representations pre-trained on large corpora, such as BERT[1], outperforms the state-of-the-art results for several natural language processing tasks, including entity extraction. Pre-training on large domain-specific data can further improve the results[2, 3].

We propose a deep learning approach, using language representations pre-trained on general domain corpora and on domain-specific datasets to extract trial outcomes. We report on creating a publicly available annotated corpus for outcome extraction.

1.1 Definitions

There are substantial discrepancies in the use of the words "outcome", "endpoint", "outcome measure" etc., that we describe in detail elsewhere. In brief, there is no agreement between researchers as for the differences in the meaning and usage of these terms, in practice they are often considered to be synonyms. In our work, we follow the common practice and do not distinguish between these notions. We prefer to use the term "outcome".

Following the accepted usage[3], we define an outcome as *a variable (or measure, or parameter) monitored during a clinical trial*. Outcome in this sense is a type of *entity*, as it is understood by the standard entity recognition task. Our definition differs from that given by Demner-Fushman and colleagues[4] who defined an outcome as "*sentence(s) that best summarizes the consequences of an intervention*". In our definition an outcome is usually shorter than a sentence, and it does not refer to trial results ("consequences of an intervention"): the results are the *values* of outcomes. Our definition is in line with other works on outcome extraction (see the Related Works section), as many applications, such as summarization of trial results, require extracting outcomes on the entity level to allow for further analysis of data for each individual outcome, which is not possible if extracting only sentences containing outcomes (cf. [5]).

We introduce here the definitions for two types of outcome mentions that are important for our work: declared and reported outcomes.

Declared outcomes are the mentions of outcomes that occur in contexts that explicitly state which variables were measured in a trial, e.g. (outcomes are in bold):

The primary outcome of this study was ***health-related quality of life****.*

Secondary outcomes included ***changes in the 6-minute walk distance (6MWD)*** *and* ***adverse events****.*

In our study, we were most interested in ***changes in PHQ-9 scores after the 12-week trial****.*

Declared outcomes can be further classified according to their importance as stated by the authors (primary, secondary, tertiary, or undefined).

Reported outcomes are the mentions of outcomes that occur in contexts that report the results for the outcomes, e.g. (outcomes are in bold):

The HRQoL *was higher in the experimental group.*

The mean incremental QALY *of intervention was 0.132 (95% CI: 0.104–0.286).*

1.2 Applications

Extraction of trial outcomes is an important part of systematic review process[6], clinical question answering[7], assessment of an article for distorted reporting practices such as bias[8], outcome switching[9] and spin[10]. For us, the main application of interest is spin detection.

In general, spin is defined as presenting research results as being more positive than the experiments proved. In particular, in randomized controlled trials (RCTs) assessing a new intervention, spin consists in exaggerating the beneficial effects (efficacy and/or safety) of the studied intervention. As RCTs are the main source of information for

[3] e.g. https://rethinkingclinicaltrials.org/chapters/design/choosing-specifying-end-points-outcomes/choosing-and-specifying-endpoints-and-outcomes-introduction/

Evidence-Based Medicine, spin in RCTs presents a serious threat to the quality of healthcare. The presence of spin makes clinicians overestimate the effects of the treatment in question[11], and provokes spin in health news and press releases[12, 13], which can affect public expectations regarding the treatment.

One of the most common forms of spin is *selective reporting of trial outcomes* – reporting only the outcomes that prove the hypothesis of the authors. To automatically detect this form of spin, declared and reported trial outcomes need to be extracted, and declared outcomes must be compared to the reported outcomes to check for mismatches: declared outcomes that are not reported, or reported outcomes that were not declared.

Our current work presents the first step of the algorithm of selective outcome reporting detection and deals with the extraction of declared and reported outcomes. The second step consists in assessing semantic similarity between pairs of declared and reported outcomes and will be presented elsewhere.

2 Related work

As the volume of published biomedical articles grows exponentially[14], manual extraction of clinical trial information becomes infeasible. Several works addressed the extraction of outcome-related information for facilitating systematic reviews or supporting clinical question-answering systems.

A number of works addressed extraction of information on clinical trials using the PICO - Patient/Problem, Intervention, Comparison, Outcome - framework[15] or its extensions. The majority of works using the PICO framework treat the task as sentence classification[16, 17, 18, 19, 20, 21, 22, 23, 24]. F-measure for outcome sentence extraction varies between 54% and 88% for different methods (see the systematic review[6]).

Demner-Fushman and colleagues[7, 4] also treated the task of outcome extraction as text classification. The authors trained several classifiers on a dataset of 633 MEDLINE citations. Naive Bayes classifier outperformed linear SVM and decision-tree classifier. The accuracy of outcome sentence identification ranged from 88% to 93%.

However, for some tasks (including spin detection), identification of relevant sentences is not enough and extracting outcomes at the entity level is required. This task has been addressed by fewer works than the PICO classification. It is important to distinguish the works addressing the extraction of declared (primary and secondary) outcomes from those targeting reported outcomes.

De Bruijn and colleagues[25] aimed at extracting declared (primary and secondary) outcomes and their time points, along with other elements of trial design. They point out that the outcomes of a trial can be poorly defined by referring to "main outcomes" instead of primary and secondary ones. The authors also note that it is necessary to analyse the whole article, not only the abstract, e.g. to find secondary outcomes. The system uses a two-step approach: first, a classifier is applied to identify sentences containing a given type of information; second, regular expression rules are used to find text fragments corresponding to the target information. The dataset used in this work consists of 88 randomly selected full-text articles from five medical journals. For primary and secondary outcomes, only the first step (sentence classification) was implemented. Performance for identification of sentences containing outcomes is reported to be lower that for the other elements. For the primary outcomes, the sentence classification reaches precision of 87% and recall of 90%; for secondary outcomes, precision was 57% and recall was 90%.

In their following work[26], the authors further develop their approach and add rules for extracting text fragments for primary and secondary outcomes. This work used a different dataset: the initial corpus consisted of 78 manually annotated articles from five clinical journals that were considered to be representative of general medicine, to which 54 articles from a wider selection of journals were added, resulting in a final training set of 132 articles from 22 clinical journals. The test set contained 50 full-text articles reporting RCTs from 25 journals. The results were assessed at the sentence and fragment levels. The sentence classification performance for outcomes is as follows: precision was 66%, recall was 69% for primary outcomes; precision was 69% and recall 79% for secondary outcomes. For fragment extraction, the authors report for primary outcomes a precision and recall of 97% for both overlapping and exact matches; for secondary outcomes, precision and recall for exact matches are 93% and 88% respectively, and for overlapping matches, both precision and recall are 100%.

Summerscales and colleagues[27] addressed the task of identifying treatments, patient groups and reported outcomes in abstracts of medical articles. The authors created a corpus of 100 abstracts of articles published in the BMJ[4], extracted from PubMed[5]. The corpus of 1,344 sentences contained 1,131 outcomes, 494 out of which were unique. Outcomes vary in length from 1 to 14 tokens (mean = 3.6). The examples of outcomes given in the article are noun phrases, but the authors did not specify whether they annotated only noun phrases or included other syntactic constituents (e.g. verb

[4] https://www.bmj.com/

[5] http://www.ncbi.nlm.nih.gov/pubmed/

phrases, adjectives). The authors note that the boundaries of entities are often ambiguous and annotating each variant is not optimal; thus, they suggest to evaluate both exact and partial matches. The authors trained a Conditional Random Field (CRF) classifier to label each word, using features such as the word form, its POS tag, corresponding Medical Subject Heading ID, its semantic tag(s) (anatomy, time, disease, symptom, drug, procedure and measurement terms, assigned using lists of terms), the title of the enclosing section, and four words to the left and right of the word with their POS and semantic tags. The token-levels results for outcomes are: precision 75%, recall 62%, and F-measure 68%.

In their following work[28], the authors enlarge their dataset to 263 abstracts of BMJ articles. A first-order linear-chain CRF classifier on this set yielded a precision of 56%, recall of 34%, and F-measure of 42% for outcome extraction.

Blake and Lucic[29] aimed at extracting noun phrases for three items in comparative sentences: two compared entities (the agent and the object) and the ground for comparison (endpoint, or outcome). The dataset for this work included the sentences containing all the three items (agent, object and endpoint), selected from over 2 million sentences from full-text medical articles. 100 sentences with 656 noun phrases constituted the training set. First the algorithm finds comparative sentences with the use of a set of adjectives and lexico-syntactic patterns. Then two classifiers - SVM and Generalized Linear model (GLM) —– are used to predict the roles (agent, object, endpoint) of noun phrases. SVM showed better results that GLM on the training set (for endpoint, precision=67%, recall=94% and F-measure=78%). However, on the test set the results were significantly lower: SVM achieved precision of 42%, recall of 64% and F-measure 51% for endpoint detection. The performance was evaluated separately on shorter sentences (up to 30 words), where it was higher than on longer sentences.

The following work[30] used the information whether the head noun of the candidate noun phrase denotes an amount or a measure, in order to improve the detection of the first entity and of the endpoint. The annotation of the corpus was enriched by the corresponding information, which resulted in an improvement of endpoint detection: precision was 56% on longer sentences and 58% on shorter ones; recall was 71% on longer sentences and 74% on shorter ones.

A recent work of Nye et al[31] describes the development of a crowd-sources corpus of nearly 5000 abstracts with annotations for patients, interventions and outcomes. The authors provide the results of two baseline algorithms for extracting these entities. A linear CRF model, using current, previous and next tokens, pos-tags, and character information as features, achieved the precision of 83%; recall of 17% and F-measure of 29%. A neural model, based on a bi-directional LSTM passing distributed vector representations of input tokens to a CRF, yielded the precision of 69%, recall of 58% and F-measure of 63%.

3 Dataset

In the course of our work on spin detection, we annotated a corpus of declared and reported outcomes. The reason for creating this new corpus is the absence of any available resource with the annotation for these two types of outcome mentions. The only currently available corpus with outcome annotation, to our knowledge, is that introduced by Nye et al[31], which was not available at the time of the beginning of our work and which does not distinguish between declared and reported outcomes, while this distinction is of vital importance for our project.

Our corpus is based on a dataset of 3,938 PMC[6] articles, selected from a larger corpus (119,339) of PMC articles on the basis of being assigned the PubMed publication type "Randomized controlled trial". We annotated a corpus of sentences from full-text articles for two types of entities: declared outcomes and reported outcomes. For declared outcomes, we annotated only primary outcomes, as they are the most important for our final goal of spin detection (omission or change of the primary outcome is one of the most common types of spin). The annotation and extraction of secondary outcomes is one of the directions of future work on this task.

As it proved to be impossible to run a large-scale annotation project with several expert annotators, the annotation was performed by AK with guidance by domain experts. We developed an annotation tool[32] for the sake of simplicity, ease of format conversion and customizing. The annotation uses a CoNLL-like representation scheme with B (begin) - I (inside) - O (outside) elements.

3.1 Declared outcome annotation

Misreporting of the trial outcomes is most often related to the primary outcome of a trial, thus, the primary outcome presents the highest interest for spin detection. We annotated the declared outcomes only in the contexts that explicitly state that the outcome was the primary one in the given trial, e.g. (the outcome is in bold):

*The primary outcome was **the PHQ-9**.*

[6] https://www.ncbi.nlm.nih.gov/pmc/

Information about the primary outcome can sometimes be expressed implicitly, in statements about objectives or in descriptions of measured variables. For example, in the absence of explicit definition of the primary outcome (while secondary outcomes are clearly defined), the readers can infer that "*the human gastrointestinal microbiota*" and "*metabolic markers of health*" were the primary outcomes in:

*We aimed to assess the impact of walnut consumption on **the human gastrointestinal microbiota** and **metabolic markers of health**. Fecal and blood samples were collected at baseline and at the end of each period to assess secondary outcomes of the study, including effects of walnut consumption on fecal microbiota and bile acids and metabolic markers of health.*

In the following example "*Cognitive functioning*" can be interpreted as the trial's primary outcome as no other outcome is defined in the abstract:

***Cognitive functioning** was measured at baseline and after 12 weeks.*

To assess the need of including these types of statement in our corpus, we conducted a qualitative corpus study and consulted our medical advisors (Isabelle Boutron, Patrick Bossuyt and Liz Wager) in the course of the supporting project (2016-2019). We compared the variables mentioned in the statements on measured variables ("***X** was measured*", "*We aimed at measuring **X***", etc.) to the variables explicitly stated to be the primary outcomes in the same text. Our corpus study showed that the variables described in this type of statements differ from the explicitly declared primary outcomes: in particular, statements of objectives usually contain more general description of what was studied (e.g. "efficacy") compared to outcomes (e.g. "survival" or "quality of life"). Thus, we concluded that these types of statement do not define a primary outcome. Furthermore, absence of an explicit definition of the trial's primary outcome does not conform with good reporting practices[33, 34]. Hence we excluded these types of sentences from our corpus.

To create the corpus for declared primary outcome annotation, we searched the full-text articles for sentences where both the word "*primary*" (or its synonyms: "*principal*", "*main*", etc.) and the word "*outcome*" (or its synonyms: "*end-point*", "*measure*", etc.) occur, the former precedes the latter, and the distance between them is no more than 3 tokens. Regular expressions were used to search for the terms and Python NLTK library[7] was used for sentence splitting. Out of the sentences corresponding to our criteria, we randomly selected 2,000 sentences, coming from 1,672 articles.

We created two versions of the corpus which differ in annotation of coordinated outcomes. In the first version of our corpus, we annotated coordinated outcomes as one entity for the sake of simplifying the annotation task. This version contains 1,253 occurrences of declared primary outcomes. Further, we created a more elaborated version of the annotation, re-marking coordinated outcomes as separate entities. This second version contains 1,694 occurrences of declared primary outcomes. This version will be further used in our target application. The two versions serve to assess the capabilities of our algorithms to correctly analyse coordinated entities.

A definition of a primary outcome can include time points, measurement tool, etc. We annotated the longest continuous text span containing all the relevant information about the trial's primary outcome. Declared primary outcomes are most typically represented by noun phrases, but can also be expressed by verb phrases or clauses:

*Our primary outcome measures will be (a) **whether the TUPAC guideline recommendations are implemented**, and (b) if implemented, the estimated time used for the counselling.*

3.2 Reported outcome annotation

We annotated reported outcomes in the abstracts of the articles for which we annotated the primary outcomes, to allow for further annotation of spin related to incomplete outcome reporting. We extracted the Results and Conclusions sections of the corresponding articles (using rules and regular expressions). A number of articles in our corpus are not RCT reports but trial protocols, thus their abstracts did not contain Results and Conclusions. These abstracts were excluded from the reported outcomes corpus. The final corpus contains 1,940 sentences from 402 articles. A total of 2,251 reported outcomes was annotated.

The ways of reporting outcomes differ, and the same outcome can be reported in several ways, e.g. the sentence:

***Mean total nutrition knowledge score** increased by 1.1 in intervention (baseline to follow-up : 28.3 to 29.2) and 0.3 in control schools (27.3 to 27.6).*

can be rewritten as:

*The increase in **mean total nutrition knowledge score** was 1.1 in intervention (baseline to follow-up : 28.3 to 29.2) and 0.3 in control schools (27.3 to 27.6).*

[7] http://www.nltk.org

While both sentences report the same outcome, the structure is different, and for the second sentence both "*The increase in mean total nutrition knowledge score*" and "*mean total nutrition knowledge score*" can be considered to represent the outcome. Besides, there is a choice whether to include the aggregation method ("*mean*") into the outcome, or annotate simply "*total nutrition knowledge score*". In order to preserve uniformity throughout the annotation of reported outcomes, we decided to annotate the smallest possible text fragment referring to an outcome ("*total nutrition knowledge score*" for the given example) as it allows to annotate the same text fragment for all the variants of outcome reporting.

Reported outcomes are characterized by high variability from the syntactic point of view. They can be represented either by a noun phrase:

*Overall **response rate** was 39.1% and 33.3% in 3-weekly and weekly arms.*

a verb phrase:

*No patients were **reintubated**.*

or an adjective:

*The CSOM and MA appeared less **responsive** following a GLM-diet.*

One of the challenges in annotating reported outcomes is classifying reported variable either as a trial outcome or as a independent variables or covariates[8]. We decided to annotate all the mentions of variables (outcomes or not) unless the context of the sentence or the semantics of the phrase allows to classify it as a non-outcome variable. For example, in the sentence:

*Adjustments for **age**, **gender**, and **treatment group** were performed, but did not change the results.*

the context allows to categorize all the variables as covariates.

It should be noted that this annotation decision leads to some counter-intuitive annotations: e.g. it can be expected that in a set of coordinated entities, either all the entities should be annotated as outcomes, or none of them. However, consider the following example:

***Age**, **gender** and **disease status** distribution was similar in both groups.*

Here "*Age*" and "*gender*" are considered to be independent variables due to their semantics, while "*disease status*" can be a dependent variable and will be the only entity annotated as outcome in this sentence.

There are a few differences between our corpus and the datasets used in previous works on outcome extraction. We address both declared (primary) and reported outcomes (annotation and extraction of secondary outcomes has not been covered yet and is a direction for our future work). We do not limit our dataset in terms of specific types of sentences (e.g. comparative) or constituents to be annotated (e.g. noun phrases). Our corpus is not limited to specific journals or topics. Our corpus is publicly available[35].

4 Methods

4.1 Baseline

We developed a simple rule-based baseline system, combining syntactic and sequential rules that cover the most typical patterns in which declared and reported outcomes can occur. For declared outcomes, sequential rules search for patterns such as:

*DET ADJ outcome was **DET ADJ* NN***

where DET denotes a determiner, ADJ is an adjective, and NN is a common noun. The sequence matched by "*DET ADJ* NN*" here is considered to be the outcome. Sequential rules use the information on tokens, lemmas and pos-tags of the words in the input text.

Syntactic rules search for similar patterns, but use the syntactic dependency graph (tags and directions of syntactic relations) instead of sequential information, to capture the cases where the target phrase is separated from the cue phrase (e.g. "*DET ADJ outcome was*") by other words.

Our rule-based baseline was designed to detect the declared outcomes in the first version of the corpus only (coordinated outcomes annotated as single entity). A simple rule-based approach can hardly be successful in a complex task such

[8] https://methods.sagepub.com/Reference//encyc-of-research-design/n85.xml

as dividing coordinated entities, hence we did not build a rule-based baseline for the second version of the declared outcomes corpus.

For reported outcomes, the searched patterns include the expressions with comparative meaning, e.g.:

DET ADJ* NN *increased.*

DET ADJ* NN *was higher in the NN arm.*

DET ADJ* NN *was NUM*,

where NUM denotes a numeral. Fragments matched by the patterns in bold are tagged as reported outcomes.

For pos-tagging and dependency parsing, we used spaCy dependency parser[36].

4.2 Bi-LSTM-CRF-char algorithm

Our second approach is inspired by the work of Ma and Hovy[37] and uses the implementation of this method proposed by G.Genthial[9]. First, the model gets character-level representations of words from character embeddings using a bi-directional LSTM (bi-LSTM). After that, the model combines the character-level representation with a GloVe[38] word vector representation and passes the combined representations to a bi-LSTM to build contextual representations of words. Finally, a linear chain CRF is applied to decode the labels. Table 1 shows the values of the parameters used in the configuration of the model.

Table 1: Training parameters

Parameter	Value
dim_word	300
dim_char	100
train_embeddings	False
nepochs	15
dropout	0.5
batch_size	20
lr_method	"adam"
lr	0.001
lr_decay	0.9
clip	-1
nepoch_no_imprv	3
hidden_size_char	100
hidden_size_lstm	300

4.3 BERT-based algorithms: the use of deep pre-trained Language models

Recently, language models pre-trained on large corpora with complex neural network architectures have been shown to be useful for several downstream NLP tasks such as question answering, named entity recognition, natural language inference, etc. by ELMO[39], OpenAI's GPT[40] and Google's BERT[1]. Intuition is to build a model trained on a large corpus for a relatively simple task of language modelling, which can further be modified for complex NLP tasks. There are two approaches to employing these pre-trained models for supervised downstream tasks:

1. feature-based approach (used in ELMO) relies on task-specific architecture, where pre-trained representations are included as additional features to existing neural network models;
2. fine-tuning approach (used in OpenAI GPT and BERT) does not require extensive task-specific parameters, it simply fine-tunes the pre-trained parameters on a downstream task.

We compared a number of recent deep pre-trained language models. First, we employed the BERT (Bidirectional Encoder Representations from Transformers) models which are well documented with openly available pre-trained weights for the models[10]. In brief, BERT uses a masked language model (MLM), randomly masking some input tokens, which allows to pre-train a deep bidirectional Transformer using both left and right contexts. Representation of a token combines the corresponding token, segment and position embeddings. The advantage of BERT compared to ELMO

[9] https://github.com/guillaumegenthial/sequence_tagging

[10] https://github.com/google-research/bert

and OpenAI GPT is the deep bi-directionality of the representations and the size of the training corpus. We chose to use BERT because it outperformed ELMO and OpenAI GPT on a number of tasks[1]. There are several versions of BERT models: cased and uncased models, differing in the preprocessing of the input data (lower-cased vs unchanged); and base and large models, differing in the model sizes.

BioBERT[2], a domain-specific analogue of BERT, was pre-trained on a large (18B words) biomedical corpus: PubMed abstracts and PMC full-text articles, in addition to BERT training data. BioBERT is based on the cased BERT base model. Another domain-specific version of BERT is SciBERT[3], trained on a corpus of scientific texts (3.1B) added to BERT training data. SciBERT provides both cased and uncased models, with two versions of vocabulary: BaseVocab (the initial BERT general-domain vocabulary) and SciVocab (the vocabulary built on the scientific corpus). Both BioBERT and SciBERT outperformed BERT on some tasks of biomedical natural language processing.

Table 2 summarizes the training data of BERT, BioBERT and SciBERT.

Table 2: Training data for BERT/BioBERT/Scibert

	BERT	SciBERT	BioBERT
Training data	BooksCorpus English Wikipedia	BooksCorpus English Wikipedia + Semantic Scholar (Biomedical, Computer Science)	BooksCorpus English Wikipedia + PubMed Abstracts PMC Full-text articles
Volume of training data (words)	3.3B	6.4B	21.3B

The models we evaluated in our experiments include: BERT-base models, both cased and uncased; BioBERT model; SciBERT models, both cased and uncased, with the SciVocab vocabulary (recommended by the authors). We did not perform experiments with BERT-Large due to the lack of resources.

We explored two approached of employing the BERT-based language models for our sequence labelling task. The first approach, suggested by the developers of BERT and BioBERT[1, 2], employs a simple fine-tuning of the models on our annotated datasets. The principle behind this approach is that the pre-trained BERT models can be fine-tuned for a supervised task with one additional output layer. The one additional layer parameters along with the whole BERT model parameters are fine-tuned for the intended task. Table 3 summarizes the hyperparameters used for BERT-based models training and evaluation.

Table 3: BERT/BioBERT/SciBERT hyperparameters

Hyperparameter	**Value**	**Definition**
init_checkpoint	None	Initial checkpoint (usually from a pre-trained BERT model)
do_lower_case	True/False	Whether to lower case the input text
max_seq_length	128	The maximum total input sequence length after WordPiece tokenization
do_train	True	Whether to run training
use_tpu	False	Whether to use TPU or GPU/CPU
train_batch_size	32	Total batch size for training
eval_batch_size	8	Total batch size for eval
predict_batch_size	8	Total batch size for predict
learning_rate	5e-5	The initial learning rate for Adam
num_train_epochs	10.0	Total number of training epochs to perform
warmup_proportion	0.1	Proportion of training to perform linear learning rate warmup for
save_checkpoints_steps	1000	How often to save the model checkpoint
iterations_per_loop	1000	How many steps to make in each estimator call
master	None	TensorFlow master URL

The second approach, suggested by the SciBERT developers[3], uses minimal task-specific architecture on top of BERT-based embeddings. A representation of each token in this model is built by concatenating its BERT embedding and a CNN-based character embedding. Similarly to the method of Ma and Hovy[37], a multilayer bi-LSTM is applied to token embeddings, and a CRF is used on top of the bi-LSTM[11].

We compared performance of all the models both with unaltered input data and with lower-cased input data. It is expected that cased models perform better with unaltered input data, while uncased models perform better with lower-cased data. All the algorithms were evaluated on the token level. Machine-learning algorithms were assessed using 10-fold cross-validation (train-dev-test split was done in proportion 8:1:1). We report the averaged results. We used Tensorflow for our experiments.

5 Results and discussion

Tables 4, 5, 6 show the performance of the tested algorithms (all evaluations are done at the token level). The True value of the do_lower_case flag indicates lower-cased input data. The suffix "_biLSTM-CRF" for BERT-based model indicates the results of the approach using CRF on top of bi-LSTM.

Algorithm	**do_lower_case**	**Precision**	**Recall**	**F1**
SciBERT-cased	False	89.32	87.87	88.52
BioBERT	True	88.83	88.24	88.45
BioBERT	False	88.44	88.11	88.2
SciBERT-uncased	True	88.74	87.51	88.06
SciBERT-cased	True	88.34	87.82	88
BERT-cased	False	87.73	86.94	87.23
SciBERT-uncased	False	88.05	86.24	87.06
BERT-uncased	True	87.19	86.55	86.71
BERT-cased	True	88.46	85.12	86.68
BERT-uncased	False	86.97	86.08	86.42
SciBERT-uncased_biLSTM-CRF	True	85.01	83.76	84.3
SciBERT-cased_biLSTM-CRF	True	83.93	83.88	83.88
BioBERT_biLSTM-CRF	False	83.43	84.25	83.79
SciBERT-cased_biLSTM-CRF	False	83.37	83.64	83.49
BioBERT_biLSTM-CRF	True	83.12	83.78	83.42
SciBERT-uncased_biLSTM-CRF	False	80.59	81.76	81.15
BERT-uncased_biLSTM-CRF	True	80.26	81.52	80.87
BERT-cased_biLSTM-CRF	False	80.04	80.87	80.38
BERT-cased_biLSTM-CRF	True	78.3	80.97	79.58
BERT-uncased_biLSTM-CRF	False	78.49	79.37	78.87
Rule-based	-	78.6	68.98	73.51
Bi-LSTM-CRF-char	-	59.14	63.41	61.07

Table 4: Primary outcome extraction - version 1: results

5.1 Comparison of approaches

Our rule-based system showed reasonable performance for extracting primary outcomes (on the first version of the corpus), but not reported outcomes, as the latter are highly diverse and thus rule-based approach is not optimal. The Bi-LSTM-CRF-char algorithm using character and GloVe token embeddings did not show high performance for our tasks. All BERT-based models outperformed the Bi-LSTM-CRF-char algorithm and the rule-based baseline with a large absolute improvement (see Tables 4, 5, 6).

The fine-tuning approach employing BERT-based models consistently showed better performance than the approach using CRF on top of bi-LSTM with BERT-based embeddings. For all the tasks, even the best results of the model with CRF were inferior to the lowest results achieved by fine-tuning. This fact shows that a simple architecture (fine-tuning) can be superior to more complex (bi-LSTM-CRF) architectures for entity extraction task.

[11]The configuration and hyperparameters used for training the model can be found at: https://github.com/allenai/scibert/blob/master/allennlp_config/ner.json.

Algorithm	do_lower_case	Precision	Recall	F1
BioBERT	False	86.99	90.07	88.42
SciBERT-cased	False	87.52	89.07	88.21
SciBERT-uncased	True	87.49	88.92	88.1
SciBERT-cased	True	87.39	88.64	87.92
BioBERT	True	87.01	88.96	87.9
SciBERT-uncased	False	86.57	88.3	87.35
BERT-cased	False	86.96	87.41	87.14
BERT-uncased	True	86.6	87.39	86.91
BERT-uncased	False	86.96	86.87	86.84
BERT-cased	True	86.71	87.12	86.81
BioBERT_biLSTM-CRF	False	78.82	82	80.34
SciBERT-uncased_biLSTM-CRF	True	77.52	81.15	79.22
SciBERT-cased_biLSTM-CRF	False	77.23	80.89	78.95
BioBERT_biLSTM-CRF	True	77.86	80.12	78.9
BERT-cased_biLSTM-CRF	False	78.2	78.84	78.47
SciBERT-cased_biLSTM-CRF	True	77.05	79.73	78.29
SciBERT-uncased_biLSTM-CRF	False	77.54	78.67	78.07
BERT-uncased_biLSTM-CRF	True	76.72	78.73	77.63
BERT-cased_biLSTM-CRF	True	75.35	77.61	76.41
BERT-uncased_biLSTM-CRF	False	75.23	76.62	75.79
Bi-LSTM-CRF-char	-	49.16	52.21	50.55

Table 5: Primary outcome extraction - version 2: results

Algorithm	do_lower_case	Precision	Recall	F1
SciBERT-uncased	True	81.17	78.09	79.42
BioBERT	True	80.38	77.85	78.92
BioBERT	False	79.61	77.98	78.6
SciBERT-cased	False	79.6	77.65	78.38
SciBERT-cased	True	79.24	76.61	77.64
SciBERT-uncased	False	79.51	75.5	77.26
BERT-uncased	True	78.98	74.96	76.7
BERT-cased	False	76.63	74.25	75.18
BERT-cased	True	76.7	73.97	75.1
BERT-uncased	False	77.28	72.25	74.46
SciBERT-uncased_biLSTM-CRF	True	68.44	73.47	70.77
BioBERT_biLSTM-CRF	False	70.18	71.43	70.63
BioBERT_biLSTM-CRF	True	69.09	71.57	70.24
SciBERT-cased_biLSTM-CRF	False	67.98	72.52	70.11
SciBERT-cased_biLSTM-CRF	True	66.11	71.16	68.37
SciBERT-uncased_biLSTM-CRF	False	67.25	69.59	68.18
BERT-cased_biLSTM-CRF	False	65.98	65.54	65.64
BERT-uncased_biLSTM-CRF	True	64.6	66.73	65.4
BERT-cased_biLSTM-CRF	True	64.73	66.49	65.37
BERT-uncased_biLSTM-CRF	False	62.07	64.98	63.29
Bi-LSTM-CRF-char	-	51.12	44.6	47.52
Rule-based	-	26.69	55.73	36.09

Table 6: Reported outcome extraction: results

5.2 Comparison of BERT-based models

Out of all the tested approaches, fine-tuned BioBERT model showed the best performance for the second version of primary outcome extraction; fine-tuned SciBERT model outperformed other systems for the first version of primary outcome extraction (cased model) and reported outcome extraction (uncased model).

As expected, BERT and SciBERT uncased models performed better with lower-cased input, while cased models performed better with unchanged input data. On the contrary, BioBERT model (cased) performed slightly better with lower-cased input for two out of three tasks. A possible explanation is that BioBERT has the largest amount of training data, where the majority of the input is naturally in lower case, thus the learnt representations show similar performance for lower-cased or unchanged input.

Overall, SciBERT and BioBERT outperformed BERT, supporting the hypothesis that motivated their creation: while pre-training of language representations on large corpora gives good results, adding domain-specific corpora to the pre-training data further improves the performance they yield. SciBERT and BioBERT show comparable performance, demonstrating that addition of domain-specific corpus of 3.1B words to the training data (as done by SciBERT) is sufficient and leads to similar improvements as adding 18B words (as done by BioBERT).

The performance for the second version of the primary outcomes corpus is very close to the performance of corresponding models for the first version of the task. These results show that deep pre-trained language representations successfully handle the extraction of coordinated entities, which makes it a promising approach to extraction of secondary outcomes, most often represented by coordinated syntactic groups.

It is difficult to compare our results directly with the previous works on outcome extraction: all the works used corpora that vary in volume and that were built on different principles regarding the selection of sentences and text fragments to annotate; besides, evaluation in different approaches was performed on different level (sentence, entity, token). Taking into account this limitations, we can still state that our results are better than the reported results in the previous comparable works[27, 28, 29, 30, 31]. To allow for transparency and reproducibility of outcome extraction, we released our corpus annotated for declared (primary) and reported outcomes. It has, however, some limitations: e.g., annotating only explicit definitions of declared primary outcomes; annotating reported outcome in abstracts only; annotation done by one annotator.

6 Conclusions

Automatic extraction of primary and reported outcomes of clinical trials is a vital task for automating systematic review process, clinical question answering, and assessment of biomedical articles for bias and spin.

We proposed a deep learning approach to trial outcome extraction and tested a number of pre-trained language representations. Our results show that language models pre-trained on large general-domain corpora can be successfully employed for extracting complex and varied entities, even with limited amount of domain specific training data. Pre-training language models on domain-specific data further improves the performance. Our approach does not require manual feature engineering or any other task-specific settings.

7 Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*, 2019.

[3] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. 2019.

[4] D. Demner-Fushman, B. Few, S.E. Hauser, and G. Thoma. Automatically identifying health outcome information in medline records. *Journal of the American Medical Informatics Association*, 13:52–60, 2006.

[5] Catherine Blake and Rebecca Kehm. Comparing breast cancer treatments using automatically detected surrogate and clinically relevant outcomes entities from text. *Journal of Biomedical Informatics: X*, 1:100005, 2019.

[6] Siddhartha Reddy Jonnalagadda, Pawan Goyal, and Mark D Huffman. Automating data extraction in systematic reviews: a systematic review. In *Systematic reviews*, 2015.

[7] Dina Demner-Fushman and Jimmy Lin. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103, 2007.

[8] Julian P T Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savović, Kenneth F Schulz, Laura Weeks, Jonathan A C Sterne, Cochrane Bias Methods Group, and Cochrane Statistical Methods Group. The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343, 2011.

[9] B. Goldacre, H. Drysdale, A. Powell-Smith, A. Dale, I. Milosevic, E. Slade, P. Hartley, C. Marston, K. Mahtani, and C. Heneghan. The compare trials project. 2016.

[10] Isabelle Boutron, Susan Dutton, Philippe Ravaud, and Douglas Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 303(20):2058–2064, 05 2010.

[11] Isabelle Boutron, Douglas Altman, Sally Hopewell, Francisco Vera-Badillo, Ian Tannock, and Philippe Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the spiin randomized controlled trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 32, 11 2014.

[12] Romana Haneef, Clement Lazarus, Philippe Ravaud, Amelie Yavchitz, and Isabelle Boutron. Interpretation of results of studies evaluating an intervention highlighted in google health news: A cross-sectional study of news. *PloS one*, 10:e0140889, 10 2015.

[13] Amélie Yavchitz, Isabelle Boutron, Aida Bafeta, Ibrahim Marroun, Pierre Charles, Jean Mantz, and Philippe Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLOS Medicine*, 9(9):1–11, 09 2012.

[14] Ritu Khare, Robert Leaman, and Zhiyong lu. Accessing biomedical literature in the current information landscape. *Methods in molecular biology (Clifton, N.J.)*, 1159:11–31, 05 2014.

[15] W.S. Richardson, M.C. Wilson, J. Nishikawa, and R.S. Hayward. The well-built clinical question: a key to evidence-based decisions. *ACP J Club*, 123(3), 1995.

[16] Dina Demner-Fushman and Jimmy Lin. Knowledge extraction for clinical question answering : Preliminary results. In *Proc of the AAAI 2005 Workshop on Question Answering in Restricted Domains*, 2005.

[17] Florian Boudin, Jian-yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. Combining classifiers for robust pico element detection. *BMC medical informatics and decision making*, 10:29, 05 2010.

[18] Ke-Chun Huang, Charles Liu, Shung-Shiang Yang, Furen Xiao, Jau-Min Wong, Chun-Chih Liao, and I-Jen Chiang. Classification of pico elements by text features systematically extracted from pubmed abstracts. In *Proceedings - 2011 IEEE International Conference on Granular Computing, GrC 2011*, pages 279–283, 11 2011.

[19] Su Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. Automatic classification of sentences to support evidence based medicine. *BMC bioinformatics*, 12 Suppl 2:S5, 03 2011.

[20] Yong gang Cao, James J. Cimino, John Ely, and Hong Yu. Automatically extracting information needs from complex clinical questions. *Journal of Biomedical Informatics*, 43(6):962 – 971, 2010.

[21] Mathias Verbeke, Vincent Van Asch, Roser Morante, Paolo Frasconi, Walter Daelemans, and Luc De Raedt. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learnin*, pages 579–589, 07 2012.

[22] Ke-Chun Huang, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. Pico element detection in medical text without metadata: Are first sentences enough? *Journal of Biomedical Informatics*, 46(5):940 – 946, 2013.

[23] Hamed Hassanzadeh, Tudor Groza, and Jane Hunter. Identifying scientific artefacts in biomedical literature: The evidence based medicine use case. *Journal of Biomedical Informatics*, 49:159 – 170, 2014.

[24] Di Jin and Peter Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[25] B. De Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim. Automated information extraction of key trial design elements from clinical trial publications. In *AMIA Annual Symposium*, 2008.

[26] S. Kiritchenko, B. De Bruijn, S. Carini, J. Martin, and I. Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC Med Inform Decis Mak.*, 2010.

[27] R.L. Summerscales, S. Argamon, J. Hupert, and A. Schwartz. Identifying treatments, groups, and outcomes in medical abstracts. In *Sixth Midwest Computational Linguistics Colloquium (MCLC)*, 2009.

[28] Rodney L. Summerscales, Shlomo Engelson Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. Automatic summarization of results from clinical trials. *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377, 2011.

[29] C. Blake and A. Lucic. Automatic endpoint detection to support the systematic review process. *J. Biomed. Inform*, 56:42–56, 2015.

[30] A. Lucic and C. Blake. Improving endpoint detection to support automated systematic reviews. In *AMIA Annu Symp Proc.*, page 1900–1909, 2016.

[31] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[32] Anna Koroleva and Patrick Paroubek. Demonstrating construkt, a text annotation toolkit for generalized linguistic contructions applied to communication spin. In *The 9th Language and Technology Conference (LTC 2019) Demo Session*, 2019.

[33] D. Rennie. Consort revised – improving the reporting of randomized trials. *JAMA*, 285:2006–7, 2001.

[34] Moher D Schulz KF, Altman DG. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, 8 18, 2010.

[35] Anna Koroleva. Annotated corpus for primary and reported outcomes extraction, May 2019.

[36] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proc. of EMNLP 2015*, pages 1373–1378, Lisbon, Portugal, September 2015. ACL.

[37] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August 2016. ACL.

[38] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. of EMNLP 2014*, pages 1532–1543, Doha, Qatar, October 2014. ACL.

[39] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

[40] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.