Text Classification for Identifying Eligibility Criteria in Clinical Trial Protocols using Pre-trained Deep Learning Models

Miao Chen[1], PhD; Victor Lobanov[1], PhD; Aaron Mackey[2], PhD
[1]Covance, Princeton, NJ, USA; [2] Flagship, USA

Background

Information extraction from clinical protocols

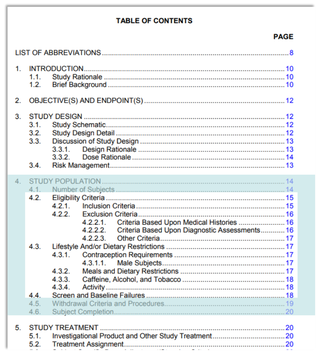


TABLE OF CONTENTS

PAGE

LIST OF ABBREVIATIONS 8

1. INTRODUCTION 10
   1.1. Study Rationale 10
   1.2. Brief Background 10
2. OBJECTIVE(S) AND ENDPOINT(S) 12
3. STUDY DESIGN 12
   3.1. Study Schematic 12
   3.2. Study Design Detail 12
   3.3. Discussion of Study Design 13
      3.3.1. Design Rationale 13
      3.3.2. Dose Rationale 14
   3.4. Risk Management 13
4. STUDY POPULATION 14
   4.1. Number of Subjects 14
   4.2. Eligibility Criteria 15
      4.2.1. Inclusion Criteria 15
      4.2.2. Exclusion Criteria 16
         4.2.2.1. Criteria Based Upon Medical Histories 16
         4.2.2.2. Criteria Based Upon Diagnostic Assessments 16
         4.2.2.3. Other Criteria 17
   4.3. Lifestyle And/or Dietary Restrictions 17
      4.3.1. Contraception Requirements 17
         4.3.1.1. Male Subjects 17
      4.3.2. Meals and Dietary Restrictions 17
      4.3.3. Caffeine, Alcohol, and Tobacco 18
      4.3.4. Activity 18
   4.4. Screen and Baseline Failures 18
   4.5. Withdrawal Criteria and Procedures 19
   4.6. Subject Completion 20
5. STUDY TREATMENT 20
   5.1. Investigational Product and Other Study Treatment 20
   5.2. Treatment Assignment 20

Importance of protocols and downstream tasks of protocol analytics

- For understanding operational requirements
- For understanding systemic challenges
- Implying probability of success
- cost implications and business planning

Challenges

- Clinical protocols = semi-structured or unstructured document
- Lack of training and labeled

Methods

Pretrained on large corpus

- *optimized towards language model and next sentence prediction*

Bi-directional language representations

- *as opposed to unidirectional representation in other pre-trained models*

Context-sensitive deep neural networks

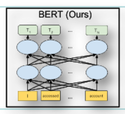
- *compared to context-insensitivity of non-pretrained neural networks*

End-to-end machine learning paradigm

- *Without classical NLP pipelines, e.g., tokenization -> feature engineering -> feature extraction -> machine learning…*

Vocabulary
- Base vocab
- Scibert vocab

Pretrained BERT model
- Bert
- BioBert
- SciBert



Text classification
- Softmax layer

Data & Methods

Eligibility criteria data from clinicaltrials.gov (ct.gov)

- A small subset of ct.gov XML documents
- A XML doc contains multiple fields about a protocol, e.g. eligibility criteria, intervention, condition, etc.
- Used the eligibility criteria (EC) section of the XML docs as positive samples, and other sections (non-EC) as negative samples

Eligibility criteria (EC)	Non-eligibility (non-EC)
Train – 14,000 Test – 6,000	Train - 14000 Test – 6,000

Binary text classification task using pretrained deep learning model

- Using Bidirectional Encoder Representations from Transformers (BERT) as the pretrained language model
- Fine tuning the binary text classification task based on the BERT pretrained model
- Experimented with variants of BERT pretrained models: Bert/BioBert/Scibert

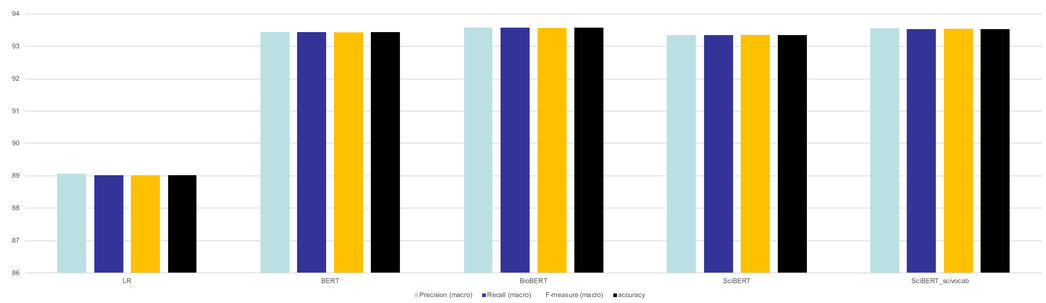
Positive sample

```
<eligibility>
  <criteria>Inclusion Criteria (patient): - Willingness to treat the target lesion according to the DCB only…</criteria>
</eligibility>
```

Negative sample

```
<Intervention_name>drug coated balloon angioplasty<Intervention_name>
</Intervention>
```

Results & Conclusions



- Overall the BERT family models outperform the logistic regression baseline due to the deeper language representation and fine tuning
- The BERT family models perform close to each other on our data set, likely due to the fact it was a small classification data set
- Ongoing NLP efforts in Covance – entity recognition, syntactic relations, multi-task learning

Contact: contact@ohdsi.org

poster url: https://www.ohdsi.org/2019-us-symposium-showcase-32/
BERT image from https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Title here

Authors and affiliations

Abstract

Put text here

Methods

Put text here

Results

Put text here

Background

Put text here

Results

Put text here

Conclusions

Put text here

Contact: contact@ohdsi.org