Extracting relations between outcomes and significance levels in Randomized Controlled Trials (RCTs) publications

Anna Koroleva, Patrick Paroubek

Extracting relations between outcomes and significance levels in Randomized Controlled Trials (RCTs) publications

Anna Koroleva
LIMSI, CNRS, Université Paris-Saclay,
F-91405 Orsay, France
Academic Medical Center, University of Amsterdam,
Amsterdam, the Netherlands
koroleva@limsi.fr

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

Abstract

Randomized controlled trials assess the effects of an experimental intervention by comparing it to a control intervention with regard to some variables - trial outcomes. Statistical hypothesis testing is used to test if the experimental intervention is superior to the control. Statistical significance is typically reported for the measured outcomes and is an important characteristic of the results. We propose a machine learning approach to automatically extract reported outcomes, significance levels and the relation between them. We annotated a corpus of 663 sentences with 2,552 outcome - significance level relations (1,372 positive and 1,180 negative relations). We compared several classifiers, using a manually crafted feature set, and a number of deep learning models. The best performance (F-measure of 94%) was shown by the BioBERT fine-tuned model.

1 Introduction

In clinical trials, outcomes are the dependent variables that are monitored to assess how they are influenced by other, independent, variables (treatment used, dosage, patient characteristics). Outcomes are a central notion for clinical trials.

To assess the impact of different variables on the outcomes, statistical hypothesis testing is commonly used, giving an estimation of statistical significance – the likelihood that a relationship between two or more variables is caused by something other than a chance (Schindler, 2015). Statistical significance levels are typically reported along with the trial outcomes as p-values, with a certain set threshold, where a p-value below the threshold means that the results are statistically significant, while a p-value above the threshold presents non-significant results. Hypothesis testing in clinical trials is used in two main cases:

1. In a trial comparing several treatments given to different groups of patients, a difference in value of an outcome observed between the groups at the end of the trial is evaluated by hypothesis testing to determine if the difference is due to the difference in medication. If the difference is statistically significant, the null hypothesis (the difference between treatments is due to a chance) is rejected, i.e. the superiority of one treatment over the other is considered to be proved.

2. When an improvement of an outcome is observed within a group of patients taking a treatment, hypothesis testing is used to determine if the difference in the outcome at different time points within the group is due to the treatment. If the results are statistically significant, it is considered to be proven that the treatment has a positive effect on the outcome in the given group of patients.

Although p-values are often misused and misinterpreted (Head et al., 2015), extracting significance levels for trial outcomes is still vital for a number of tasks, such as systematic reviews, detection of bias and spin. In particular, our application of interest is automatic detection of spin, or distorted reporting of research results, that consists in presenting an intervention studied in a trial as having higher beneficial effects than the research has proved. Spin is an alarming problem in health care as it causes overestimation of the intervention by clinicians (Boutron et al., 2014) and unjustified positive claims regarding the intervention is health news and press releases (Haneef et al., 2015; Yavchitz et al., 2012).

Spin is often related to a focus on significant outcomes, and occurs when the primary outcome (the main variable monitored during a trial) is not significant. Thus, to detect spin, it is important to identify the significance of outcomes, and espe-

Proceedings of the BioNLP 2019 workshop, pages 359–369
Florence, Italy, August 1, 2019.

cially of the primary outcome. To our best knowledge, no previous work addressed the extraction of the relation between outcomes and significance levels. In this paper, we present our approach towards extracting outcomes, significance levels and relations between them, that can be incorporated into a spin detection pipeline.

2 State of the art

Extraction of outcome - significance level relations consists of two parts: entity extraction (reported outcomes and significance levels) and extraction of the relationship between the entities. In this section, we present the previous works on these or similar tasks.

2.1 Entity extraction

The number of works addressing automatic extraction of significance levels is limited.

(Hsu et al., 2012) used regular expressions to extract statistical interpretation, p-values, confidence intervals, and comparison groups from sentences categorized as "outcomes and estimation". The authors report precision of 93%, recall of 88% and F-measure of 90% for this type of information.

(Chavalarias et al., 2016) applied text mining to evaluate the p-values reported in the abstracts and full texts of biomedical articles published in 1990 – 2015. The authors also assessed how frequently statistical information is presented in ways other than p-values. P-values were extracted using a regular expression; the system was evaluated on a manually annotated dataset. The reported sensitivity (true positive rate) is 96.3% and specificity (true negative rate) is 99.8%. P-values and qualitative statements about significance were more common ways of reporting significance than confidence intervals, Bayes factors, or effect sizes.

A few works focused on extracting outcome-related information, addressing it either as a sentence classification, or as entity extraction task.

(Demner-Fushman et al., 2006) defined an outcome as *"The sentence(s) that best summarizes the consequences of an intervention"* and thus adopted a sentence classification approach to extract outcome-related information from medical articles, using a corpus of 633 MEDLINE citations. The authors tested Naive Bayes, linear SVM and decision-tree classifiers. Naive Bayes showed the best performance. The reported classification accuracy ranged from 88% to 93%.

One of the notable recent works addressing outcome identification as an entity extraction task, rather than sentence classification, is (Blake and Lucic, 2015). The authors addressed a particular type of syntactic constructions – comparative sentences – to extract three items: the compared entities, referred to as the agent and the object, and the ground for comparison, referred to as the endpoint (synonymous to outcome). The aim of this work was to extract corresponding noun phrases. The dataset was based on full-text medical articles and included only the sentences that contain all the three entities (agent, object and endpoint). The training set comprised 100 sentences that contain 656 noun phrases. The algorithm proceeds in two steps: first, comparative sentences are detected with the help of a set of adjectives and lexico-syntactic patterns. Second, the noun phrases are classified according to their role (agent, object, endpoint) using SVM and generalized linear model (GLM). On the training set, SVM showed better performance than GLM, with an F-measure of 78% for the endpoint. However, on the test set the performance was significantly lower: SVM showed an F-measure of only 51% for the endpoint. The performance was higher on shorter sentences (up to 30 words) than on the longer ones.

A following work (Lucic and Blake, 2016) aimed at improving the recognition of the first entity and of the endpoint. The authors propose to use in the classification the information on whether the head noun of the candidate noun phrase denotes an amount or a measure. The annotation of the corpus was enriched by the corresponding information. As a result, precision of the endpoint detection improved to 56% on longer sentences and 58% on shorter ones; recall improved to 71% on longer sentences and 74% on shorter ones.

2.2 Relation extraction

To our knowledge, extraction of the relation between outcomes and significance levels has not been addressed yet. In this section, we overview some frameworks for relation extraction and outline some common features of different approaches in the biomedical relation extraction.

A substantial number of works addressed extracting binary relations, such as protein-protein interactions or gene-phenotype relation, or com-

plex relations, such as biomolecular events. A common feature of the works in this domain, noted by (Zhou et al., 2014; Lever and Jones, 2017) and still relevant for recent works e.g. (Peng and Lu, 2017; Asada et al., 2017), consists in assuming that entities of interest are already extracted and provided to the relation extraction system as input. Thus, the relation extraction is assessed separately, without taking into account the performance of entity extraction. We adopt this approach for relation extraction evaluation in our work, but we provide separate assessment for our algorithms of entity extraction.

One of the general frameworks for relation extraction in the biomedical domain is proposed by (Zhou et al., 2014). The authors suggest using trigger words to determine the type of a relation, noting that for some relation types trigger words can be extracted simply with a dictionary, while for other types, rule-based or machine-learning approaches may be required. For relation extraction, rule-based methods can be applied, often employing regular expressions using words or POS tags. Rules can be crafted manually or learned automatically. The machine learning approaches to binary relation extraction, as the authors note, usually treat the task as a classification problem. Features for classification often use output of textual analysis algorithms such as POS-tagging and syntactic parsing. Machine learning approaches can be divided into feature-based approaches (using syntactic and semantic features) and kernel approaches (calculating similarity between input sequences based on string or syntactic representation of the input). Supervised machine learning is a highly successful approach for binary relation extraction, but its main drawback consists in the need of large amount of annotated data.

A framework for pattern-based relation extraction is introduced by (Peng et al., 2014). The approach aims at reducing the need for manual annotation. The approach is based on a user-provided list of trigger words and specifications (the definition of arguments for each trigger). Variations of lexico-syntactic patterns are derived using this information and are matched with the input text, detecting the target relations. Some interesting features of the framework include the following: the use of text simplification to avoid writing rules for all existing constructions; the use of referential relations to find the best phrase referring to an entity. The authors state that their system is characterized by good generalizability due to the use of language properties and not of task-specific knowledge.

A recent work (Björne and Salakoski, 2018) reports on the development of convolutional neural networks (CNNs) for event and relation extraction, using Keras (Chollet et al., 2015) with Tensorflow backend (Abadi et al., 2016). Parallel convolutional layers process the input, using sequence windows centered around the candidate entity, relation or event. Vector space embeddings are built for input tokens, including features such as word vectors, POS, entity features, relative position, etc. The system was tested on several tasks and showed improved performance and good generalizability.

3 Our dataset

3.1 Corpus creation and annotation

In our previous work on outcome extraction, we manually annotated a corpus for reported outcomes comprising 1,940 sentences from the Results and Conclusions sections of PMC article abstracts. We used this corpus as a basis for a corpus with annotations for outcome significance level relations.

Our corpus contains 2,551 annotated outcomes. Out of the sentences with outcomes, we selected those where statistical significance levels are supposedly reported (using regular expressions) and manually annotated relations between outcomes and significance levels. The annotation was done by one annotator (AK), in consultation with a number of domain experts, due to infeasibility of recruiting several annotators with sufficient level of expertise within a reasonable time frame.

The final corpus contains 663 sentences with 2,552 annotated relations, out of which 1,372 relations are positive (the significance level is related to the outcome) and 1,180 relations are negative (the significance level is not related to the outcome). The corpus is publicly available (Anna, 2019).

3.2 Data description

There are three types of data relevant for this work: outcomes, significance levels, and relationship between them. In this section, we describe these types of data and the observed variability in the ways of presenting them.

1. Outcomes

A trial outcome is, in broad sense, a measure or variable monitored during a trial. It can be binary (presence of a symptom or state), numerical ("*temperature*") or qualitative ("*burden of disease*"). Apart from the general term denoting the outcome, there are several aspects that define it: a measurement tool (questionnaire, score, etc.) used to measure the outcome; time points at which the outcome is measured; patient-level analysis metrics (change from baseline, time to event); population-level aggregation method (mean, median, proportion of patients with some characteristic).

Generally, there are two main contexts in which outcomes of a clinical trial can be mentioned: a definition of what the outcomes of a trial were ("***Quality of life*** *was selected as the primary outcome*."), and reporting results for an outcome ("***Quality of life*** *was higher in the experimental group than in the control group.*"). In both cases, a mention of an outcome may contain the aspects listed above, but does not necessarily include all of them. In this work, we are interested in the second type of context.

The ways of reporting outcomes are highly diverse. Results for an outcome may be reported as a value of the outcome measure: for binary outcomes, it refers to presence/absence of an event or state; for numerical outcome, it is a numerical value; for qualitative outcome, it is often a value obtained on the associated measurement tool. As the primary goal of RCTs is to compare two or more interventions, results for an outcome can be reported as a comparison between the interventions/patient groups, with or without actual values of the outcome measure. Syntactically, an outcome may be represented by a noun phrase, a verb phrase, an adjective or a clause. We provide here some examples of outcome reporting, to give an idea of variability of expressions.

The outcome is reported as a numerical value:

a) *The median* ***progression-free survival*** *was 32 days.*

The outcome is reported as a comparison between groups, without the values for groups:

b) *MMS resulted in more* ***stunting*** *than standard Fe60F ($p = 0.02$).*

The outcome is reported as a numerical value with comparison between groups:

c) *The average* ***birth weight*** *was 2694 g and* ***birth length*** *was 47.7 cm, with no difference among intervention groups.*

d) *The crude incidence of* ***late rectal toxicity*** $\geq$ ***G2*** *was 14.0% and 12.3% for the arm A and B, respectively.*

e) *More than 96% of patients who received DPT were* ***apyrexial*** *48 hours after treatment compared to 83.5% in the AL group ($p < 0.001$).*

f) *The proportion of patients who* ***remained relapse-free at Week 26*** *did not differ significantly between the placebo group (5/16, 31%) and the IFN beta-1a 44 mcg biw (6/17, 35%; p = 0.497), 44 mcg tw (7/16, 44%; p = 0.280) or 66 mcg tw (2/18, 11%; p = 0.333) groups.*

In the latter case, the variation is especially high, and the same outcome may be reported in several different ways (cf. the examples **d**, **e** and **f** that all talk about a percentage of patients in which a certain event occurred, but the structure of the phrases differs).

Identifying the textual boundaries of an outcome presents a challenge: for the example **d**, it can be "*the crude incidence of late rectal toxicity* $\geq$ *G2*" or "*late rectal toxicity* $\geq$ *G2*"; for the example **f**, it can be "*the proportion of patents who remained relapse-free at Week 26*", or "*remained relapse-free at Week 26*", or simply "*relapse-free*". This variability poses difficulties for both annotation and extraction of reported outcomes. In our annotation, we aimed at annotating the minimal possible text span describing an outcome, not including time points, aggregation and analysis metrics.

2. Significance levels

   The ways of presenting significance levels are less diverse than the ways of reporting outcomes. Typically, significance levels are reported via p-values. Another way of determining significance of the results is the confidence interval (CI), where a CI comprising

zero denotes non-significant results. In this work, we do not address CIs as they are less frequently reported (Chavalarias et al., 2016).

Statistical significance can be reported as an exact value of P ("*p=0.02*"), as P-value relative to a pre-set threshold ("*p<0.05*"), or in qualitative form ("*significant*"/"*non-significant*"). We address all these forms of reporting significance.

Although in general the ways of presenting statistical significance are rather uniform, there are a few cases to be noted:

- Coordinated p-values:
  For the non-HPD stratum, the intent-to-treat relative risks of spontaneous premature birth at < 34 and < 37 weeks' gestation were 0.33 (0.03, 3.16) and 0.49 (0.17, 1.44), respectively, and they were non-significant (ns) with ***p = 0.31 and 0.14****.*
- Significance level in score of a negation:
  The respiratory rate, chest indrawing, cyanosis, stridor, nasal flaring, wheeze and fever in both groups recorded at enrollment and parameters ***did not differ significantly*** *between the two groups.*
  A particular difficulty is presented by the cases in which a negation marker occurs in the main clause and a significance level in the dependent clause, thus the significance level is within the scope of the negation, but there is a big linear distance between them:
  Results There was ***no evidence*** *that an incentive (52% versus 43%, Risk Difference (RD) -8.8 (95%CI 22.5, 4.8); or abridged questionnaire (46% versus 43%, RD 2.9 (95%CI 16.5, 10.7);* ***statistically significantly*** *improved dentist response rates compared to a full length questionnaire in RCT A.*

3. Relationship between outcomes and significance levels

   The correspondence between outcomes and significance levels in a sentence is often not one-to-one: multiple outcomes can be linked to the same significance level, and vice versa. Several outcomes are linked to one significance level when outcomes are coordinated:
   No significant *improvements in* ***lung function****,* ***symptoms****, or* ***quality of life*** *were seen.*

   Several significance levels can be associated to one outcome in a number of cases:

   - one outcome is linked to two significance levels when a significance level is presented in both qualitative and numerical form:
     Results ***The response rates*** *were* ***not significantly*** *different Odds Ratio 0.88 (95% confidence intervals 0.48 to 1.63)* ***p = 0.69****.*
   - in the case of comparison between patient groups taking different medications, when there are more than 2 groups, significance can be reported for all pairs of groups;
   - significance level for difference observed within groups of patients receiving a particular medication:
     [***Na***] *increased* ***significantly*** *in the 0.9% group (+0.20 mmol/L/h [IQR +0.03, +0.4];* ***P = 0.02****) and increased, but* ***not significantly****, in the 0.45% group (+0.08 mmol/L/h [IQR -0.15, +0.16];* ***P = 0.07****).*
   - significance reported for both between- and within-group comparison:
     PTEF *increased* ***significantly*** *both after albuterol and saline treatments but the difference between the two treatments was* ***not significant*** *(****P = 0.6****).*
   - significance for differences within subgroups of patients (e.g. gender or age subgroups) receiving a medication;
   - significance for different types of analysis: intention-to-treat / per protocol:
     Results For ***BMD****,* ***no*** *intent-to-treat analyses were* ***statistically significant****; however, per protocol analyses (ie, only including TC participants who completed ≥ 75% training requirements) of* ***femoral neck BMD*** *changes were* ***significantly*** *different between TC and UC (+0.04 vs -0.98%; P = 0.05).*
   - significance for several time points:
     Results A ***significant*** *main effect of time (p < 0.001) was found for* ***step-counts*** *attributable to significant increases in steps/day between: pre-intervention (M*

= 6941, SD = 3047) and 12 weeks (M = 9327, SD = 4136), t (78) = - 6.52, ***p*** *<* ***0.001****, d = 0.66; pre-intervention and 24 weeks (M = 8804, SD = 4145), t (78) = - 4.82,* ***p*** *<* ***0.001****, d = 0.52; and pre-intervention and 48 weeks (M = 8450, SD = 3855), t (78) = - 4.15,* ***p*** *<* ***0.001****, d = 0.44.*

- significance level for comparison of various analysis metrics (mean, AUC, etc.)

4 Methods

To extract the relation between an outcome and its significance level, we propose a 3-step algorithm: 1) extracting reported outcomes; 2) extracting significance levels; 3) classification of pairs of outcomes and significance levels to detect those related to each other.

As significance levels are not characterized by high variability, we follow the previous research in using rules (regular expressions and sequential rules using information from pos-tagging) to extract significance levels.

We present our methods and results for outcome extraction in detail elsewhere, here we provide a brief summary. We tested several approaches: a baseline approach using sequential rules using information from pos-tagging; an approach using rules based on syntactic structure provided by spaCy dependency parser (Honnibal and Johnson, 2015); a combination of bi-LSTM, CNN and CRF using GloVe(Pennington et al., 2014) word embeddings and character-level representations (Ma and Hovy, 2016); and a fine-tuned bi-LSTM using BERT (Devlin et al., 2018) vector word representations.

BERT (Bidirectional Encoder Representations from Transformers) is a recently introduced approach to pre-training language representations, using a masked language model (MLM) which randomly masks some input tokens, allowing to pre-train a deep bidirectional Transformer using both left and right context. The pre-trained BERT models can be fine-tuned for supervised downstream tasks by adding one output layer.

BERT was trained on a dataset of 3.3B words combining English Wikipedia and BooksCorpus. Two domain-specific versions of BERT are available, pre-trained on a combination of the initial BERT corpus and additional domain-specific datasets: BioBERT (Lee et al., 2019), adding a large biomedical corpus of PubMed abstracts and PMC full-text articles comprising 18B tokens; and SciBERT (Beltagy et al., 2019), adding a corpus of 1.14M full-text papers from Semantic Scholar with the total of 3.1B tokens. Both BioBERT and SciBERT outperform BERT on biomedical tasks.

BERT provides several models: uncased (trained on lower-cased data) and cased (trained on unchanged data); base and large (differing in model sizes). BioBERT is based on the BERT-base cased model and provides three versions of models: pre-trained on PubMed abstracts, on PMC full-text articles, or on combination of both. SciBERT has both cased and uncased models and provides two versions of vocabulary: BaseVocab (the initial BERT vocabulary) and SciVocab (the vocabulary from the SciBERT corpus). We fine-tuned and tested the BioBERT model trained on the whole corpus, and both cased and uncased base models for BERT and SciBERT (using SciVocab). We did not perform experiments with BERT-Large as we do not have enough resources. We used the code provided by BioBERT for the entity extraction task[1].

The relation extraction assumes that the entities have already been extracted and are given as an input to the algorithm, with the sentence in which they occur. To predict the tag for outcome - significance level pair, we use machine learning.

As the first approach, we compared several classifiers available in the Python scikit-learn library (Pedregosa et al., 2011): Support Vector Machine (SVM) (Cortes and Vapnik, 1995); DecisionTreeClassifier (Rokach and Maimon, 2008); MLPClassifier(von der Malsburg, 1986); KneighborsClassifier (Altman, 1992); GaussianProcessClassifier (Rasmussen and Williams, 2005); RandomForestClassifier (Breiman, 2001); AdaBoostClassifier (Freund and Schapire, 1997); ExtraTreesClassifier (Geurts et al., 2006); GradientBoostingClassifier (Friedman, 2002). Feature engineering was performed manually and was based on our observations on the corpus.

Evaluation was performed using 10-fold cross-validation. To account for different random states, the experiments were run 10 times, we report the average results of the 10 runs. We performed hyperparameters tuning via exhaustive grid search (with the help of the scikit-learn GridSearchCV

[1] https://github.com/dmis-lab/biobert/blob/master/run_ner.py

function).

As the second approach, we employed a deep learning approach to relation extraction, fine-tuning BERT-based models on this task. We tested the same models as for the outcome extraction. We used the code provided by BioBERT for relation extraction task[2]. The algorithm takes as input sentences with the two target entities replaced by masks ("@outcome$" and "@significance$") and positive/negative relation labels assigned to the sentence.

Hyperparameters for entity and relation extraction with BERT-based algorithms are shown in the Table 1. We tested both possible values (True/False) of the hyperparameter "do_lower_case" (lower-casing the input) for all the models.

Hyperparameter	Entity extraction	Relation extraction
max_seq_length	128	
train_batch_size	32	
eval_batch_size	8	
predict_batch_size	8	
use_tpu	False	
learning_rate	5e-5	2e-5
num_train_epochs	10.0	3.0
warmup_proportion	0.1	
save_checkpoints_steps	1000	
iterations_per_loop	1000	
tf.master	None	

Table 1: BERT/BioBERT/SciBERT hyperparameters

5 Features

Features are calculated for each pair of outcome and significance level. They are based both on the information about these entities (their position, text, etc.) and on the contextual information (presence of other entities in the sentence, etc.). We used the following binary (True/False) features:

1. only_out: whether the outcome is the only outcome present in the sentence. If yes, it is the only candidate that can be related to the present statistical significance values.

2. only_signif: whether the significance level is the only significance level in the sentence. If yes, it is the only candidate that can be related to the present outcomes.

3. signif_type_num: whether the significance level is expressed in the numerical form;

[2]https://github.com/dmis-lab/biobert/blob/master/run_re.py

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
Bi-LSTM-CNN-CRF		51.12	44.6	47.52
Rule-based		26.69	55.73	36.09

Table 2: Reported outcome extraction results

4. signif_type_word: whether the significance level is expressed in the qualitative form;

5. signif_exact: whether the exact value of significance level is given ($P = 0.049$), or it is presented only as comparison to a threshold ($P < 0.05$). Significance levels expressed in the word form always have "False" value for this feature. We assumed that significance levels with exact numerical value are less likely to be related to several outcomes that significance levels with inexact value: obtaining exactly same significance level for several outcomes seems unlikely.

6. signif_precedes: whether the significance level precedes the outcome. It is especially pertinent for numerical significance values as they most often follow the related outcome.

7. out_between: whether there is another outcome between the outcome and significance level in the given pair. The outcome that is closer to a significance level is a more likely candidate to be related to it.

8. signif_between: whether there is another significance level between the outcome and the significance level in a given pair. The significance level that is closer to an outcome is a more likely candidate to be related to it.

9. concessive_between: whether there are words

Classifier	Hyperparameters	Precision	Recall	F1
RandomForestClassifier	max_depth = 15, min_samples_split = 10, n_estimators = 300	90.16	92.6	91.33
ExtraTreesClassifier	default	89.74	88.53	89.08
GradientBoostingClassifier	learning_rate = 0.25, max_depth = 23.0, max_features = 7, min_samples_leaf = 0.1, min_samples_split = 0.2, n_estimators = 200	88.44	89.8	89.07
RandomForestClassifier	default	89.54	88.64	89.03
GaussianProcessClassifier	1.0 * RBF(1.0)	86.99	90.38	88.64
GradientBoostingClassifier	default	87.75	89.14	88.4
SVC	C = 1000, gamma = 0.0001, kernel = 'rbf'	86.14	89.65	87.79
DecisionTreeClassifier	default	87.85	86.83	87.27
MLPClassifier	activation = 'tanh', alpha = 0.0001, hidden_layer_sizes = (50, 100, 50), learning_rate = 'constant', solver = 'adam'	84.06	85.15	84.44
MLPClassifier	default	84.4	83.34	83.47
KNeighborsClassifier	n_neighbors = 7, p = 1	83.37	81.27	82.21
AdaBoostClassifier	learning_rate = 0.1, n_estimators = 500	81.34	83.09	82.16
AdaBoostClassifier	default	80.85	82.36	81.53
KNeighborsClassifier	default	81.39	79.88	80.55
GaussianProcessClassifier	default	79.41	78.86	79.1
SVC	default	87.24	64.06	73.77
baseline (majority class)		53.76	100	69.92

Table 3: Results of classifiers

Feature	Weight
only_signif	0.21663222
signif_type_num	0.21341347
signif_exact	0.15207938
signif_type_word	0.10103105
dist_min_out_preceding	0.0919397
out_between	0.05683003
dist_min_out_following	0.04683059
concessive_between	0.04260114
only_out	0.02336161
dist	0.02043495
dist_min_graph	0.01794923
signif_precedes	0.01631646
signif_between	0.00058017

Table 4: Feature ranking

Algorithm	do_lower_case	Precision	Recall	F1
BioBERT	True	94.3	94	94
SciBERT cased	True	93.9	93.6	93.8
SciBERT cased	False	93.5	93.1	93.3
SciBERT uncased	False	94.2	92.3	93.3
SciBERT uncased	True	94	92.8	93.2
BioBERT	False	92.8	89.7	91.1
BERT cased	False	91.6	90.2	90.9
BERT uncased	True	90.9	90.9	90.8
BERT uncased	False	90.4	89.8	90
BERT cased	True	89.6	90.5	89.8

Table 5: Results of relation extraction with BERT/BioBERT/SciBERT

(conjunctions) with consessive semantics (*but*, *however*, *although*, etc.) between the outcome and the significance level in the pair.

We used the following numerical features:

1. dist: the distance in characters between the outcome and the significance level in the pair;

2. dist_min_graph: the minimal syntactic distance between the words in the outcome and the words in the significance level;

3. dist_min_out_preceding: the distance from the outcome of the pair to the nearest preceding outcome.

4. dist_min_out_following: the distance from the outcome of the pair to the nearest following outcome. The two last features are designed to reflect the information about coordination of outcomes (the distances between

coordinated entities is typically small), as coordinated outcomes are likely to be related to the same significance level.

We assessed the importance of the features with the attribute "feature_importances_" of the RandomForestClassifier classifier. The results are presented in the Table 4.

6 Evaluation

6.1 Entity extraction

The rule-based extraction of significance levels shows the following per-token performance: precision of 99.18%, recall of 96.58% and F-measure of 97.86%.

The results of all the tested approaches to the extraction of reported outcomes are reported in the Table 2. The best performance was achieved by the fine-tuned SciBERT uncased model: precision was 81.17%, recall was 78.09% and F-measure was 79.42%.

6.2 Relation extraction

The baseline value is based on assigning the majority (positive) class to all the entity pairs. Baseline precision is 53.76%, recall is 100% and F-measure is 69.95%.

The results of the classifiers are presented in the Table 3. We present the performance of the default classifiers and of the classifiers with tuned hyperparameters. All the classifiers outperformed the baseline. Random Forest Classifier with tuned hyperparameters (max_depth = 15, min_samples_split = 10, n_estimators = 300) showed the best results, with F-measure of 91.33%, which is by 21.41% higher than the baseline.

It is interesting to compare the deep learning approach using BERT-based fine-tuned models (Table 5) to the feature-based classifiers: none of the Google BERT models outperformed the Random Forest Classifier, neither did BioBERT with unchanged input data. However, all the SciBERT fine-tuned models and the BioBERT model with lower-cased input outperformed the Random Forest Classifier. Interestingly, BioBERT, which only has a cased model pre-trained on unchanged data and is thus meant to work with unchanged input, showed the best performance on lower-cased input for the relation extraction task, achieving the F-measure of 94%.

7 Conclusion and future work

In this paper, we presented a first approach towards the extraction of the relation between outcomes of clinical trials and their reported significance levels. We presented our annotated corpus for this task and described the ways of reporting outcomes, significance levels and their relation in a text. We pointed out the difficulties posed by the high diversity of the data.

We crafted a feature set for relation extraction and trained and tested a number of classifiers for this task. The best performance was shown by the Random Forest classifier, with the F-measure of 91.33%. Further, we fine-tuned and evaluated a few deep learning models (BERT, SciBERT, BioBERT). The best performance was achieved by the BioBERT model fine-tuned on lower-cased data, with F-measure of 94%.

Our relation extraction algorithm assumes that the entities have been previously extracted and provided as input. An interesting direction for future experiments is building an end-to-end system extracting both entities and relations, as proposed by (Miwa and Bansal, 2016) or (Pawar et al., 2017).

As in our algorithm the extraction of the relevant entities (reported outcomes and significance levels) is essential for extracting the relations, we reported the results of our experiments for extracting this task. Extraction of significance levels reaches the F-measure of 97.86%, while the extraction of reported outcomes shows the F-measure of only 79.42%. Thus, improving the outcome extraction is the main direction of the future work.

Besides, a very important task for clinical trial data analysis consists in determining the significance level for the primary outcome. This task requires two additional steps: 1) identifying the primary outcome, and 2) establishing the correspondence between the primary outcome and a reported outcome. We will present our algorithms for these tasks in a separate paper.

8 Acknowledgements

We thank Sanjay Kamath for his help in conducting experiments with BERT.

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

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. 2016. Tensorflow: A system for large-scale machine learning.

N.S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician - AMER STATIST*, 46:175–185.

Koroleva Anna. 2019. Annotated corpus for the relation between reported outcomes and their significance levels.

Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2017. Extracting drug-drug interactions with attention CNNs. In *BioNLP 2017*, pages 9–18, Vancouver, Canada,. Association for Computational Linguistics.

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

Jari Björne and Tapio Salakoski. 2018. Biomedical event extraction using convolutional neural networks and dependency parsing. In *BioNLP 2018 workshop*, pages 98–108, Melbourne, Australia. ACL.

C. Blake and A. Lucic. 2015. Automatic endpoint detection to support the systematic review process. *J. Biomed. Inform.*

Isabelle Boutron, Douglas Altman, Sally Hopewell, Francisco Vera-Badillo, Ian Tannock, and Philippe Ravaud. 2014. 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*.

Leo Breiman. 2001. Random forests. *Mach. Learn.*, 45(1):5–32.

David Chavalarias, Joshua D Wallach, Alvin Ho Ting Li, and John P. A. Ioannidis. 2016. Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA*, 315 11:1141–8.

François Chollet et al. 2015. Keras.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

D. Demner-Fushman, B. Few, S.E. Hauser, and G. Thoma. 2006. Automatically identifying health outcome information in medline records. *Journal of the American Medical Informatics Association*.

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

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139.

Jerome H. Friedman. 2002. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4):367–378.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Mach. Learn.*, 63(1):3–42.

R. Haneef, C. Lazarus, P. Ravaud, A. Yavchitz, and I. Boutron. 2015. Interpretation of results of studies evaluating an intervention highlighted in google health news: a cross-sectional study of news. *PLoS ONE*.

Megan L Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The extent and consequences of p-hacking in science. In *PLoS biology*.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proc. of EMNLP 2015*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

William Hsu, William Speier, and Ricky K. Taira. 2012. Automated extraction of reported statistical analyses: Towards a logical representation of clinical trial literature. *AMIA Annual Symposium*, 2012:350–359.

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

Jake Lever and Steven Jones. 2017. Painless relation extraction with kindred. In *BioNLP 2017*, pages 176–183, Vancouver, Canada,. Association for Computational Linguistics.

A. Lucic and C. Blake. 2016. Improving endpoint detection to support automated systematic reviews. In *AMIA Annu Symp Proc.*

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

Christoph von der Malsburg. 1986. Frank rosenblatt: Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. *Brain Theory*, pages 245–248.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.

Sachin Pawar, Pushpak Bhattacharyya, and Girish Palshikar. 2017. End-to-end relation extraction using neural networks and Markov logic networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 818–827, Valencia, Spain. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Yifan Peng and Zhiyong Lu. 2017. Deep learning for extracting protein-protein interactions from biomedical literature. In *BioNLP 2017*, pages 29–38, Vancouver, Canada,. Association for Computational Linguistics.

Yifan Peng, Manabu Torii, Cathy Wu, and K Vijay-Shanker. 2014. A generalizable nlp framework for fast development of pattern-based biomedical relation extraction systems. *BMC bioinformatics*, 15:285.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP 2014*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Lior Rokach and Oded Maimon. 2008. *Data Mining with Decision Trees: Theroy and Applications*. World Scientific Publishing Co., Inc., River Edge, NJ, USA.

Thomas M. Schindler. 2015. Hypothesis testing in clinical trials. *AMWA Journal*, 30(2).

A. Yavchitz, I. Boutron, A. Bafeta, I. Marroun, P. Charles, J. Mantz, and P. Ravaud. 2012. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLoS Med*.

Deyu Zhou, Dayou Zhong, and Yulan He. 2014. Biomedical relation extraction: From binary to complex. In *Comp. Math. Methods in Medicine*.