Extracting Semantic Aspects for Structured Representation of Clinical Trial Eligibility Criteria

[1*]Ishani Mondal ,[*2*]Tirthankar Dasgupta , [2]Abir Naskar, [2] Sudeshna Jana, [2]Lipika Dey
[1] Microsoft Research Labs, India [2]TCS Research and Innovation Labs, India
[1]t-imonda@microsoft.com,
[2][dasgupta.tirthankar, abir.naskar, sudeshna.jana, lipika.dey]@tcs.com

Abstract

Eligibility criteria in the clinical trials specify the characteristics that a patient must or must not possess in order to be treated according to a standard clinical care guideline. As the process of manual eligibility determination is time-consuming, automatic structuring of the eligibility criteria into various semantic categories or aspects is the need of the hour. Existing methods use hand-crafted rules and feature-based statistical machine learning methods to dynamically induce semantic aspects. However, in order to deal with paucity of aspect-annotated clinical trials data, we propose a novel weakly-supervised co-training based method which can exploit a large pool of unlabeled criteria sentences to augment the limited supervised training data, and consequently enhance the performance. Experiments with 0.2M criteria sentences show that the proposed approach outperforms the competitive supervised baselines by 12% in terms of micro-averaged F1 score for all the aspects. Probing deeper into analysis, we observe domain-specific information boosts up the performance by a significant margin.

1 Introduction

Clinical trials (CTs) are research studies that are aimed at evaluating a medical, surgical, or behavioral intervention (Embi et al., 2008), (Shivade et al., 2015). Through such trials, researchers aim to find out whether a new treatment, like a new drug or diet or medical device is more effective than the existing treatments for a particular ailment. From an organization's perspective, a successful completion of a trial depends on achieving a significant sample size of patients enrolled for the trial within a limited time period.

Total bilirubin less than or equal to 1.5 mg/dl, except in patients with history of anaemia. Have had their ileostomy or colostomy for at least 3 months. Subjects must be between the age of 18-65 yr old and must not intake alcohol.

Categories of Semantic Aspects are represented using the colors: Health Status ; Lab Test ; Demography ; Life Style ; Treatment Status

However, recruiting enough number of eligible patients to participate in a trial can be a bottleneck. If suitable patients are not found then the trials might get cancelled or delayed significantly. In this case a patient queries the sites like *clinicaltrial.gov* to retrieve suitable trials. Due to the complexity of the task which involves repeated reading of the patient's Electronic Health Record (EHR) and the trial criteria for multiple trials, this is not only a labor-intensive and time-consuming task but also prone to human errors. In addition to this, the eligibility criteria often uses complex language structures and medical jargons mentioned in either semi-structured or unstructured way.

Previous works (Koopman and Zuccon, 2016) have formulated the problem of retrieving relevant document collection based on patient query. However, we demonstrate an approach in which the primary eligibility aspects are identified initially for further screening of the patients in terms of inclusion or exclusion strategy, which is the first step towards matching patients with the relevant trials.

In this paper, we propose an effective method which automatically identifies and segregates the clinical trial eligibility criteria into five semantic aspects. Also, the criteria texts speak volume about multiple aspects of the patients that includes *demographic information*, *health status*, *treatment history*, *laboratory test reports* and *life-style*. However, there has been a dearth of annotated crite-

The first two authors contributed equally.

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 243–248

ria. Since, prior methods on neural clinical entity recognition models rely on the presence of a large annotated corpora and due to the high cost associated with manual tagging of semantic aspects and limited availability of labeled datasets (Najafabadi et al., 2015), it is difficult to train a deep neural network effectively for such a task. We attempt to combat this difficulty by proposing a novel semi-supervised method based on deep co-training (Blum and Mitchell, 1998) which can harness a large pool of unlabeled clinical trial criteria that are more economical to collect. To the best of our knowledge, we are the first to introduce such a co-training-based method and demonstrate its effectiveness in aspect categorization of clinical trials in comparison to stand-alone sequence-labelling in isolation. The end-product of our experiments is a clinical trial-register that contain details of the different aspects across conditions and interventions.

2 Problem Formulation

Given an eligibility criteria sentence in the form of a word sequence x = (x_1 ...., x_n) , where n is the maximum length of the sequence, the task is to predict an output sequence y = (y_1 , ..., y_n) in which each y_i is encoded using standard sequence labeling encoding scheme. Each y_i might take one of the following aspects :

1. **Health Status (Health)**: describes the present medical condition like pregnancy status, disease affected, etc.

2. **Treatment (Trt)**: contains information about the intervention, surgery or therapy related information of the patients.

3. **Lab-Test (Lab)**: It deals with the lab-tests or experimental results.

4. **Demography(Demo)**: This class primarily deals with the age, gender related to the patients undergoing clinical trials.

5. **Life-Style(Life)**: This class primarily deals with the information of the patients regarding their daily habits like diet, exercise etc

6. **Other**: It contains none of the above classes.

Figure 1 illustrates the overview of semantic aspect extraction.

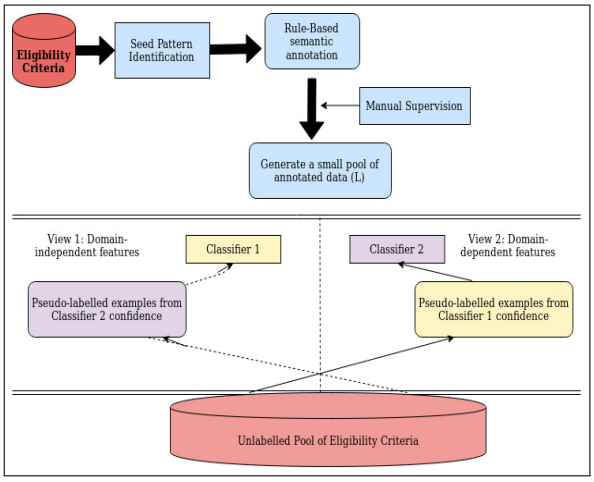

Figure 1: Working Pipeline of Semantic Aspect Extraction from Eligibility Criteria using Co-Training.

	Health	Trt	Lab	Life	Demo
TS	345	323	456	280	600
ASL	8	6	9	8	5

Table 1: Statistics of the Manually annotated Dataset. Here, *TS* indicates the total Number of Sequences for each of the different aspects and *ASL* indicates Average length of each Sequence. *Life* indicates Life-Style aspect, *Trt* indicates Treatment aspect, *Lab* indicated Lab-Test Results and *Demo* indicates demography aspect.

3 Data Annotation

To induce semantic categorization of aspects in the eligibility criteria, we generate a small pool of annotated data by manually examining some of the most frequently used n-gram patterns such as *history of*, *upper limit of normal*, *treated by*, *Allergy to* as specified in (Luo et al., 2011) in the initial phase. During pre-processing, we filter out the most frequently occurring n-grams (n=2, n=3, n=4, n=5) present in the criteria of the patients. Secondly, the criteria sentences are also tagged with CliNER Tagger (Boag et al., 2015) for extracting out the diseases and drugs. Further details of data are provided in the supplementary material [1]. After these two steps, finally, the false positives are being removed during manual supervision by four independent domain-expert annotators. These include annotations for each of the different categories. The mean Cohen's Kappa (McHugh, 2012) was 0.82, which indicate good inter-annotator agree-

[1] https://github.com/Ishani-Mondal/Clinical-Trials-Aspect-Extraction

Algorithm 1 Aspect Extraction using Co-Training Algorithm

Input U : Large amount of unlabelled criteria sentences, τ : Co-Training threshold, V^1 , V^2: Two views of labelled Aspect Annotated Criteria Sentences
Output Model Parameters : $\theta_{\text{BiLSTM-CRF}}$, $\theta_{\text{BiGRU-CRF}}$

```
T^1, T^2 ← V^1, V^2
Initialize the model parameters θ_BiLSTM−CRF, θ_BiGRU−CRF randomly.
while (stopping criteria is not met) do
    C^1 ← Train BiLSTM-CRF on T^1 (minimize the Aspect Loss)
    C^2 ← Train BiGRU-CRF on T^2 (minimize the Aspect Loss)
    for i=1 to |U| do
        if C^1.score(U_i) ≥ τ then
            T^2 ← T^2 ∪ U_i , U = U U_i
        end if
        if C^2.score(U_i) ≥ τ then
            T^1 ← T^1 ∪ U_i , U = U U_i
        end if
    end for
end while
```

ment. 1500 clinical trial documents from ClinicalTrials.gov[2] are annotated with an average of 16 sentences per document. The manually labelled dataset statistics with the class distributions are specified in the Table 1. While manually inspecting the co-occurence statistics of different aspects in the same criteria sentence of the manually annotated dataset, we observe that around 30% of the eligibility criteria contains more than one aspect, with 65% containing health, life-style, demography aspects, while the remaining 35% contains demography and treatment. For facilitating further research, we will also provide some sample examples of the annotated corpus.

4 Methodology

In this work, we experiment with two different methods of aspect extraction. One of the following being the traditional supervised setup of using BiLSTM-CRF/Bi-GRU CRF with input representation optimized using categorical cross-entropy loss (Zhang and Sabuncu, 2018). The second one being the Co-Training (Blum and Mitchell, 1998) method to extract the semantic aspects which has been outlined in Algorithm 1. The later method uses two conditionally independent feature views of the same dataset illustrated below:

1. **Domain-independent**: The contextual pre-trained language models such as, BERT (Devlin et al., 2019) (**E1**) (or word2vec (Mikolov et al., 2013) trained on GoogleNews Corpus [3] (**E2**)) embeddings followed by a BiLSTM-CRF (C^1) (Huang et al., 2015) feature extractor.

2. **Domain-dependent:** Bio-BERT embeddings (Lee et al., 2020) (**E3**) (or word2vec trained on PubMed [4]) (**E4**) followed by BiGRU-CRF (C^2) (Lerner et al., 2020) feature extractor.

At each step of co-training, the classifiers C^1 and C^2 are trained on respective views of training sets V^1 and V^2, thereby minimizing the loss function. Each instance from the unlabeled samples (U) is scored using a scoring function computed as follows. First, the current classifier is used to decode the output label distribution for each word in the unlabeled instances. For each word in the output, we choose the output label which has the maximum probability. We compute the score for the sample as the multiplication of the probabilities of each label type for all labeled words in sequence normalized by the total number of words in the sentence. If this confidence score of the sample is greater than some pre-defined threshold τ, the sample has been added to the training set of the other classifier along with its output labels as generated by the classifier. This is the process of generation of weak labels for each sequence. Due to interchange of training data, both classifiers can learn from mistakes of each other and work in synergy.

5 Experimental Details

We implement the model using Pytorch 0.3.0. The two classifiers considered for co-training are C^1 : Bi-LSTM-CRF and C^2 : Bi-GRU-CRF. For both supervised and co-training methods, the training data is divided according to 70-30% train-validation split. The two different views of co-training setup are explained as follows:

Hyper-parameters for two independent views: We run two experiments based on co-training, one using contextual embeddings (C-CTr) and the other using context-independent embeddings (NC-CTr). The hyper-parameter settings for the two views as required by the co-training method are as follows:

[2] https://clinicaltrials.gov/

[3] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

[4] http://bio.nlplab.org/

View 1: For the first view (V^1), we use Bi-LSTM-CRF (Huang et al., 2015) with domain-independent word embeddings. We experiment with both a) (NC-CTr) Word2vec embeddings trained on GoogleNews Corpus with dimension 300 b) (C-CTr) pre-trained *bert-base* (12 layers, 12 attention heads, and 110 million parameters).

View 2: For the second view (V^2), we use Bi-GRU classifier with domain-dependent word embeddings. We experiment with both a) (NC-CTr) Word2vec embeddings trained on PubMed Corpus with dimension 200 b) (C-CTr) contextualized pre-trained Bio-BERT embeddings.

For both the classifiers, the hidden unit dimensions are set to 300. During training, we use Adam (Kingma and Ba, 2015) optimizer with a learning rate of 0.001 and a batch size of 64. For co-training, τ has been set to 0.5, epoch size to 200, with early-stopping employed based on the performance of validation set. All the results are reported based on the best hyper-parameter settings after an exhaustive grid search over parameter space.

Methods	Health	Trt	Lab	Demo	Life
	F1	F1	F1	F1	F1
Baseline-1	0.78	0.73	0.72	0.70	0.80
Baseline-1(1)	0.73	0.64	0.68	0.65	0.76
Baseline-1(2)	0.75	0.60	0.68	0.61	0.73
Baseline-2	0.73	0.43	0.66	-	-

Table 2: Macro-F1 score for all the aspects using prior methods with some additional features

Methods	Health	Trt	Lab	Demo	Life
	F1	F1	F1	F1	F1
C^1+E1	0.72	0.70	0.65	0.80	0.70
C^1+E2	0.68	0.61	0.62	0.75	0.67
C^2+E3	**0.73**	**0.70**	**0.66**	**0.81**	**0.72**
C^2+E4	0.70	0.64	0.63	0.77	0.67

Table 3: Feature ablations on our supervised setup on the train-validation split of our dataset.

6 Results and Analysis

In this section, we have provided a detailed analysis of the various results and findings that we have observed during experimentation. There are various criteria on which we have tried to evaluate our semi-supervised approach.

Methods	Health	Trt	Lab	Demo	Life
	F1	F1	F1	F1	F1
w/o CTrain	0.73	0.70	0.66	0.81	0.72
C-CTr(8K)	**0.85**	0.80	**0.83**	0.90	0.80
C-CTr(10K)	0.83	0.84	0.82	0.90	0.85
C-CTr(15K)	0.77	0.82	0.75	**0.92**	**0.88**
C-CTr (20K)	0.85	**0.86**	0.79	0.90	0.83
NC-CTr (8K)	0.76	0.77	0.74	0.74	0.75
NC-CTr (10K)	0.78	0.81	0.75	0.78	0.80
NC-CTr (15K)	0.77	0.76	0.77	0.81	0.77
NC-CTr (20K)	0.78	0.77	0.77	0.82	0.79

Table 4: Results showing various co-training methodology with different size of unlabelled instances. Trt=Treatment aspect. The scores are reported in the table based on exact match F1-score for all aspects.

Comparison with the baselines:
The results of the baseline methods are enumerated in Table 2. We report the results based on exact match of each type of the aspects using F1-score. Following (Luo et al., 2011), we implement the same (**Baseline-2**) on our dataset with UMLS (Bodenreider, 2004) feature representation and "bag-of-words" (BoW) features, and report results for various aspects. Although (Luo et al., 2011) assumes each criteria sentence essentially belongs to a single aspect, we have done an ablation of Baseline-2 without UMLS features (**Baseline-1(1)**) and without BoW (**Baseline-1(2)**). We observe that UMLS feature representation boosts up the performance due to inclusion of domain-specific information. We observed that this work finds resonance with (Chalapathy et al., 2016) in which the corpus uses multiple annotations. Due to availability of their working code, we have experimented with their stand-alone Bi-LSTM-CRF approach, used them as **Baseline-2** and report results for each of the first three annotated aspects.

Feature ablation on model architecture:
For the purpose of fair comparison, we experiment with different ablations of feature extractor and types of input representation (in the supervised setup) and present the results of Macro-averaged F1-score in Table 3. It has been observed that Bi-LSTM CRF with domain-specific input representation as Bio-BERT outperforms other ablations.

Impact of using co-training:
It is also evident from table 4, when the two independent views consist of contextualized embeddings (C-CTr), the model outperforms the

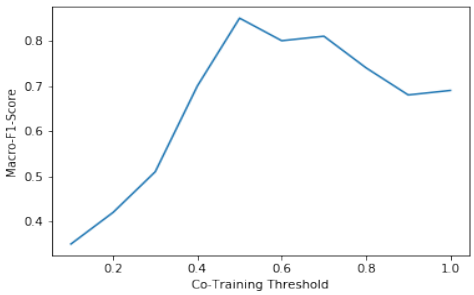

Figure 2: Testing the optimum Co-Training Threshold

non-contextualized features (NC-CTr) by an average margin of 6% F1-Scores. Also, we compare our best architecture for supervised setup with co-training approach. Given that the co-training model trains each classifier separately on different subsets of the training set, it can be sensitive to the choice of V^1 and V^2. In order to address this issue, we experiment with repeating the same experimnts with various random sampling of the two training subsets. We observe an average F1-score standard deviation (across multiple sampling) of 0.064 for Health class, 0.091 for Treatment class, 0.116 for Lab-Test Results class, 0.055 for Demography class and 0.008 for Life-style class.

Sensitivity of co-training parameters:
In figure 1, Macro-F1 score (across all aspects) of the co-trained model has been evaluated based on the values of co-training threshold. The values have been chosen from 0 to 1 at an interval of 0.1, in which the optimum value has been observed as 0.5. The sensitivity of co-training parameters has been shown in figure 2.

Effect of unlabelled data size:
Moreover, the results are fairly constant even when the unlabeled data size varies (enumerated in Table 4) which demonstrates the robustness of our approach. The contextualized representations when augmented with fair amount of semi-automatically annotated samples outperforms the supervised baseline setup.

7 Conclusion

In this paper, we have proposed a semi-supervised co-training method to tackle the scarcity of annotated data for the semantic clinical aspect extraction. This method augments a limited pool of annotated data with a large number of unlabeled clinical eligibility criteria outperforming pure supervised approaches. To the best of our knowledge, we are the first to provide an effective semi-supervised approach to detect the semantic aspects from clinical eligibility criteria which is a promising direction for further research on automatic linking of the patient Electronic Health Records (EHR) to clinical eligibility criteria with promising performance. As a future work, we aim to propose an end-to-end automatic matching system for patient-based clinical trial eligibility with low-cost data annotation.

Acknowledgments

This work has been done by the first author, Ishani, during her internship with the TCS Research Labs, India. Besides, the authors would like to thank the anonymous reviewers for their valuable feedback.

References

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA. Association for Computing Machinery.

William Boag, Kevin Wacome, Tristan Naumann, and Anna Rumshisky. 2015. Cliner : A lightweight tool for clinical named entity recognition.

O. Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. Bidirectional LSTM-CRF for clinical concept extraction. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 7–12, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter J. Embi, Anil K. Jain, and C. Martin Harris. 2008. Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment: A survey. *BMC Medical Informatics and Decision Making*, 8:13 – 13.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Bevan Koopman and Guido Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 669–672, New York, NY, USA. Association for Computing Machinery.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, D. Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Ivan Lerner, N. Paris, and Xavier Tannier. 2020. Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of biomedical informatics*, page 103356.

Zhihui Luo, Meliha Yetisgen-Yildiz, and Chunhua Weng. 2011. Dynamic categorization of clinical research eligibility criteria by hierarchical clustering. *J. of Biomedical Informatics*, 44(6):927–935.

M. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22:276 – 282.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Maryam Najafabadi, Flavio Villanustre, Taghi Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2.

Chaitanya Shivade, Courtney Hebert, Marcelo Lopetegui, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2015. Textual inference for eligibility criteria resolution in clinical trials. *J. of Biomedical Informatics*, 58(S):S211–S218.

Zhilu Zhang and Mert R. Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8792–8802, Red Hook, NY, USA. Curran Associates Inc.