Exploring ClinicalTrials.gov to Improve Information Extraction for Medical Literature

Abstract

Medical literature may be annotated into Participant (P), Intervention (I), Condition (C) and Outcomes (O) categories. Together, these are known as PICO elements. Our aim is to improve information extraction of PICO elements for medical literature. Specifically, we want to incorporate data from ClinicalTrials.gov into existing state of the art LSTM-CRF classifiers in order to improve their accuracy and recall. ClinicalTrials.gov is the directory of registered clinical trials which eventually lead to published medical literature. We are interested in this particular dataset because it is large, easily available and manually annotated. Even though there is not an exact translation between trials and studies, we hope to explore the commonalities between the two.

Background

The problem arises because PICO elements are described in abstracts in non-standard ways[1]. Even though 61.8% of PUBMED studies have structured abstracts and 38.2% have unstructured abstracts, the structure does not directly lend itself to be broken down into PICO elements. The most common structure is IMRAD (Introduction, Methods, Results, and Discussion) which is used 66.5% of the time for structured abstracts and the second most common structure is the 8-heading format (objective, design, setting, patients, intervention, main outcome measures, results, and conclusions) which is used only 33.5% of the time. Therefore, there is a need for automated solutions to better support intelligent medical search.

The current implementation of the LSTM-CRF is trained on the EBM-NLP corpus of 4,741 medical article abstracts from PUBMED which has been manually crowd-sourced for PICO annotations via Amazon Mechanical Turk[2]. For testing purposes, we used a set of 191 abstracts annotated for P, I, and O by medical professionals on Upwork. For the rest of this report, we will refer to the manually annotated EBM-NLP corpus as EBM-NLP and that clinicaltrials.gov data as CT.

---

[1] Nakayama, T., Hirai, N., Yamazaki, S., & Naito, M.F. (2005). Adoption of structured abstracts by general medical journals and format for a structured abstract. *Journal of the Medical Library Association : JMLA, 93 2*, 237-42.

[2] Patel, Roma, et al. "Syntactic Patterns Improve Information Extraction for Medical Search." *arXiv preprint arXiv:1805.00097* (2018).

Data

We scraped clinicaltrials.gov for all registered trials. We then filtered trials by those that had the PMID (PUBMED ID) field filled in which resulted in a total of 18,887 studies which is a meager 6.7% of the entire dataset. However, many trials have several studies associated which results 68,137 total linked abstracts.

The next step was processing the data. Once we established the trials that had associated PMIDs, we did exact string matching to see if the the annotated PICO elements in the trials matched with phrases in the abstract. We were careful to convert everything to lowercase while distinguishing between acronyms. For the 8625 abstracts where there is a one-to-one mapping between trials and studies, we got the following results:

Category	**Match**	**No match**
Interventions	54.8% (4451)	45.2% (3663)
Outcomes	20.4% (1607)	79.6% (6270)

For the 59512 abstracts where there is a one-to-many mapping between trials and studies, we get the following results:

Category	**Match**	**No match**
Interventions	19.3% (11,514)	80.7% (39,662)
Outcomes	11.2% (6,121)	88.8% (50,332)

This illustrates that we were able to find more exact matches for interventions and outcomes and that we were able to find more matches for both interventions and outcomes when there was a one-to-one mapping between studies and abstracts. Even though the percentage of matches are low, the absolute numbers are significant and we decided to work with this data. We tried to use cosine similarity to gather some more matches for outcomes specifically, but the gains were nominal.

Methodology

Our first approach was to combine the CT data with the EBM data to train the classifier and compare the results with classifiers trained with just EBM (state of the art) and just CT.

For interventions, we found that the CT data did much better on its own and did worse when combined with EBM data:

	F1	Precision	Recall
EBM	42.72	46.96	39.18
CT	**53.75**	**75.77**	**41.65**
Combined	42.51	49.31	37.36

For conditions, we found that the CT data did very poorly and the EBM-trained LSTM-CRF did much better:

	F1	Precision	Recall
EBM	**39.13**	**43.96**	**35.26**
CT	5.14	34.51	2.78
Combined	36.27	44.77	30.48

We concluded that using both the EBM and CT datasets do not give us better results except in the case of interventions. Our next stop was to try look-up tables and see if we could directly lookup extract PICO elements from abstracts. On average, there are 2.72 conditions, 8.05 interventions and 7.34 outcomes in each abstract. For our test set, we used 195 abstracts that are annotated by medical professionals and Amazon Mechanical Turk workers alike[3].

We used medical ontologies like ICD10 and SNOMED, as well as manually entered conditions data from ClinicalTrials.gov (which was subsequently arranged by frequency of occurrence), as different lookup tables to extract conditions. The results from ICD10 were dismal (an F1 of 20.5%) and are not reported below. An example of a condition

[3] Nye, Benjamin, et al. "A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature." *arXiv preprint arXiv:1806.04185* (2018).

missed by ICD is 'autism' where the ICD10 code is \emph{autistic spectrum disorder}. We also trained an LSTM-CRF with 2000 annotated abstracts as training data and analyze the results below.

The best overall results are achieved with SNOMED but highest precision of 65.40% comes with the LSTM-CRF whereas the highest recall comes from the Clinical Trials gazetteer. Both the LSTM-CRF and SNOMED completely miss the majority of the conditions and are therefore unsuitable to be deployed for any large-scale deployment of medical literature.

We used the SNOMED corpora, the lookup table obtained by aggregating clinical trials for **conditions** (ordered by frequency) and compared that against results from the LSTM-CRF and Amazon MTurk crowd-sourced annotations:

Model	Completely found	Partially found	Completely missed	F1	Recall	Precison
Amazon MTurk	30.50%	4.51%	46.32%	41.54%	35.00%	51.09%
SNOMED	27.11%	18.26%	54.04%	38.86%	45.37%	33.99%
Clinical Trials - frequency	35.40%	38.04%	21.65%	12.20%	73.44%	6.65%
LSTM-CRF	24.85%	1.12%	66.10%	37.18%	25.97%	65.40%

We used the SNOMED corpora, the lookup table obtained by aggregating clinical trials and then filtering by the EBM corpus for **interventions** and compared that against results from the LSTM-CRF and Amazon MTurk crowd-sourced annotations:

Model	Completely found	Partially found	Completely missed	F1	Recall	Precison
Amazon MTurk	49.77%	0%	0%	48.28%	49.77%	46.87%
SNOMED	35.73%	16.76%	46.00%	38.75%	52.49%	30.71%
Clinical Trials/EBM filtered	40.48%	15.46%	41.13%	40.23%	55.94%	31.41%
LSTM-CRF	37.10%	0.65%	22.02%	42.66%	37.75%	49.03%

We used the SNOMED corpora, the lookup table obtained by aggregating clinical trials for **outcomes** (ordered by longest) and compared that against results from the LSTM-CRF and Amazon MTurk crowd-sourced annotations:

Model	Completely found	Partially found	Completely missed	F1	Recall	Precison
Amazon MTurk	56.23%	0%	0%	54.32%	56.23%	52.53%
SNOMED	2.70%	13.82%	83.39%	21.84%	16.52%	33.28%
Clinical Trials - longest	30.43%	42.40%	22.02%	34.34%	72.83%	22.47%
LSTM-CRF	35.28%	2.49%	23.73%	42.14%	37.77%	47.66%

Conclusions

Through various experiments, we first determined that using the data from CT and relying on PICO annotations during the trial registration stage would not be a viable approach to extending the EBM corpus. Except for interventions, we found that the CT data does much poorly than the EBM corpus. Alternatively, we also tried various lookup tables to automatically annotate for PICO elements - including ICD10 codes, SNOMED corpora - and found mixed results. While lookup tables generally give higher recall, their precision is worse than using an LSTM-CRF trained on EBM.